OPSO79-1-UCSH2021

Estadística descriptiva I

08/10/2021

Estadística descriptiva I

Describir variables

Lo hacemos para conocer un conjunto de datos particular (e.g. ESI o CASEN)

Describimos las variables conociendo su distribución, como las observaciones caen en una serie de valores posibles.

El método que usamos para analizar los datos dependerá del tipo de variables

- Cuantitativas: será clave describir el centro y la variabilidad
- Categóricas: importará describir el número de observaciones que caen en cada categoría

La descripción se puede hacer mediante **resúmenes numéricos**, **tablas** o **gráficos**.

Describir categóricas

Lo más simple es mediante tablas, con valores absolutos o relativos.

```
esi <- haven::read_sav("data/esi-2019---personas_s.sav")
```

```
sjmisc::frq(esi$ocup_form)
```

```
##
## Ocupados según formalidad (x) <numeric>
## # total N=96240  valid N=42856  mean=1.31  sd=0.46
##
## Value |             Label |     N | Raw % | Valid % | Cum. %
## -----------------------------------------------------------
##     1 |    Ocupado formal | 29607 | 30.76 |   69.08 |  69.08
##     2 |  Ocupado informal | 13249 | 13.77 |   30.92 | 100.00
##  <NA> |              <NA> | 53384 | 55.47 |    <NA> |   <NA>
```

Los valores perdidos (NA) son relevantes de considerar (por eso Raw % y Valid %)

```
table(esi$ocup_form,useNA = "ifany") # Con r base
```

Describir categóricas

Gráficamente, se pueden describir mediante gráficos **de barras** y **de tortas**.

Lo primero es hacer la tabla

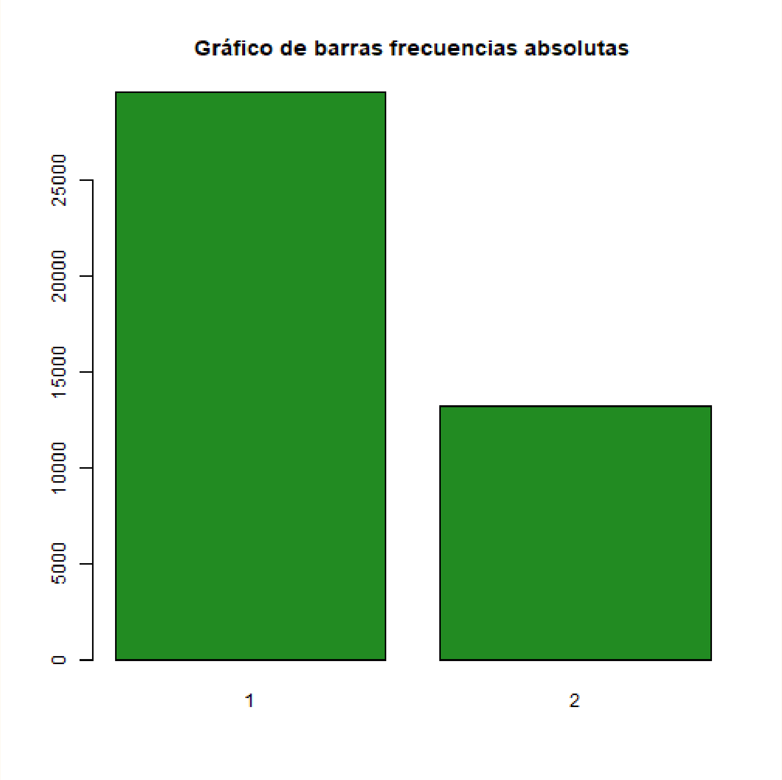
```
table(esi$ocup_form)
```

Luego introducir tabla a función de gráfico.

```
barplot(table(esi$ocup_form))
```

Describir categóricas

```
barplot(table(esi$ocup_form))
```



Describir categóricas

Cada barra vertical representa una categoría.

La altura de la barra es la frecuencia de observaciones en cada categoría.

Entre cada barra hay un espacio.

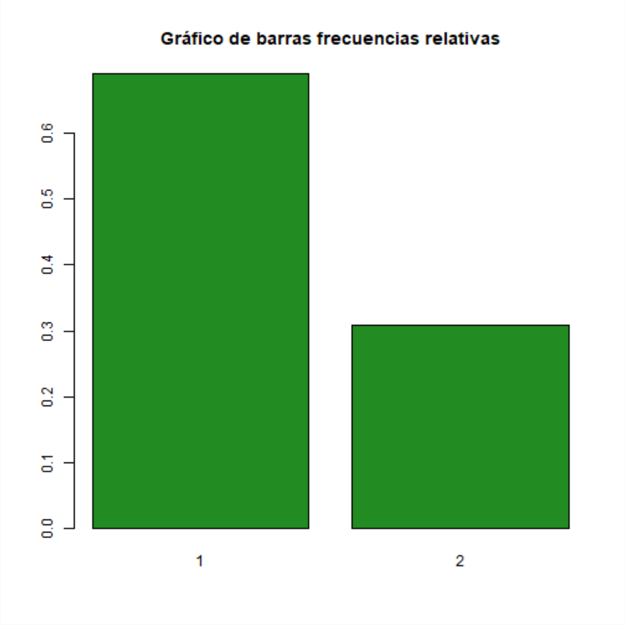
Para gráfica con frecuencia relativas hacer tabla de proporciones:

```
prop.table(table(esi$ocup_form))
```

