

ANÁLISIS UNIVARIABLE SOBRE PROBABILIDAD DE VOTO SEGÚN BARÓMETRO DEL CIS DE JULIO 2023 (ESTUDIO 3415)

Técnicas de Investigación en Ciencia Política I. UBU
Práctica 1. Apartado 4

Tomàs Ferrandis Moscardó

2023-03-22

Índice

1 INTRODUCCIÓN	2
2 OBTENCIÓN DE DATOS	3
2.1 Fichero y Datos en Abierto	3
2.2 Creación de la base de datos (<i>data frame</i>)	3
2.3 Consultar el cuestionario y el diccionario de datos	3
3 PREPARACIÓN DE LA MATRIZ DE DATOS	5
3.1 Limpieza de la base de datos.	5
3.2 Recodificación	5
4. DISTRIBUCIÓN DE FRECUENCIAS	7
4.1 Gráfico de barras	7
4.2 Tabla de frecuencias	7
5 TRATAMIENTO COMO VARIABLE CUANTITATIVA	9
5.1 Medidas de tendencia central	9
5.2 Comparación entre Moda, Media y Mediana.	11
5.3 Medidas de posición no central	12
5.4 Medidas de dispersión	15
6 RESUMEN VALORES	18
7 EJECUCIÓN DEL Rmd Y REPOSITORIO DE FICHEROS	19

1 INTRODUCCIÓN

Esta actividad corresponde al apartado 4º de la Práctica 1 de la asignatura de Técnicas de Investigación en Ciencia Política I. Se trata de un *análisis univariado* sobre la variable PROBVOTO “Escala de probabilidad de ir a votar (0-10) en las elecciones generales de 2023” de la encuesta del Barómetro de julio de 2023 (estudio 3415) vista en la actividad anterior. La pregunta en concreto es: *Como Ud. sabe, el domingo 23 de julio se van a celebrar elecciones generales en España. Para comenzar me gustaría que me dijera cuál es la probabilidad de que Ud. vaya a votar. Para contestar utilice una escala de 0 a 10 en la que el 0 significa “con toda seguridad no iría a votar” y 10 “con toda seguridad, iría a votar”*. (pregunta P1)

Al tratarse también de un *análisis univariado* este será descriptivo aunque se le dará un tratamiento distinto. Pese a ser una variable cualitativa ordinal (Valores de 0 a 10. Siendo 0, con toda seguridad no iría a votar y 10, con toda seguridad iría a votar), esta será manipulada como si fuese una variable cuantitativa para obtener más información estadística. Se trata de una práctica habitual en ciencias sociales que debe justificarse en este caso porque se observa se puede obtener más información coherente.

Obviamente al tratarse de análisis univariado solo puede ser descriptivo.

Como en la anterior actividad se incorpora el código del **lenguaje R** en los documentos renderizados (HTML y PDF) activando la opción `echo=TRUE` de los chunks. En algunos chunks no interesa la impresión del resultado por lo que se configuran con `results='hide'`. Si se desea ver el resultado de la ejecución correcta bastará con eliminar esta parametrización en el fichero Rmd descargado (ver cómo descargar en el punto 7).

```
# INSTALAR (si no están) PAQUETES de LIBRERÍAS R NECESARIAS
suppressMessages({ # Se ocultan mensajes instalación
  if (!require(epiDisplay)){install.packages("epiDisplay")}
  if (!require(e1071)){install.packages("e1071")}
  if (!require(stats)){install.packages("stats")}
  if (!require(tibble)){install.packages("tibble")}
  if (!require(knitr)){install.packages("knitr")}
  if (!require(haven)){install.packages("haven")}
})
```

2 OBTENCIÓN DE DATOS

2.1 Fichero y Datos en Abierto

```
# El fichero .sav debe estar en una subcarpeta DATOS
nombreFichero<-"datosjulio2023.sav"
directorioTrabajo<-getwd()
rutaFichero<-paste(directorioTrabajo,"DATOS",nombreFichero,sep="/")
```

2.2 Creación de la base de datos (*data frame*)

A partir del fichero de tipo **datosjulio2023.sav** descargado desde el portal del CIS, se creará el *data frame* de R. Esta será la base de datos inicial con todos los datos de la encuesta.

```
df_cisjulio<-read_sav(rutaFichero)
```

```
## Invalid date string (length=9): 25 038 23
```

2.3 Consultar el cuestionario y el diccionario de datos

Al margen de consultar el cuestionario del CIS, mediante funciones de R se puede observar la variable para intentar deducir de qué tipo se trata.

