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

Técnicas de Investigación en Ciencias Políticas-UBU.
Práctica 1. Apartado 4.

Tomàs Ferrandis Moscardó

2023-03-22

Contents

1 INTRODUCCIÓN

Este segundo estudio univariao se centrará en la variable PROBVOTO (pregunta P1) del mismo cuestionario: *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"* por lo tanto se trata también de un *análisi univariado*.

Podemos ver que, pese a ser un variable cualitativa ordinal, esta será manipulada como si fuese una cuantitativa para obtener más información estadística. Se trata de una práctica habitual en ciencias sociales que debe, en todo caso, quedar justificada.

> NOTA: En este documento se ha activado la opción *echo* de los chuncks de R a fin de poder ser valorados académicamente. La creación del fichero HTML, PDF o DOCX renderizado se haría tras la desactivación. Solo llevan título de tablas de resultados que > corresponderían a esta versión final y están maquetadas en Markdown.

```
# Para las librerías necesarias, se deben instalar los paquetes si no están ya instalados.
# Ocultamos los mensajes.
suppressMessages({
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 lo tenemos 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 *sav* se creará el *data frame* de R. En será la base de datos incial donde tendremos todos los datos del barómetro en concreto del Estudio 3415.

```
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 R se puede observar la variable para intentar deducir de qué tipo se trata.

Algunas consultas posibles podrían ser:

- Consultar algunos casos y los valores con etiquetas (en este caso) con *head()* o *tail()*
- Usar *table* que devuelve el número de casos de cada valor.
- Agrupar todos los valores con *unique*

```r
tail(df_cisjulio$PROBVOTO,5) # vemos solo 5 casos y los valores y etiquetas
```

```
## <labelled<double>[5]>: Escala de probabilidad de ir a votar (0-10) en las elecciones generales de 20
## [1] 10 10 10 10  5
##
## Labels:
##  value                                label
##      0 0 Con toda seguridad no iría a votar
##      1                                    1
##      2                                    2
##      3                                    3
##      4                                    4
##      5                                    5
##      6                                    6
##      7                                    7
##      8                                    8
##      9                                    9
##     10  10 Con toda seguridad, iría a votar
##     98                                 N.S.
##     99                                 N.C.
```

```r
                              # de la variable puesto que es factor
table(df_cisjulio$PROBVOTO) # Nos da el total de casos que tiene cada valor
```

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

```r
unique(df_cisjulio$PROBVOTO) # Obtiene los valores sin duplicados
```

```
## <labelled<double>[13]>: Escala de probabilidad de ir a votar (0-10) en las elecciones generales de 2
##  [1] 10  6  0 99  5  8  2  9  7  4  3 98  1
##
## Labels:
##  value                                label
##      0 0 Con toda seguridad no iría a votar
##      1                                    1
##      2                                    2
##      3                                    3
##      4                                    4
##      5                                    5
##      6                                    6
##      7                                    7
##      8                                    8
##      9                                    9
##     10  10 Con toda seguridad, iría a votar
##     98                                 N.S.
##     99                                 N.C.
```

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 rncol(df_cisjulio)‘ 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) # devuelve el número de columnas del data frame. Usado en Rmd
```

Hemos pasado de rncol(df_cisjulio)‘ a 1 sola columna.

3.2 Recodificción

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

Duplicamos la columna

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

```
df_cisProbVoto$recPROBVOTO<-df_cisProbVoto$PROBVOTO
```

Valores válidos y valores no válidos

Valores no válidos: * 98 N.S. * 99 N.C. Valores válidos: 0..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 de qué porcentaje de casos de la muestra se trata.

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

Obtenemos estos valores con