Esta tabla debe ir dentro de `barplot()`

Describir categóricas

```
barplot(prop.table(table(esi$ocup_form)))
```

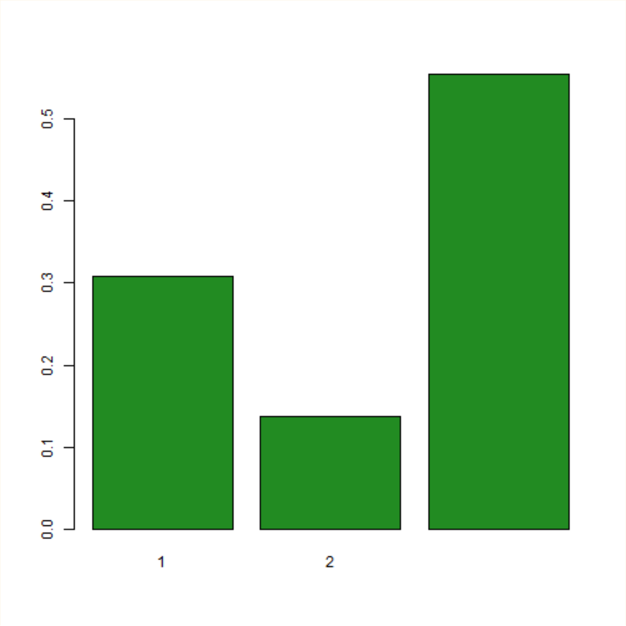


Describir categóricas

Podemos incluir los NA.

```
prop.table(table(esi$ocup_form, useNA = "ifany"))
```

```
barplot(prop.table(table(esi$ocup_form, useNA = "ifany")),
        col="forestgreen")
```



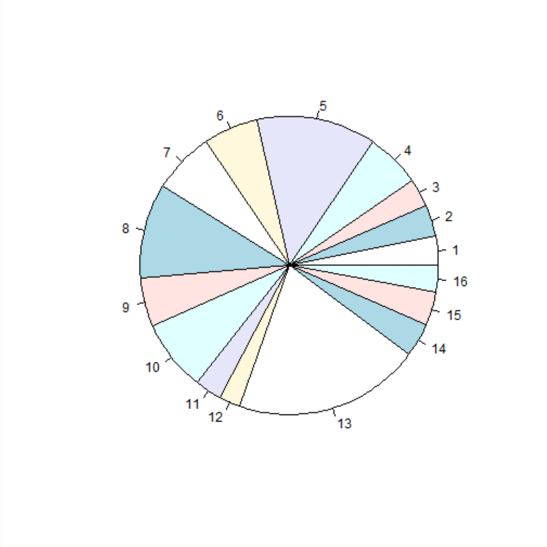
Describir categóricas

Toda una discusión respecto de los **gráficos de tortas**.

Consisten en un circulo, donde hay un "trozo" de pie para cada categoría.

El tamaño del trozo de pie corresponde al porcentaje de observaciones en cada categoría.

```
pie(prop.table(table(esi$region)))
```



Describir ordinales

¿Como presentar figuras de variables categóricas ordinales? (tipo Escala Likert)

Cargar paquete y datos de Latinobarómetro.

```
library(sjPlot)
data <- readRDS("data/Latinobarometro_2018_Esp_R_v20190303.rds")
```

Identificar y seleccionar variables

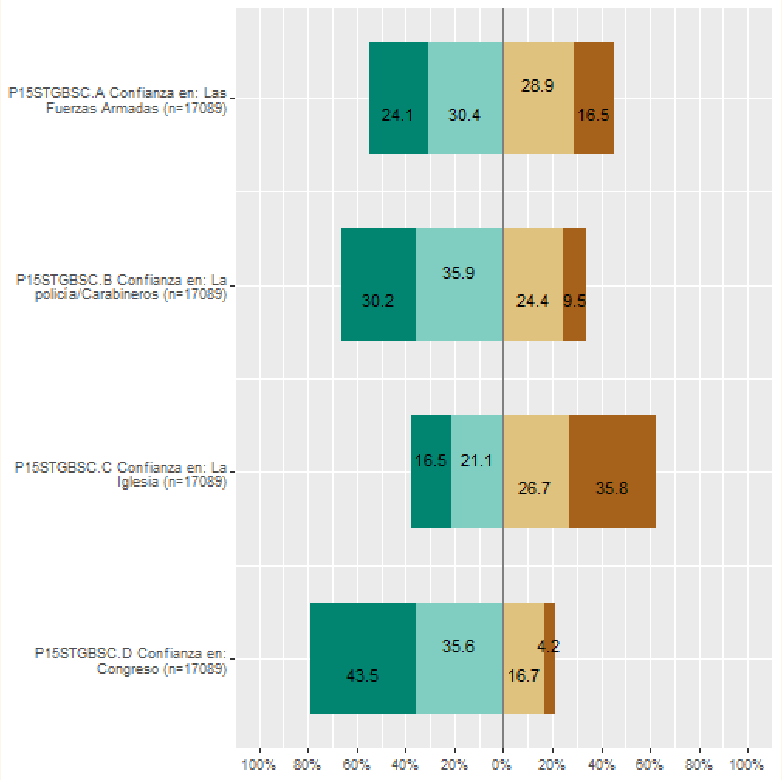
```
sjmisc::find_var(data,"Confianza")[2:5,3]
```

```
## [1] "P15STGBSC.A Confianza en: Las Fuerzas Armadas"
## [2] "P15STGBSC.B Confianza en: La policía/Carabineros"
## [3] "P15STGBSC.C Confianza en: La Iglesia"
## [4] "P15STGBSC.D Confianza en: Congreso"
```

```
seleccion <- data %>% select(starts_with("P15STGB")) %>%
  filter(P15STGBSC.A>0,P15STGBSC.B>0,P15STGBSC.C>0,P15STGBSC.D>0)
```

Describir ordinales

```
plot_likert(seleccion[,1:4],catcount = 4) +
  ggplot2::theme(legend.position = "none")
```



Describir cuantitativas

Medidas de tendencia central

Media

La suma de las observaciones dividido por el n.

$$\bar{x} = \frac{\sum x}{n}$$

```
a <- c(1,4,5,6,7,8,8,9,3,5,6,7,8)
```

```
sum(a)/length(a)
```

```
## [1] 5.923077
```

```
mean(a)
```

```
## [1] 5.923077
```

Describir cuantitativas

Moda

Es el valor más frecuente (puede ser uno o varios).