Algunas consultas posibles podrían ser:

1. Consultar algunos casos y ver sus valores y etiquetas con *head()* o *tail()*
2. Agrupar todos los valores con la función *unique*
3. Usar la función *table* que devuelve el número de casos de cada valor.

```
# 1. Muestra los 5 últimos casos. Los valores y etiquetas
# tail(df_cisjulio$PROBVOTO,5)
# 2. Obtiene los valores sin duplicados
# unique(df_cisjulio$PROBVOTO)
# 3. table(df_cisjulio$PROBVOTO) # Nos da el total de casos que tiene
# cada valor

# Uso de formato RMarkdown
```

```
kable(table(df_cisjulio$PROBVOTO),column.width=c("30%","70%"),
      col.names = c("Valor","Repeticiones"),caption="Repeticiones de cada valor")
```

Tabla 1: Repeticiones de cada valor

Valor	Repeticiones
0	249
1	35
2	31
3	32
4	23
5	183
6	75
7	117
8	303
9	574
10	7145
98	16
99	15

3 PREPARACIÓN DE LA MATRIZ DE DATOS

Una vez obtenida la matriz inicial de datos, se deberá limpiar y simplificar esta base de datos inicial.

3.1 Limpieza de la base de datos.

El data frame inicial contiene 141 columnas correspondientes a todas las preguntas de la encuesta. Se puede prescindir del resto de columnas del data frame innecesarias para nuestro análisis univariado.

```
df_cisProbVoto<-df_cisjulio [, "PROBVOTO"]  
# ncol(df_cisjulio) # Para ver el número de columnas
```

Se ha reducido de 141 columnas a 1 sola.

3.2 Recodificación

En este caso hay pocos valores válidos como es habitual en una variable cualitativa (11 valores que son [0..10]), no se va a agrupar valores pero si se debe decidir qué valores no son válidos en la estadística y excluirlos.

Se duplica la columna

Antes de modificar cualquier campo, se debe hacer un duplicado de este (una nueva columna en el data frame) y trabajar sobre él.

```
# Duplicación de columna  
df_cisProbVoto$recPROBVOTO<-df_cisProbVoto$PROBVOTO
```

Valores válidos y valores no válidos

Valores no válidos:

- 98 N.S.
- 99 N.C.

Valores válidos:

- 0,1,2,3,4,5,6,7,8,9,10

Los valores no válidos tienen un doble tratamiento:

- Valorar su importancia
- Excluirlos de los cálculos

Importancia de los valores perdidos Para valorar la incidencia o relevancia de los valores no válidos se puede ver de qué porcentaje de casos sobre el total de la muestra se trata.

Existen 31 de un total de 8798. Un porcentaje inferior a un 0.4 % de los casos y, por tanto, insignificante.

Asignar NA para excluir en el análisis R permite excluir en sus funciones estadísticas los valores perdidos con parametrizaciones al estilo *MISSING=TRUE* o *na.rm = TRUE*. Para ello, previamente se tiene que asignar el valor NA a estos campos.

```
# Los valores 98 y 99 (NS, NC ) pasan a NA
df_cisProbVoto$recPROBVOTO[df_cisProbVoto$recPROBVOTO>=98]<-NA
```