```
#r nrow(df_cisProbVoto[df_cisProbVoto$recPROBVOTO>=98,])
```

```
#r nrow(df_cisProbVoto)
```

```
#r nrow(df_cisProbVoto[df_cisProbVoto$recPROBVOTO>=98,])/nrow(df_cisProbVoto)*100
```

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.

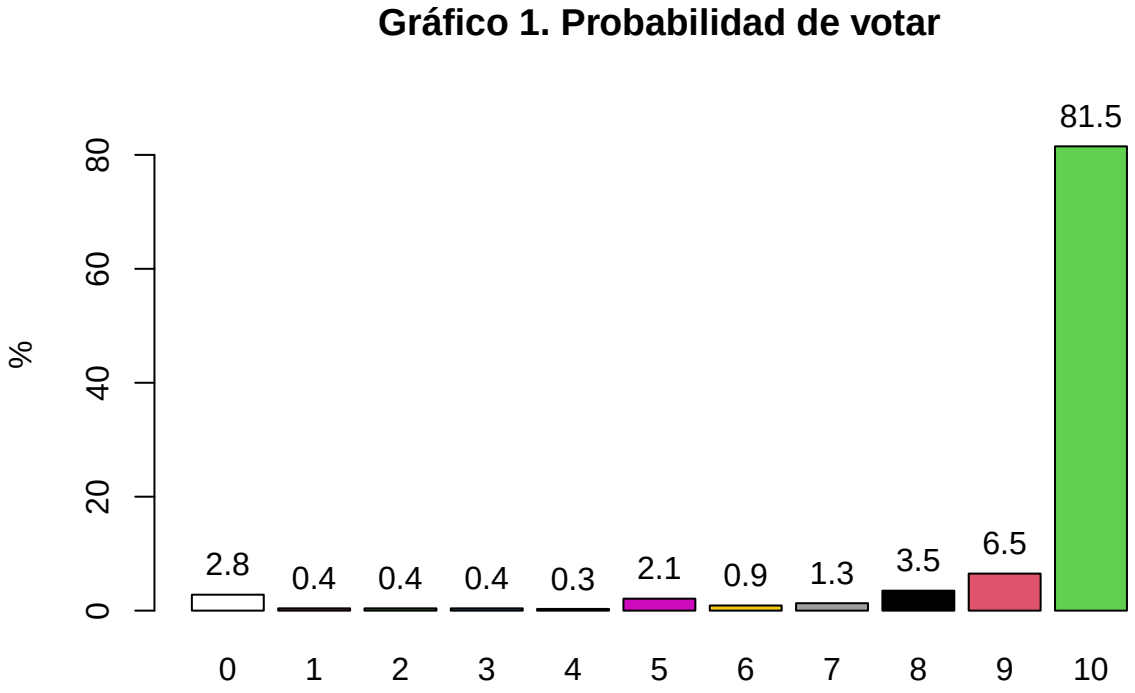
```
df_cisProbVoto$recPROBVOTO[df_cisProbVoto$recPROBVOTO>=98]<-NA
```

4. DISTRIBUCIÓN DE FRECUENCIAS

En esta apartado se verá el uso de tablas de distribución de frecuencias y gráficos, métodos fundamnetales que nos muestran as constituyen una primera forma de dar cuenta de la información sobre las variables.

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")
```



4.2 Tabla de frecuencias

En la tabla de frecuencias siguiente se muestra el total casos por cada valor, la frecuencia en que aparecen y la frecuencia acumulada. Las frecuencias puede calcularse respecto al totas 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 mejorar la comprensión se puede cambiar los nombres de las columnas e imprimir con MD
colnames(tabla)<-c("Frecuencia","%","%_acumulado", "%_válido","%_acumulado válido")
knitr::kable(tabla,column.width=c("10%","20%","20%","20%","20%"),caption="Tabla1.
             Tabla de distribución de frecuencias")
```

Table 1: Tabla1. 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

	Frecuencia	%	%_acumulado	%_válido	%_acumulado válido
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

# Para calcular las frecuencias 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 artiméticas 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. No obstante se trata de una interpretación que no debe sacarse del contexto del 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)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000  10.000  10.000   9.297  10.000  10.000      31
```

Media, Mediana y Moda

Estas medidas de posición central se pueden obtener mediante funciones específicas con posibilidad de especificar opciones en su cálculo (borrar NA, número de decimales…) y guardar su valor en una variable para su posterior uso.