Útil cuando la variable toma pocos valores.

```
table(a)
```

```
## a
## 1 3 4 5 6 7 8 9
## 1 1 1 2 2 2 3 1
```

```
table(a) %>% as.data.frame() %>% arrange(-Freq) %>% slice(1)
```

```
##   a Freq
## 1 8    3
```

Describir cuantitativas

Medidas de tendencia central

Mediana

Valor que se encuentra en la mitad de las observaciones cuando las observaciones son ordenadas de la más pequeñas a la más larga.

```
a[order(a)]
```

```
##  [1] 1 3 4 5 5 6 6 7 7 8 8 8 9
```

```
a[order(a)][round((length(a))/2)]
```

```
## [1] 6
```

```
median(a)
```

```
## [1] 6
```

Describir cuantitativas

Medidas de tendencia central

Usalmente la media no es igual a ningún valor de los observados en la variable.

La mediana, por el contrario, siempre es igual a uno de los valores observados.

La media es fuertemente influenciada por valores atípicos (outlier), mientras que la mediana no.

Un **valor atípico** es una observación que está muy por encima o muy por debajo del volumen general de los datos.

```
a <- c(1,4,5,6,7,8,8,9,3,5,6,7,8,500) ## sesgada a la derecha
mean(a)
```

```
## [1] 41.21429
```

```
median(a)
```

```
## [1] 6.5
```

Describir cuantitativas

Medidas de tendencia central

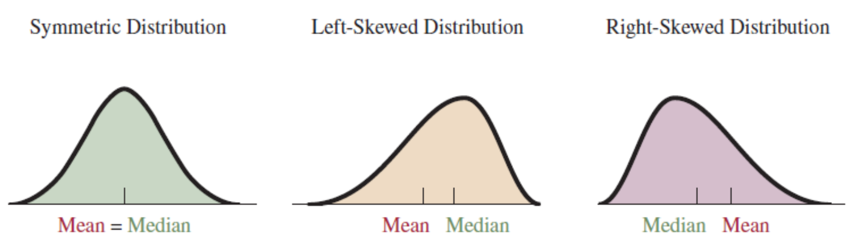
Valor extremo negativo (distribución **sesgada** a la izquierda)

```
a <- c(-1000,1,4,5,6,7,8,8,9,3,5,6,7,8)
mean(a)
```

```
## [1] -65.92857
```

```
median(a)
```

```
## [1] 6
```



▲ **Figure 2.9** Relationship Between the Mean and Median. Question For skewed distributions, what causes the mean and median to differ?

Describir cuantitativas

Medidas de variabilidad

El rango es la diferencia entre la observación más alta y la más pequeña.

```
min(a)
```

```
## [1] -1000
```

```
max(a)
```

```
## [1] 9
```

```
summary(a)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1000.00     4.25     6.00   -65.93     7.75     9.00
```

Describir cuantitativas

Medidas de variabilidad

Un mejor resumen de la variabilidad consiste en usar todo el dato.

¿Cuál es la distancia típica en la que caen las observaciones respecto de la media?

¿Que tanto se desavían las observaciones respecto de la media?

La desviación de una observación respecto de su media es $x - \bar{x}$

El promedio de las desviaciones sería la **desviación estándar**:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Describir cuantitativas

Medidas de variabilidad

En R

```
a <- c(1,4,5,6,7,8,8,9,3,5,6,7,8)
sd(a)
```

```
## [1] 2.289889
```

Las observaciones de a se desvían tipicamente 2,29 unidades respecto de la media a a

```
a <- c(1,4,5,6,7,8,8,9,3,5,6,7,8,500)
sd(a)
```

```
## [1] 132.0659
```

Con el valor 500 dentro de a, la variabilidad de los datos aumenta en cerca de 130 unidades.

Describir cuantitativas

Medidas de variabilidad

Mientras mayor sea la variabilidad de los datos respecto de la media, mayor será s

La desviación estándar sería 0 si todas las observaciones tienen el mismo valor (la menor variabilidad posible para una muestra)

Al igual que la media, s es sensible a la presencia de outliers o valores atípicos.

La **varianza**

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Describir cuantitativas

Medidas de variabilidad

Coeficiente de variación permite comparar la variabilidad de dos variables con unidades de medida distintas y medias diferentes.

Dado que `s` está en la unidad de medida de una variable particular, y hace referencia a la variabilidad respecto de la media, se necesita de una medida estandarizada.

Esta medida es el **coeficiente de variación**

$$cv = \frac{\sigma}{\bar{x}}$$

Para usarla los datos deben ser positivos.

A mayor valor de `cv` mayor heterogeneidad de los valores de la variable, a menor cv (30% o menos) mayor homogeneidad (media es informativa).

Describir cuantitativas

Medidas de variabilidad

```
a <- c(1,4,5,6,7,8,6,6,6,6,6,8,9,3,5,6,7,8)
sd(a)/mean(a)
```

```
## [1] 0.3237002
```

```
sd(esi$edad)/mean(esi$edad)
```

```
## [1] 0.5967146
```

```
sd(esi[esi$ing_t_d>0,]$ing_t_d)/mean(esi$ing_t_d)
```

```
## [1] 3.281442
```

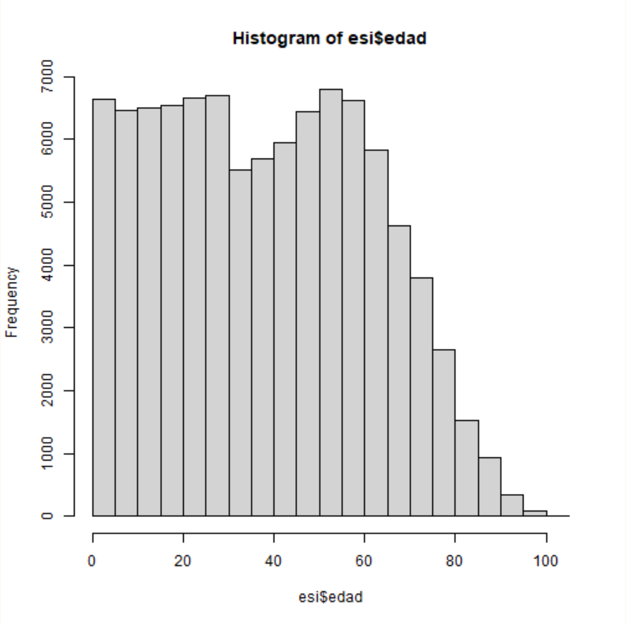
Si el C.V es menor o igual al 30%, significa que la media aritmética es representativa del conjunto de datos, por ende el conjunto de datos es "Homogéneo".