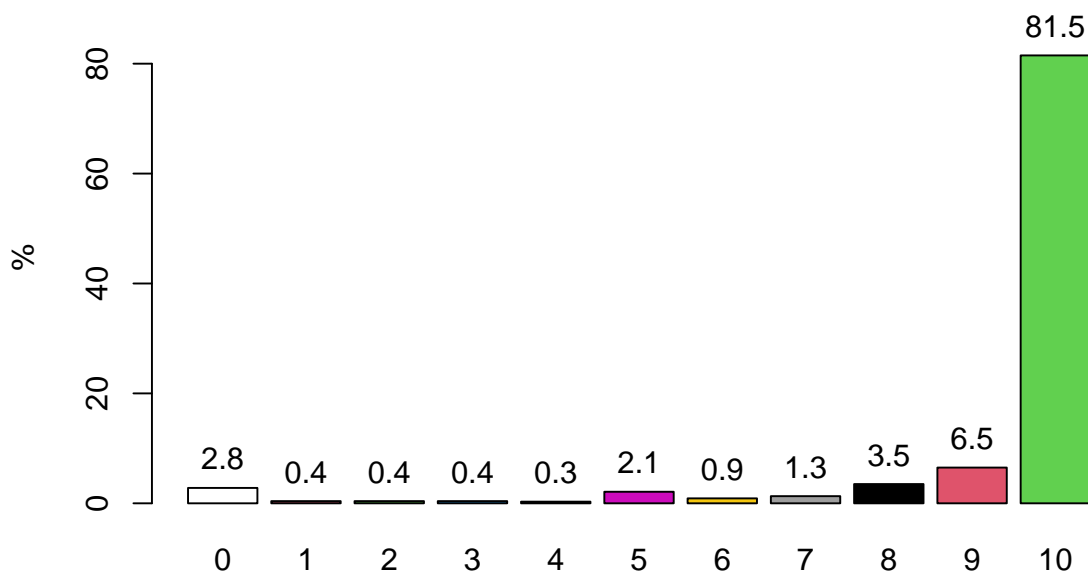
4. DISTRIBUCIÓN DE FRECUENCIAS

En este apartado se verá el uso de tablas de distribución de frecuencias y gráficos. Ambos nos dan una idea de la frecuencia con que se repiten los diferentes valores.

4.1 Gráfico de barras

```
tablaGrafico <- tab1(df_cisProbVoto$recPROBVOTO, cum.percent = TRUE,  
                     bar.values = "percent", missing = FALSE,  
                     xlab="% probabilidad", decimal=1,  
                     main="Gráfico 1. Probabilidad de votar")
```

Gráfico 1. Probabilidad de votar



4.2 Tabla de frecuencias

En la tabla de frecuencias siguiente se muestra el total casos de la muestra por cada valor, la frecuencia en que aparecen y la frecuencia acumulada. Las frecuencias puede calcularse respecto al total de casos o respecto al total de casos válidos. En este caso, como ya se ha apuntado anteriormente, la diferencia es insignificante.

```
# Mediante el atributo output.table se exporta una tabla
tabla<-tablaGrafico$output.table
#Para mejor la comprensión se puede cambiar los nombres de las columnas e
# imprimir la tabla con Rmd
colnames(tabla)<-c("Frecuencia","%", "%_acumulado", "%_válido",
                  "%_acumulado válido")

kable(tabla,column.width = c("10%", "20%", "20%", "20%", "20%"),
      caption="Tabla de distribución de frecuencias")
```

Tabla 2: Tabla de distribución de frecuencias

	Frecuencia	%	%_acumulado	%_válido	%_acumulado válido
0	249	2.8	2.8	2.8	2.8
1	35	0.4	3.2	0.4	3.2
2	31	0.4	3.6	0.4	3.6
3	32	0.4	3.9	0.4	4.0
4	23	0.3	4.2	0.3	4.2
5	183	2.1	6.3	2.1	6.3
6	75	0.9	7.1	0.9	7.2
7	117	1.3	8.5	1.3	8.5
8	303	3.4	11.9	3.5	12.0
9	574	6.5	18.4	6.5	18.5
10	7145	81.2	99.6	81.5	100.0
NA	31	0.4	100.0	0.0	100.0
Total	8798	100.0	100.0	100.0	100.0

```
# print(tabla) # En R estrictamente (no Rmd)

# Para calcular solo las frecuencias de cada valor directamente:
# round(prop.table(table(df_cisProbVoto$recPROBVOTO, useNA="no"))*100,1)
```


5 TRATAMIENTO COMO VARIABLE CUANTITATIVA

Aunque la “probabilidad de votar” en este cuestionario se trata de una variable *cualitativa ordinal*, se podría asumir que los valores válidos [0..10] representan una magnitud que admite operaciones aritméticas y darle un tratamiento de *variable cuantitativa*. En este contexto podría entenderse como de variable *razón* si se asume el valor 0 como la ausencia absoluta de intención de ir a votar (cero absoluto). No obstante se trata de una interpretación que no debe sacarse del contexto de este análisis concreto.