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 casoa con un valor superior y el otro grupo el formado por los casos con valor inferior.

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

La opción *na.rm = TRUE* indica no se debe considerar los valores no validos: remove NA Redondeamos con la función *round*, como es habitual en ciencias sociales a 1 decimal

Moda La moda es el valor de la muestra que más se repite. > 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.

Función moda

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

Una vez definida la función se puede llamar como las incorporadas en las librerías.

```r
#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 la altísima frecuencia de los valores más altos, sobretodo del máximo. Pese que la *mediana* nos de como resultado 10, el total de casos con valor inferior no llega a ser el 50%.

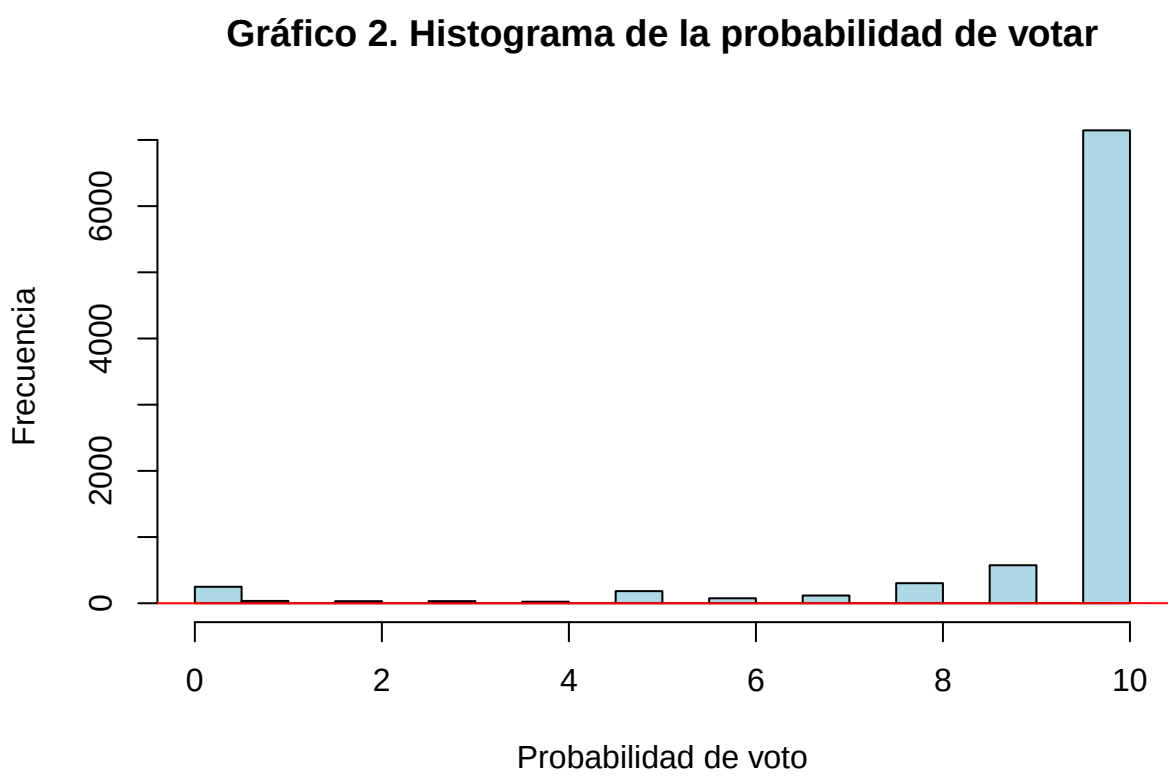
La moda sin duda es el valor 10 a la vista del porcentaje reflejado en la tabla de frecuencia.

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**

```r
# 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)
```



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 Mediada 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

```
maximo<-max(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)

minimo<-min(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

Los valores extremos son:

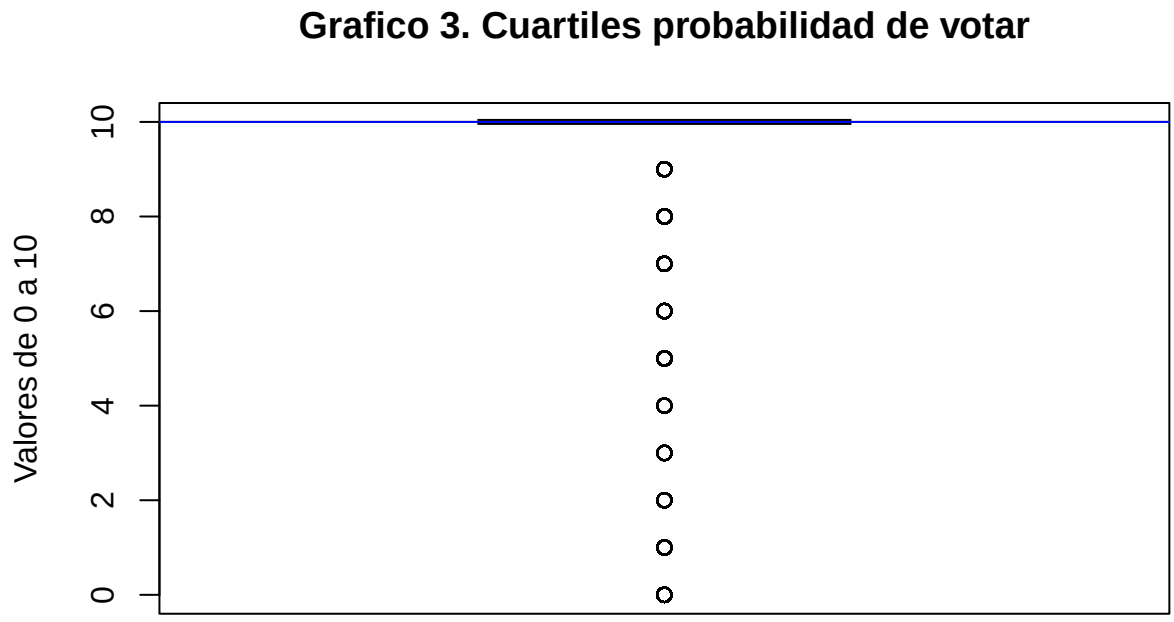
- El valor MÁXIMO es: 10
- El valor MÍNIMO es: 0

Al tratarse de una pregunta cerrada con una ordenación entre las respuestas (variable cualitativa ordinal), estos valores indican el par de respuestas más dispares. La coincidencia de máximo y mínimo con las opciones más extremas implica que ha habido casos donde el encuestado ha elegido estos valores, simplemente.

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).

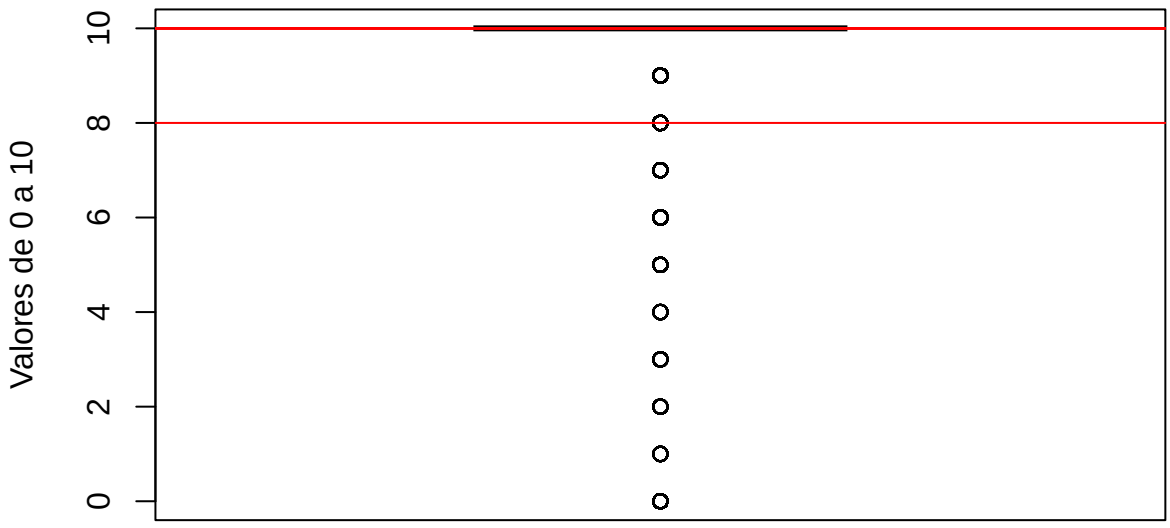
```r
# Crear el diagrama de caja
boxplot(df_cisProbVotoSinNA$recPROBVOTO,
        main = "Grafico 3. Cuartiles probabilidad de votar",
        ylab = "Valores de 0 a 10",
        col = "lightblue")

# 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"))
```