Describir cuantitativas

Histogramas

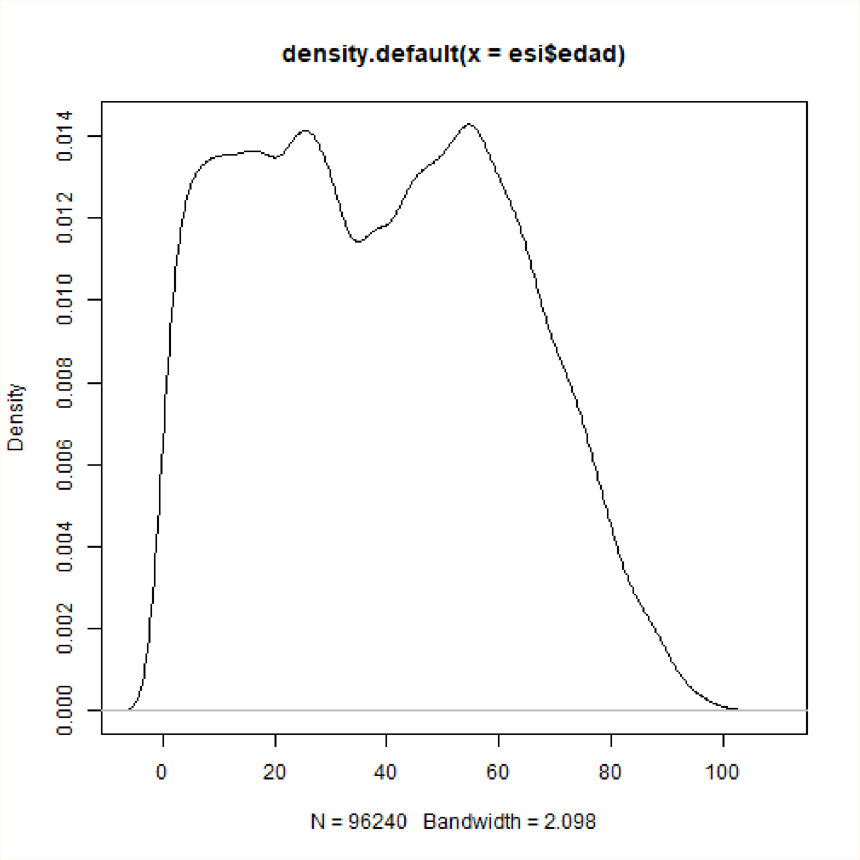
Un histograma es un gráfico que usa barras para representar las frecuencias o las frecuencias relativas de los valores de una variable cuantitativa

Distribución bimodal y no sesgada de edad. Media 38.8, Mediana 38 y s 23.1



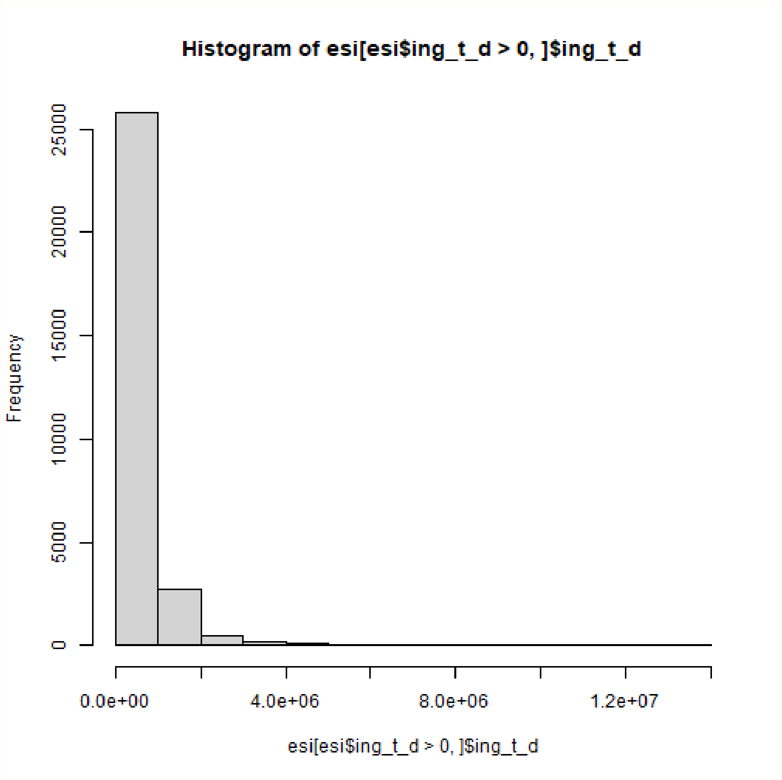
Describir cuantitativas

Gráficos de densidad



Describir cuantitativas

Histogramas



Describir cuantitativas

Histogramas

¿Cuál es el problema?