Este planteamiento, excepcional, se justifica por la necesidad de obtener los valores de las *medidas de tendencia central* más allá de la *moda*.

5.1 Medidas de tendencia central

Función *summary*

Esta función nos aporta las medidas de posición central (moda, media y mediana) y las de posición no central (valores extremos y cuartiles). El segundo cuartil equivale a la mediana.

```
summary(df_cisProbVoto$recPROBVOTO,digits=2) # 2 dígitos significativos
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
##	0.0	10.0	10.0	9.3	10.0	10.0	31

Media, Mediana y Moda

Estas medidas de posición central se pueden obtener mediante funciones de R con posibilidad de parametrizar su cálculo (borrar NA, número de decimales...) y evidentemente guardar su valor de retorno en una variable para su posterior uso.

- La opción *na.rm = TRUE* indica que no se debe considerar los valores no válidos
- El valor obtenido se puede redondear con la función *round*. En ciencias sociales lo habitual es 1 decimal

Media Promedio aritmético de todos los valores de la muestra.

```
media<-round(mean(df_cisProbVoto$recPROBVOTO, na.rm = TRUE),1)
```

Mediana Valor de la muestra que divide la muestra en dos mitades o grupos con igual cantidad de casos, siendo uno de los grupos el formado por casos con un valor superior y el otro grupo el formado por los casos con valor inferior.

```
mediana<-median(df_cisProbVoto$recPROBVOTO, na.rm = TRUE,1)
```

Moda La moda es el valor de la muestra que más se repite. Como ya se vio en la actividad anterior de esta práctica, en R no existe una función que devuelva la moda directamente similar a las vistas para mediana *median()* y para la media *mean()*.

Se debe usar una diseñada expresamente que, una vez definida, se puede llamar de igual manera que a las librerías predefinidas de R e incorporadas con la instalación de los paquetes.

Función moda

```
#declaración y desarrollo de la función de usuario
moda<- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```
#Se llama a la función previamente desarrollada
moda<-moda(df_cisProbVoto$recPROBVOTO)
```

Resumen de valores de posición central

- La Media es 9.3
- La Mediana es 10
- La Moda es 10

Conclusiones

La *media aritmética*: 9.3 está muy influida por la altísima frecuencia de los valores más altos, sobretodo del máximo (10). Pese a que la *Mediana* nos de como resultado 10, el total de casos con valor inferior no llega a ser el 50%.

La moda sin duda es el valor 10 como ya se podía deducir a partir del porcentaje reflejado en la tabla de frecuencia.

5.2 Comparación entre Moda, Media y Mediana.

Histograma

El histograma mostrará una asimetría positiva (sesgo a la izquierda) donde la media y a moda coinciden y la media es el valor más bajo

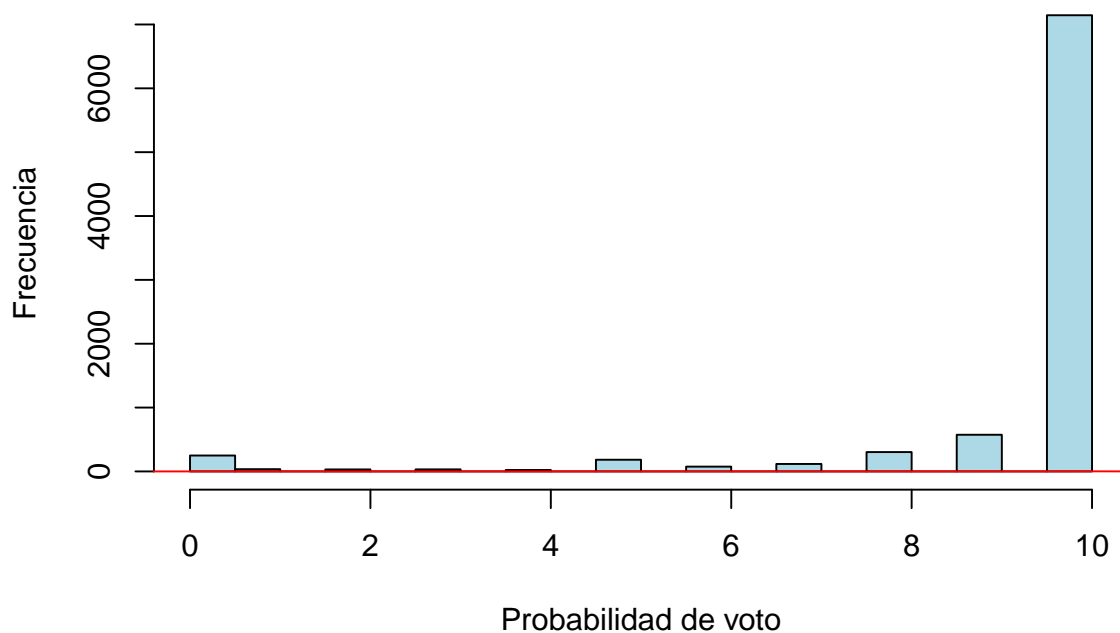
$X < Me = Mo$