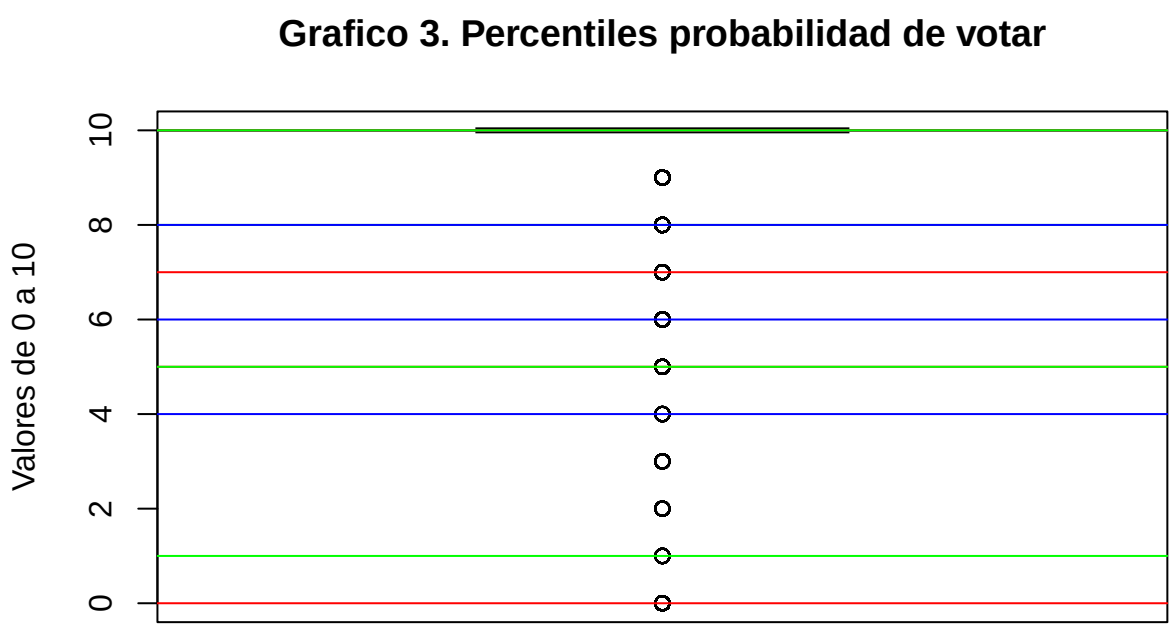
Se puede usar los percentiles para analizar mejor la información.

```r
percentiles<-quantile(df_cisProbVotoSinNA, c(.10,.20,.30,.40,.50,.60,.70,.80,.90), na.rm = TRUE)
boxplot(df_cisProbVotoSinNA$recPROBVOTO, main = "Grafico 3. Percentiles probabilidad de votar",
        ylab = "Valores de 0 a 10", col = "lightblue")
# Agregar las líneas de los los cuartiles
abline(h = percentiles, col = "red")
```