R divide el rango de valores de la variable en intervalos de igual "anchura" (width)

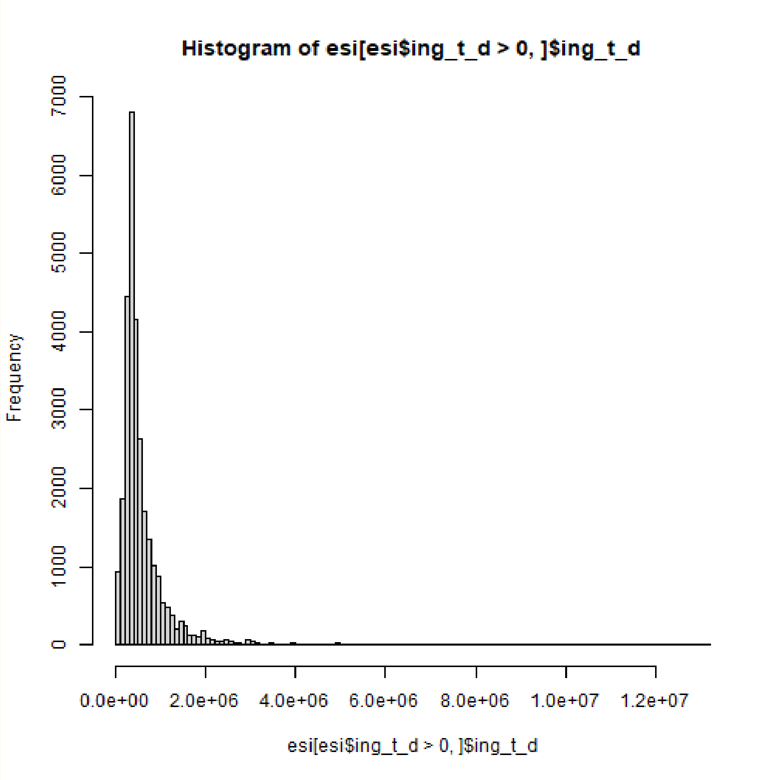
Luego cuenta el número de observaciones en cada intervalo

Luego esa tabla es graficada. Cada barra es un intervalo, su altura corresponde a la frecuencia de cada intervalo.

Por defecto dividió los datos en 5 intervalos. Hay una interesante variabilidad en la primera barra que no se puede observar.

Describir cuantitativas

```
hist(esi[esi$ing_t_d>0,]$ing_t_d, breaks = 100)
```



¿Como es la distribución de la variable ing_t_d?

Describir cuantitativas

Histogramas

`ing_t_d` tiene una distribución unimodal centrada en $413.356 (mediana).

El promedio de ingresos es $592.400, más de $150.000 más que la mediana, lo que es un indicador de una distribución sesgada a la derecha.

Valores de ingresos muy altos (outliers) están "empujando" a la media a ser mayor que la mediana.

Los ingresos varían entre $5.000 y $13.130.848, con una s de $590.429

```
sjmisc::descr(esi[esi$ing_t_d>0,]$ing_t_d)
```

```
##
## ## Basic descriptive statistics
##
##  var    type                              label     n NA.prc     mean
##   dd numeric Total ingresos sueldos y salarios 29229      0 592440.2 590428
##        se        md trimmed                               range       iqr skew
##  3453.51 413355.7  485239 13125847.61 (5000-13130847.61) 384798.2 4.87
```

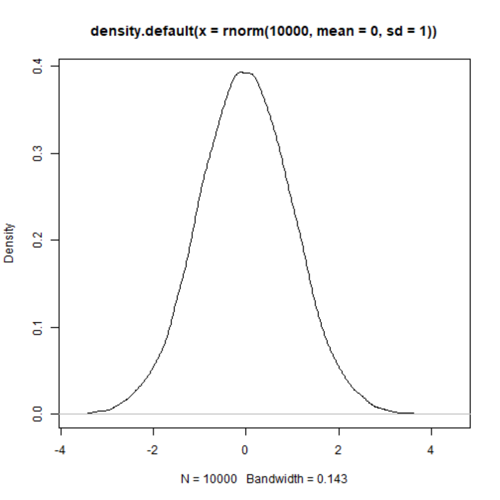
Describir cuantitativas

Medidas de forma

El sesgo (skew) de la variable se observa en el histograma, pero también puede ser numerado.

Este indicador, junto a la curtosis, nos informan sobre el grado de normalidad de la distribución

```
plot(density(rnorm(10000,mean = 0,sd=1)))
```



Describir cuantitativas

Medidas de forma

- **Asimetría:** sesgo respecto de la media (horizontal). Si una distribución es simétrica, existe el mismo número de valores a la derecha que a la izquierda de la media
- **Curtosis:** mide el grado en que las puntuaciones están agrupadas en torno al punto central (vertical)

Describir cuantitativas

Medidas de forma: asimetría

¿cuándo es demasiada la asimetría?

- entre -0.5 y 0.5 datos bastante simétricos.
- entre -1 y -0.5 (sesgado negativamente) o entre 0.5 y 1 (sesgado positivo), datos moderadamente sesgados.
- Si el sesgo es menor que -1 (sesgo negativo) o mayor que 1 (sesgo positivo), los datos están muy sesgados.

skew de ingresos es 4.87. De edad 0.16

```
aleatoria <- rnorm(10000,mean = 0,sd=1)
moments::skewness(aleatoria)
```

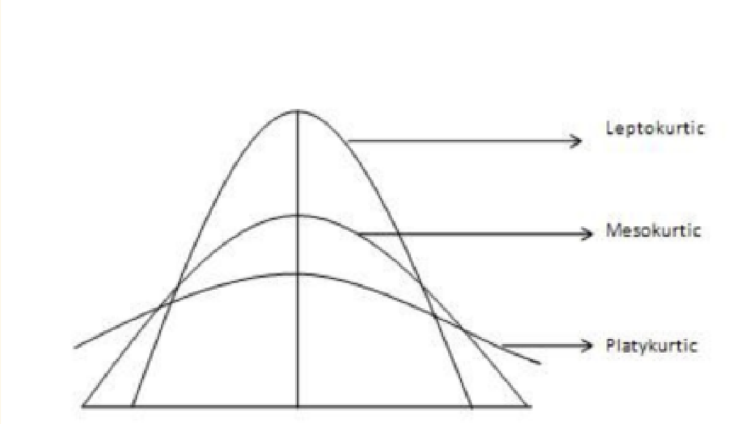
```
## [1] -0.00384941
```

Describir cuantitativas

Medidas de forma: curtosis

¿cuándo es demasiado la curtosis?