```
# Eliminar filas con NA
df_cisProbVotoSinNA <- df_cisProbVoto[complete.cases(
  df_cisProbVoto$recPROBVOTO), ]

hist(df_cisProbVoto$recPROBVOTO, col = "lightblue",
      xlab = "Probabilidad de voto",
      ylab = "Frecuencia",
      main = "Gráfico 2. Histograma de la probabilidad de votar")

lines(density(df_cisProbVotoSinNA$recPROBVOTO), col = "red", lwd = 1)
```

Gráfico 2. Histograma de la probabilidad de votar



Conclusión sobre las medidas de posición central

Viendo el sesgo (a la derecha) se entiende que la mediana tenga un valor mayor que la moda y menor que la Media. La mediana debería ser una medida de posición central mejor que la media ya que esta quedaría distorsionada por los extremos. La moda viene definida por la repetición de un valor. En nuestro caso, en cambio, la mediana coincide con la Moda.

5.3 Medidas de posición no central

Valores extremos: Máximo y Mínimo

```
# Se guardan en variables los retorno de las funciones predefinidas de R  
# así se pueden leer desde le MD  
maximo<-max(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)  
minimo<-min(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

Los valores extremos son:

- El valor MÁXIMO es: 10
- El valor MÍNIMO es: 0

Al tratarse de una pregunta cerrada con una ordenación entre las respuestas (variable cualitativa ordinal), estos valores indican el par de respuestas más dispares entre si de nuestra muestra. La coincidencia de máximo y mínimo con las opciones más extremas posibles significa simplemente que en nuestra muestra ha habido, al menos, un caso (persona encuestada) que ha elegido un extremo.