Grafico 3. 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 = "Grafico 3. 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"))
```



Conclusión: SOlo un porcentaje inferior al 9% de la muestra ha respuesto un valor inferior al '8'. Algo que se puede ver también en la columna de frecuencias acumuladas (Tabla de frecuencia) donde el valor 8 acumula el 8,5%.

5.4 Medidas de dispersión

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

Varianza o desviación típica

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

La desviación típica es: 4.2

Desviación Estándar

```
desviacionEstandar<-sd(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

La desviación estándar es: 2

Rango intercuatílico

```
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 tien sentidos en distribuciones unimodales y simétricas. No es este nuestro caso.

```
coeficienteKurtosis<-kurtosis(df_cisProbVoto$recPROBVOTO, na.rm = TRUE)
```

El coeficiente de curtosis es: 11.9497589

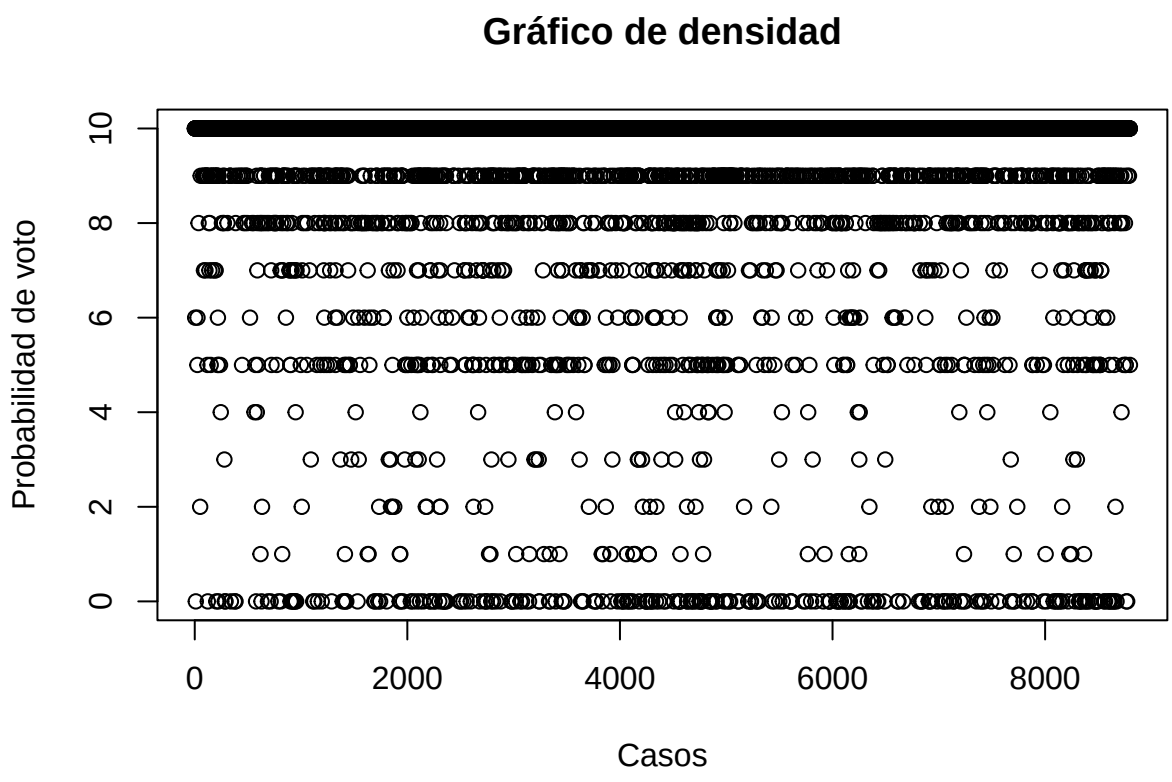
Coeficiente de Asimetría

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

El coeficiente de asimetría es: -3.5101093

Grafico de densidad

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



6 RESUMEN VALORES

MEDICIÓN	
POSICION 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