- Mesocurtica: curtosis similar a la normal
- Leptocurtica (Curtosis > 3): colas son más cortas, pico más alto y agudo (concentración de valores de la variable muy cerca de la media de la distribución)
- Platicurtica: (Curtosis < 3): representa un reducido grado de concentración alrededor de los valores centrales de la variable.



Describir cuantitativas

Medidas de forma: curtosis

```r
moments::kurtosis(rnorm(10000,mean = 0,sd=1))
```

```
## [1] 2.956642
```

```r
moments::kurtosis(esi$edad)
```

```
## [1] 2.005933
```

```r
moments::kurtosis(esi$ing_t_d)
```

```
## [1] 71.21357
```

Describir cuantitativas

Medidas de posición

La mediana es un caso especial de un repertorio general de posiciones: **percentiles**

El p-ésimo percentil es un valor tal que el p por ciento de las observaciones caen por debajo o en ese valor.

La mediana es el percentil 50: el 50% de las observaciones de una variable están en la mediana o bajo ella.

Tambien se le llama segundo cuartil.

En la ESI, el 50% de las observaciones tienen 38 o menos años.

```
median(esi$edad)
```

```
## [1] 38
```

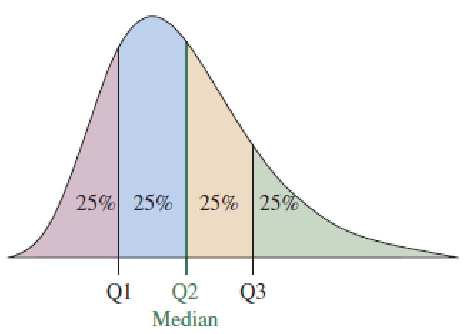
Describir cuantitativas

Medidas de posición

Hay otros dos percentiles que se ocupan bastante: percentil 25 (primer cuartil) y percentil 75 (tercer cuartil)

```
summary(esi$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   19.00   38.00   38.76   57.00  104.00
```



▲ **Figure 2.14 The Quartiles Split the Distribution Into Four Parts.** Twenty-five percent is below the first quartile (Q1), 25% is between the first quartile and the second quartile (the median, Q2), 25% is between the second quartile and the third quartile (Q3), and 25% is above the third quartile. **Question** Why is the second quartile also the median?

Describir cuantitativas

Medidas de posición

En R se puede especificar cuál percentil queremos.

```
quantile(esi$edad, prob=c(.25))
```

```
## 25%
##  19
```

```
quantile(esi$edad, prob=c(.5))
```

```
## 50%
##  38
```

```
quantile(esi$edad, prob=c(.75))
```

```
## 75%
##  57
```

Describir cuantitativas

Medidas de posición

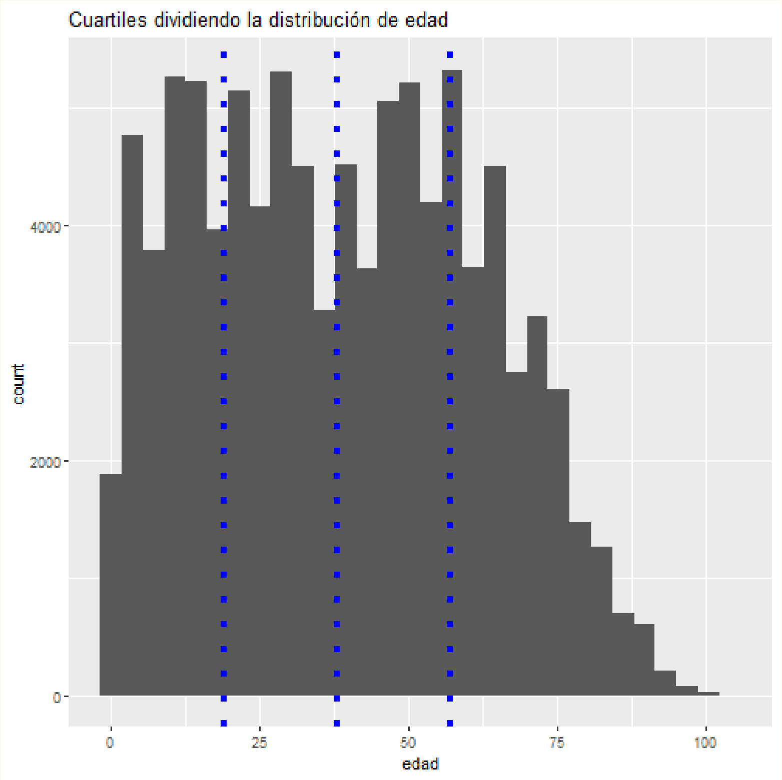
Incluso en el marco de `dplyr` y agrupamientos:

```
esi %>% group_by(sexo) %>% summarise(p25=quantile(esi$edad, prob=c(.25)
                                     p50=quantile(esi$edad, prob=c(.50)
                                     p75=quantile(esi$edad, prob=c(.75)
```