Cuartiles y percentiles

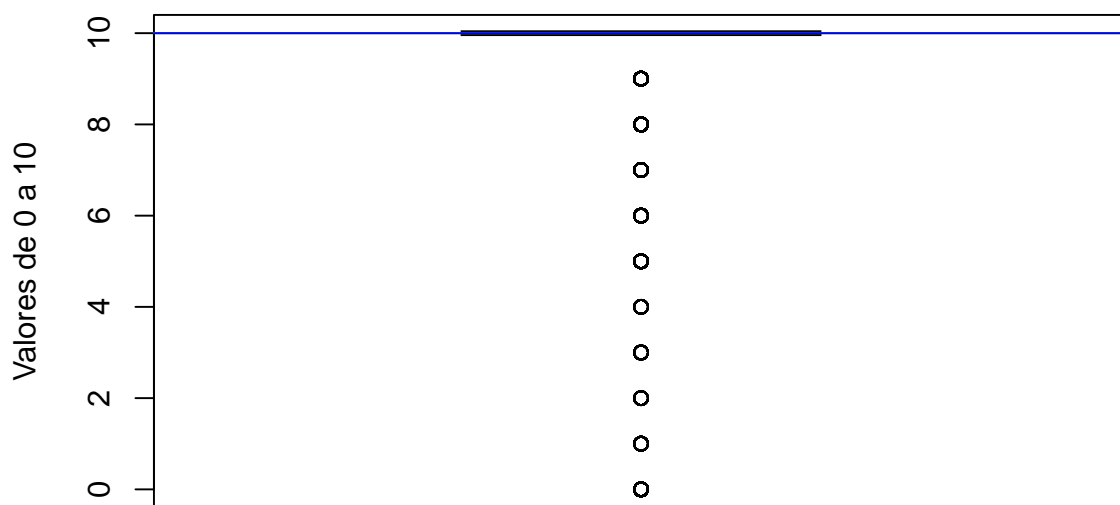
En una distribución de valores de una muestra tan sesgada, un cuartil no da demasiada información. La inmensa mayoría de valores son 10 y el resto son tratados como valores atípicos (1,5 veces la caja).

```
# Crear el diagrama de caja  
boxplot(df_cisProbVotoSinNA$recPROBVOTO,  
        main = "Gráfico 3. Cuartiles probabilidad de votar",  
        ylab = "Valores de 0 a 10",  
        col = "lightblue")
```

```
# Calcular los cuartiles
cuartiles <- quantile(df_cisProbVotoSinNA$recPROBVOTO,
                     probs = c(0.25, 0.5, 0.75))

# Agregar las líneas de los los cuartiles
abline(h = cuartiles, col = c("red", "green", "blue"))
```

Gráfico 3. Cuartiles probabilidad de votar



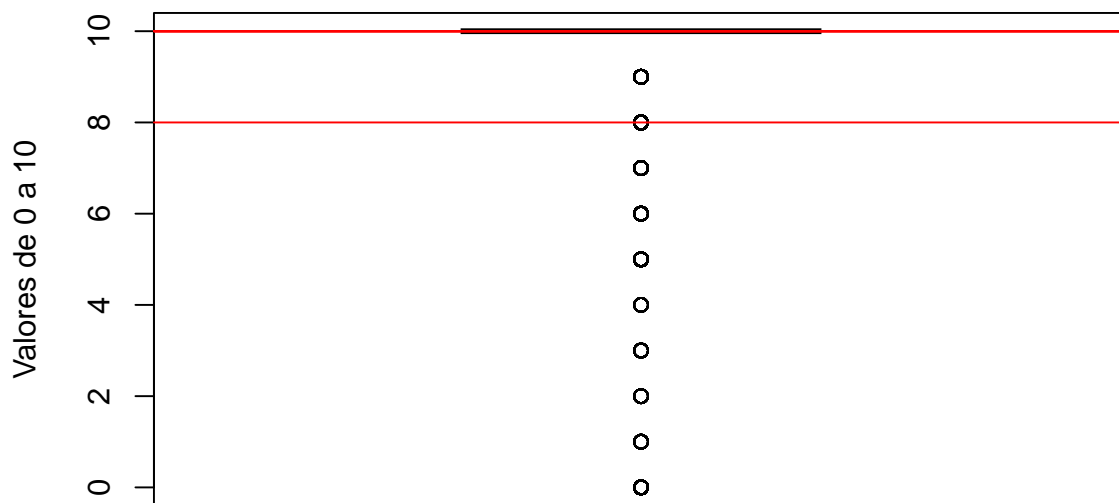
Se puede usar los percentiles para analizar mejor la información.

```
percentiles<-quantile(df_cisProbVotoSinNA, c(.10,.20,.30,.40,.50,.60,.70,
                                             .80,.90), na.rm = TRUE)

boxplot(df_cisProbVotoSinNA$recPROBVOTO,
        main = "Gráfico 4. Percentiles probabilidad de votar",
        ylab = "Valores de 0 a 10", col = "lightblue")

# Agregar las líneas de los los cuartiles
abline(h = percentiles, col = "red")
```

