

ANÁLISIS UNIVARIABLE SOBRE PREFERENCIA DE PRESIDENTE/A DE GOBIERNO SEGÚN BARÓMETRO DEL CIS DE JULIO 2023 (ESTUDIO 3415)

Técnicas de Investigación en Ciencia Política I. UBU
Práctica 1. Apartado 3

Tomàs Ferrandis Moscardó

2023-03-18

Índice

1. INTRODUCCIÓN	3
2. OBTENCIÓN DE DATOS	3
2.1 Fichero y Datos en Abierto	4
2.2 Creación de la base de datos (<i>data frame</i>)	4
2.3 Consultar el cuestionario y el diccionario de datos	4
2.4 Tipo de variable PREFPTE	4
3. PREPARACIÓN DE LA MATRIZ DE DATOS	6
3.1 Limpieza de la base de datos	6
3.2 Recodificación	7
4 MEDIDAS DE TENDENCIA CENTRAL	10
4.1 Moda	10
5. MEDIDAS DE DISPERSIÓN CENTRAL	11
5.1 Tabla de frecuencias de Preferencia de presidente/a	11
6 Gráficos	13
6.1 Gráfico de barras	13
6.2 Gráfico de sectores	15
6.3 Gráfico de Pareto	16
7 GUARDAR RESULTADOS EN FICHERO EXTERNO	17
7.1 Fichero CSV (<i>Comma-Separated Values</i>)	17

7.2 Hoja de cálculo MS Excel (xlsx)	18
8 EJECUCIÓN DEL Rmd Y REPOSITORIO DE FICHEROS	19

1. INTRODUCCIÓN

Esta actividad corresponde al apartado 3º de la Práctica 1 de Técnicas de Investigación en Ciencia Política I. Se trata de un *análisis univariado* sobre la variable PREFPTE del Barómetro del CIS de julio de 2023 (Estudio 3415). Concretamente se trata de la pregunta P8 que dice: *De los/las principales líderes políticos/as, ¿quién preferiría que fuese el/la presidente/a del Gobierno tras las elecciones?* . Obviamente al tratarse de análisis univariado solo puede ser descriptivo.

Para el análisis de los datos se usa el **lenguaje de programación R**. Se incorpora el código en los documentos renderizados (HTML y PDF) activando la opción *echo* de los chunks (*echo=TRUE*). 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 de alguno de estos chunks bastará con eliminar el parametro *results='hide'* en el fichero Rmd descargado (ver punto 8 para descarga).

2. OBTENCIÓN DE DATOS

```
# Instalar paquetes (si no están) con las funciones necesarias
# Supresión de mensajes del proceso (message=FALSE)
if(!require(haven)){install.packages("haven")} # para abrir fichero .sav
if(!require(epiDisplay)){install.packages("epiDisplay")} # para gráfico
if(!require(knitr)){install.packages("knitr")}#tabla dinámica Rmd y otros
if(!require(xlsx)){install.packages("xlsx")} # manejo de ficheros .xlsx
if(!require(qcc)){install.packages("qcc")}# diagrama Pareto
library(qcc)
library(haven)
library(epiDisplay)
library(knitr)
library(xlsx)
```

2.1 Fichero y Datos en Abierto

Del portal del CIS, se localiza el *barómetro de julio 2023* y se descargan los datos estadísticos en un formato de fichero *.sav*

Se usará una variable R (*nombreFichero*) con la ruta relativa *DATOS* que será el directorio donde se guardará los ficheros de datos.

```
# El fichero .sav se guardará 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 objeto *data frame* de R. Este *data frame* será la base de datos inicial con todos los datos obtenidos de la muestra mediante la encuesta.

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

2.3 Consultar el cuestionario y el diccionario de datos

Observando el cuestionario del CIS se puede ver qué se trata de una variable de tipo cualitativa. Una comprobación que se puede hacer mediante R es ver que los valores existentes en la base de datos coinciden coherentemente con las respuestas válidas del cuestionario. Mediante la función *unique* devuelve todos valores existentes en la columna sin duplicados, además de un listado de los valores posibles y sus respectiva etiquetas definidos. Puede ser de gran ayuda para esta comprobación.

```
unique(df_cisjulio$PREFPTE) # obtiene los valores sin duplicado
```

2.4 Tipo de variable PREFPTE

A la vista de los resultados anteriores y, tras leer la documentación del cuestionario, se puede deducir que la variable PREFTE es una variable *cualitativa nominal* con escasos valores como es habitual (11) y, donde cada uno de ellos tiene una etiqueta asociada.

Antes de obtener medidas de posición o de variación se deberá tratar previamente la matriz de datos.

3. PREPARACIÓN DE LA MATRIZ DE DATOS

A partir de la matriz obtenida se procederá a limpiar y simplificar esta base de datos inicial.

3.1 Limpieza de la base de datos

Reducción del tamaño de la matriz de datos

Al tratarse de un *análisis univariado* en que solo se va a considerar una columna es conveniente prescindir de los demás datos usando así, un *data frame* de menor tamaño (menos columnas) y más simple.

```
df_cis<-df_cisjulio [, "PREFPTE"]
```

Para comprobar el resultado se puede usar instrucciones de lectura del data frame de R como *head* o *tail*.