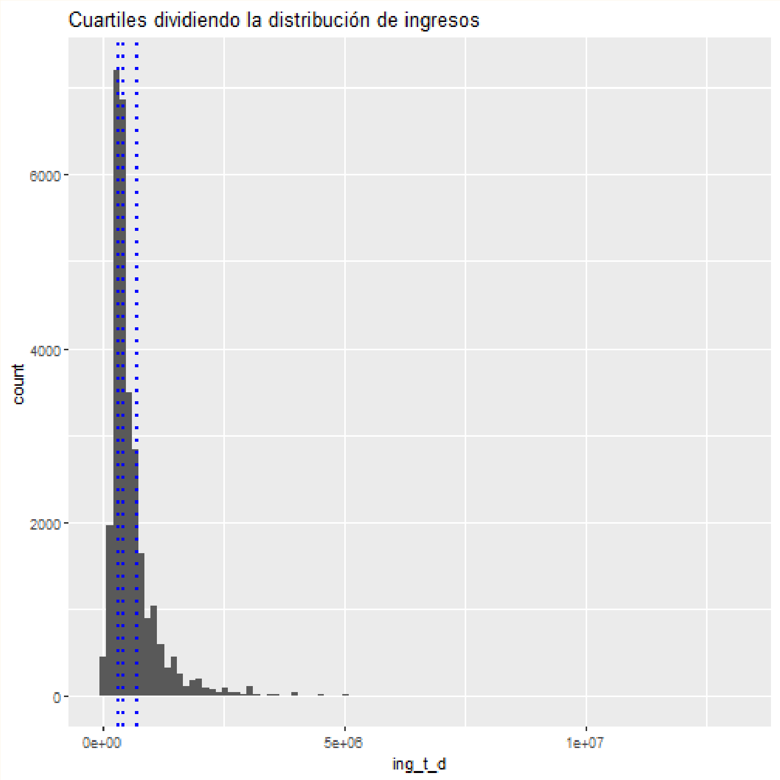
```
## # A tibble: 2 x 4
##          sexo   p25   p50   p75
##     <dbl+lbl> <dbl> <dbl> <dbl>
## 1 1 [Hombre]     19    38    57
## 2 2 [Mujer]      19    38    57
```

Ojo que los percentiles marcan posiciones. Son un estadístico. No son grupos.

Describir cuantitativas



Describir cuantitativas



Describir cuantitativas

Medidas de variabilidad

El rango intercuartílico (IQR) es la distancia entre el tercer y primer cuartil.

$$IQR = Q3 - Q1$$

La medida resume el rango de valores de la mitad central de los datos

```
summary(esi[esi$ing_t_d>0,]$ing_t_d)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     5000   300767   413356   592440   685566 13130848
```

Describir cuantitativas

Medidas de variabilidad: IQR a mano

```
685566-300767
```

```
## [1] 384799
```

```
esi %>% filter(ing_t_d>0) %>% count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 29229
```

```
esi %>% filter(ing_t_d>0) %>%
  filter(ing_t_d<=685566 & ing_t_d>=300767) %>%  count()
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 14678
```

Describir cuantitativas

Medidas de variabilidad: IQR a mano

```
14678    / 29229
```

```
## [1] 0.5021725
```

```
sjmisc::descr(esi[esi$ing_t_d>0,]$ing_t_d)
```

```
##
## ## Basic descriptive statistics
##
##   var    type                             label     n NA.prc     mean
##    dd numeric Total ingresos sueldos y salarios 29229      0 592440.2 590428
##        se       md trimmed                          range       iqr skew
##   3453.51 413355.7  485239 13125847.61 (5000-13130847.61) 384798.2 4.87
```

Describir cuantitativas

Detección de outliers

El IQR se utiliza para detectar casos atípicos

Outlier superior: si un caso se encuentra 1,5 por IQR veces sobre Q3

Outlier inferior: si un caso se encuentra -1,5 por IQR veces bajo Q3

¿El caso máximo de ingresos es un outlier?

```
max(esi$ing_t_d)
```

```
## [1] 13130848
```

Q3 + (IQR * 1,5)

```
685566 + (384799*1.5)
```

```
## [1] 1262765
```

Describir cuantitativas

Detección de outliers

¿El caso máximo de edad (104) es un outlier?

```
summary(esi$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   19.00   38.00   38.76   57.00  104.00
```

Q3 + (IQR * 1,5)

```
57 + (38*1.5)
```

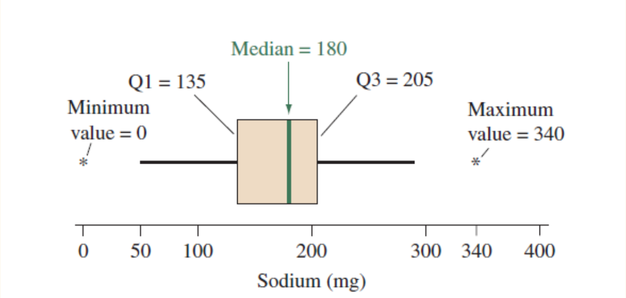
```
## [1] 114
```

El caso máximo de la edad está bajo 114, por lo que no es un outlier.

Describir cuantitativas

Gráficos de cajas y bigotes

Integra las diferentes medidas que hemos visto



La caja va desde el cuartil más bajo Q1 al quartil superior Q3

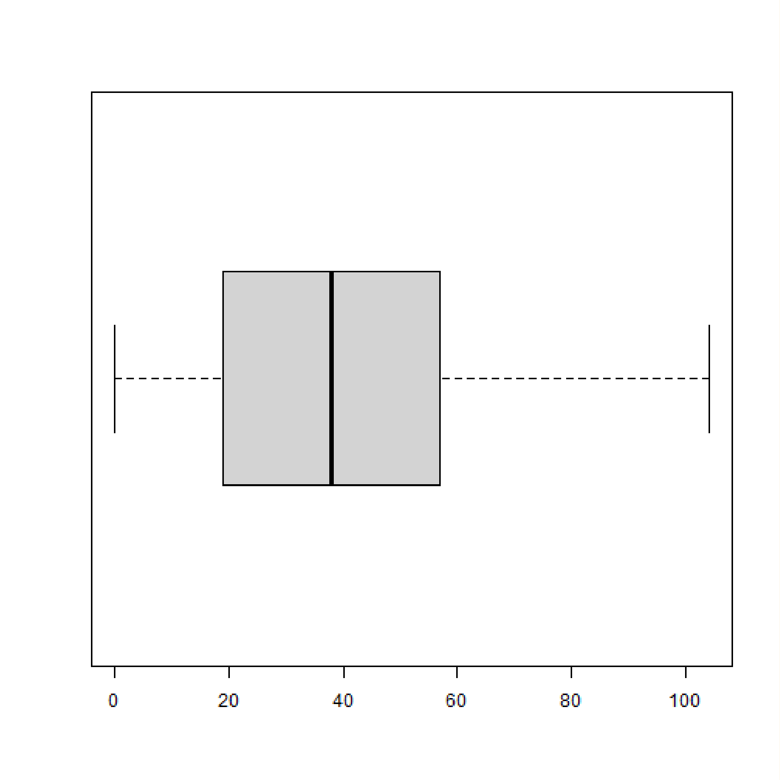
La línea dentro de la caja indica la mediana