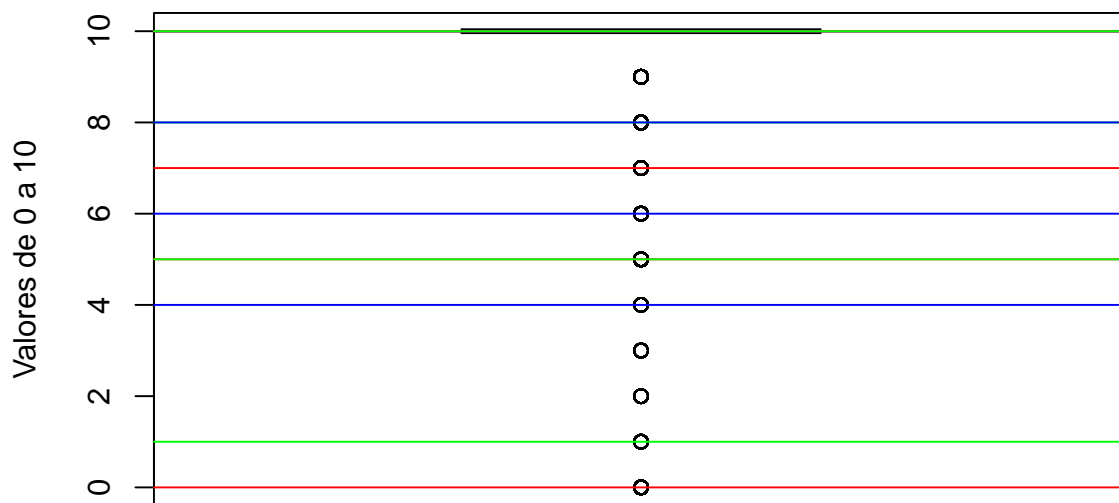
Gráfico 4. Percentiles probabilidad de votar



```
percentiles<-quantile(df_cisProbVotoSinNA, c(.02,.03,.04,.05,.06,.07,.08,
      .09,.10,.20,.30,.40,.50,.60,.70,.80,.90), na.rm = TRUE)
boxplot(df_cisProbVotoSinNA$recPROBVOTO,
      main = "Gráfico 5. Percentiles probabilidad de votar",
      ylab = "Valores de 0 a 10", col = "lightblue")

# Agregar las líneas de los los cuartiles
abline(h = percentiles, col = c("red", "green", "blue","red", "green",
      "blue","red", "green","blue","red",
      "green", "blue","red", "green",
      "blue","red", "green", "blue"))
```

Gráfico 5. Percentiles probabilidad de votar



Conclusión:

Solo un porcentaje inferior al 9% de la muestra ha respondido un valor inferior al '8'. Un hecho que se puede apreciar también en la columna de *frecuencias acumuladas* de la Tabla de frecuencias donde el valor 8 acumula el 8,5%.

5.4 Medidas de dispersión

Se procederá a obtener las medidas de variabilidad o dispersión que informan del comportamiento de la variable y complementan la información sobre la centralidad.

Varianza o desviación típica

```
# na.rm = TRUE. Si en el data frame existen valores perdidos.  
varianza<-var(df_cisProbVoto$recPROBVOTO,na.rm = TRUE)
```

La desviación típica es: 4.2

Desviación Estándar

```
# na.rm = TRUE. Si en el data frame existen valores perdidos.  
desviacionEstandar<-sd(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

La desviación estándar es: 2

Rango intercuatílico

```
# na.rm = TRUE. Si en el data frame existen valores perdidos.  
iqr<-IQR(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

El rango intercuartílico es 0

Curtosis

El coeficiente de curtosis indica el nivel de apuntamiento achatamiento que presenta una distribución de valores. Solo tiene sentidos en distribuciones unimodales y simétricas. No es este nuestro caso.

```
# na.rm = TRUE. Si en el data frame existen valores perdidos.  
coeficienteKurtosis<-kurtosis(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

El coeficiente de curtosis es: 11.9497589

Coeficiente de Asimetría

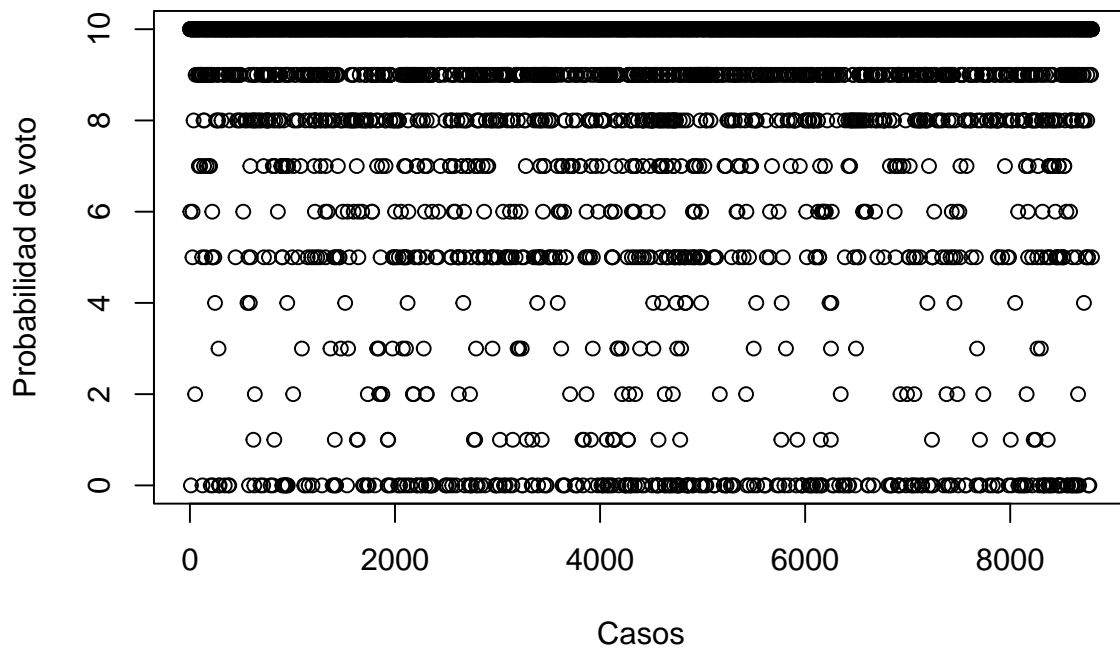
```
coeficienteAsimetria<-skewness(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

El coeficiente de asimetría es: -3.5101093

Grafico de densidad

```
plot(df_cisProbVoto$recPROBVOTO,main="Gráfico 6. Gráfico de densidad"  
      ,xlab="Casos", ylab = "Probabilidad de voto")
```


Gráfico 6. Gráfico de densidad



6 RESUMEN VALORES







Tabla 3: Resumen de valores de las mediciones

Medición	Valor
POSICIÓN CENTRAL (Medidas de tendencia central)	
MODA	10
MEDIANA	10
MEDIA	9.3
POSICIÓN (Medidas de posición no central)	
MÁXIMO	10
MÍNIMO	0
PERCENTIL 25	10
PERCENTIL 50	10
PERCENTIL 75	10
MEDIDAS DE DISPERSIÓN	
RANGO	10
RANGO INTERCUARTIL	0
DESVIACIÓN TÍPICA	4.2
DESVIACIÓN TÍPICA	2
MEDIDAS DE FORMA	
COEFICIENTE DE ASIMETRÍA	-3.5
COEFICIENTE DE KURTOSIS	12

7 EJECUCIÓN DEL Rmd Y REPOSITORIO DE FICHEROS

Para la ejecución del código R, está disponible el fichero ejecutable Rmd y el fichero de datos *sav* en el repositorio GitHub. En el mismo repositorio se encuentra este PDF y un fichero HTML renderizados a partir del Rmd.

El fichero **datosjulio2023.sav** debe ubicarse en una subcarpeta del directorio de trabajo con nombre *DATA*.

	Enlace a GitHub	
	Fichero de datos tipo sav	datosjulio2023.sav
	Fichero Rmd	probabilidadVoto.Rmd
	Fichero HTML	probabilidadVoto.html
	Fichero PDF	probabilidadVoto.pdf
	Repositorio tofermos	