Los bigotes (línea horizontal) marca la zona de valores típicos, que va entre Q1-(1,5xIQR) y Q3+(1,5xIQR)

Los casos fuera de los bigotes son los potenciales outliers.

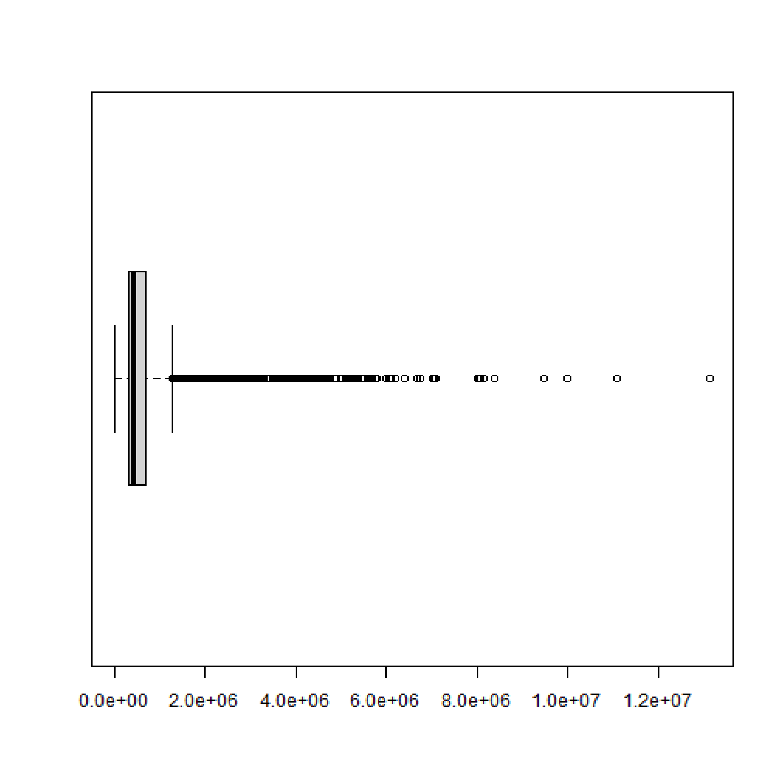
Describir cuantitativas

```
boxplot(esi$edad,horizontal = TRUE)
```

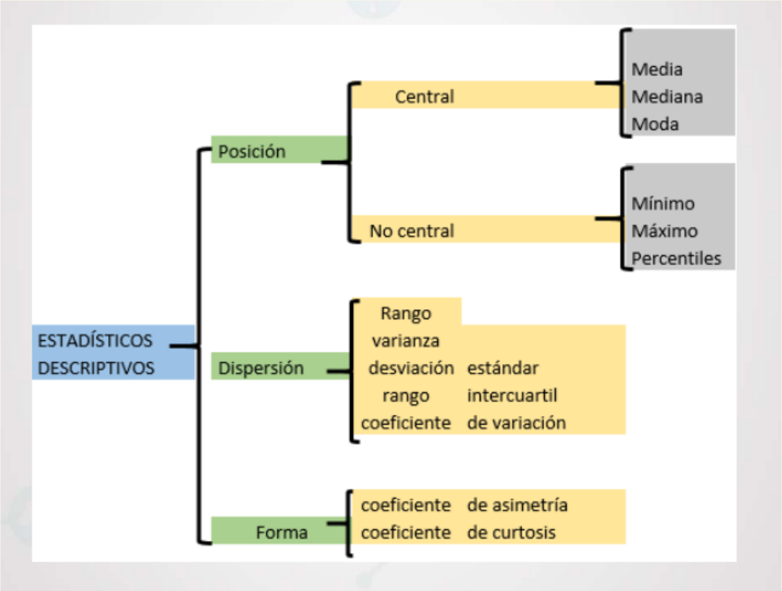


Describir cuantitativas

```
boxplot(esi[esi$ing_t_d>0,]$ing_t_d,horizontal = TRUE)
```



Resumen de lo visto



Sobre Prueba 2

22 de octubre (15 de octubre semana de receso)

Para practicar hemos dejado una guía que les permitirá además subir nota de prueba 1: https://opso791ucsh2021.netlify.app/clases/clase7/ejerciciopractico_latinobarometro.pdf

La prueba será similar.

Para esta prueba solo hay tres capítulos obligatorios:

- Mario Triola (2009) Estadísticas (cap. 2 y 3)
- Hadley Wickham (2021) R para ciencia de datos (cap 18)

Ambos para descarga en el siguiente link

Recursos web utilizados

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Para seguir aprendiendo

Diagrama de barras

Histogramas y gráficos de densidad

Asimetría y curtosis

Bibliografía utilizada

Agresti, A. and C. Franklin (2018). *Statistics the Art and Science of Learning from Data*. Pearson Education Limited.

Triola, M. F. (2009). *Estad\'istica. Décima Edición, Editorial Pearson Educación*. México, DF: Pearson.