```
tail(df_cis,4) # Muestra los 4 últimos casos  
head(df_cis,4) # Muestra los 4 primeros casos
```

Otra comprobación interesante consiste en comparar las dimensiones de ambos *data frames*, el inicial y el resumido, con funciones de R como:

- *dim()* Devuelve un vector con dos valores: el número total de casos (líneas del *data frame*) y el número de variables (columnas del *data frame*).
- *nrow()* Devuelve el número total de casos (líneas del *data frame*)
- *ncols()* el número de variables (columnas del *data frame*)

```
dim(df_cis)  
dim(df_cisjulio)  
nrow(df_cis)  
nrow(df_cisjulio)  
ncol(df_cisjulio)  
  
n<-nrow(df_cis) # La n del muestreo
```

Lógicamente el número de casos es 8798 en ambos *data frames*.

8798 es el valor **n** del muestreo

3.2 Recodificación

Duplicación de la variable

Lo más prudente cuando se va a editar datos es que estos se realicen sobre una copia. Por esta razón se duplica la columna del *data frame*

```
df_cis$recPREFPTE<-df_cis$PREFPTE
```

Valores válidos y no válidos

Una vez examinados los datos en el apartado 3.2 anterior, ya se puede decidir qué valores de la variable son válidos y cuales no interesan para este análisis.

Valores válidos: niveles y etiquetas

- Pedro Sánchez 1
- Alberto Núñez Feijóo 2
- Santiago Abascal 3
- Yolanda Díaz 4
- Ione Belarra 10
- Íñigo Errejón 7
- Isabel Díaz Ayuso 8

Valores no válidos: niveles y etiquetas

- (O LEER) Ninguno/a de ellos/as 97
- (NO LEER) Otro/a 96
- N.S. 98
- N.C. 99

Se deberá agrupar bajo la etiqueta de NA (valor ausente) los valores 96, 97, 98 y 99. Para ello, antes se duplicará esta columna, es decir, se insertará dentro del *data frame* una segunda columna con los mismos valores. Sobre esta columna se editarán los datos (**recodificación**).

Agrupación de valores NO válidos

Todos los valores que no son válidos para el análisis (96, 97, 98, 99 se actualizarán a NA.

Al mismo tiempo se puede simplificar el resto de valores en este caso, solo cambiándolos a [1,2,3,4,5,6 7].

```
df_cis0<-df_cis
df_cis$recPREFPTE[df_cis$recPREFPTE >=96]<-NA
df_cis$recPREFPTE<-factor(df_cis$recPREFPTE,
                           levels=c(1,2,3,4,5,6,7),
                           labels=c("Pedro Sánchez", "Alberto Núñez Feijóo",
                                     "Santiago Abascal","Yolanda Díaz","Ione Belarra",
                                     "Íñigo Errejón","Isabel Díaz Ayuso"))
```

El tratamiento de los valores no válidos es doble. Se excluirán del análisis pero se deberá valorar su importancia. Para ello se debe ver su peso en relación al conjunto de valores. En R se dispone de la función *summary* que devuelve los valores de posición central (media, mediana y moda) y los cuartiles. Se puede parametrizar para que cuente también los valores NA (remove missing values = FALSE)

```
numero_na<-sum(is.na(df_cis$recPREFPTE)) # Guarda valor en variable para...
porcentaje_na<-round(mean(is.na(df_cis$recPREFPTE))*100,1) #calcular % NA

# Totaliza por valores, incluyendo NA)
kable(summary(df_cis$recPREFPTE, na.rm=FALSE),
       col.names = c("Líder","Repeticiones"),caption="Summary")
```

Tabla 1: Summary

Líder	Repeticiones
Pedro Sánchez	2631
Alberto Núñez Feijóo	2576
Santiago Abascal	658
Yolanda Díaz	1300
Ione Belarra	0
Íñigo Errejón	0
Isabel Díaz Ayuso	8
NA's	1625

Importancia de los valores perdidos NA

Hay un total de 1625 lo que supone un porcentaje del 18.5 de casos no válidos sobre el total de $n=8798$.

4 MEDIDAS DE TENDENCIA CENTRAL

Al tratarse de una variable **cualitativa nominal** solo tiene sentido como medida de tendencia central la *moda*. La *mediana* y la *media*, no.

4.1 Moda

Para el cálculo de la moda, R no dispone de ninguna función en sus librerías, se puede usar una función de usuario como la aportada en los apuntes de la asignatura.

Función Moda

```
moda <- function(v) {  
  unqv <- unique(v)  
  unqv[which.max(tabulate(match(v, unqv)))]  
}
```

Una vez creada la función e incorporado su código en el documento RMarkdown (o script R) se podrá llamar y almacenar su resultado en una variable para consultas u operaciones posteriores.

```
varModa<-moda(df_cis$recPREFPTE) #llama a la función y recupera valor
```

La moda es: *Pedro Sánchez*

5. MEDIDAS DE DISPERSIÓN CENTRAL

Al tratarse de una variable cualitativa nominal no tiene sentido estudiar un comportamiento en la distribución de valores.

5.1 Tabla de frecuencias de Preferencia de presidente/a

En los siguientes análisis de datos se excluirán los valores no válidos.

Frecuencias absolutas

La frecuencia absoluta indica la suma de casos que se deciden por cada uno de los candidatos.

```
# La función *table* de R puede ignorar o no los NA (useNA="no"/"ifany")  
# por defecto. useNA="no" (los ignora, no haría falta indicarlo)  
  
# table(df_cis$recPREFPTE, useNA="no") # En R estrictamente  
  
# Uso de formato RMarkdown  
kable(table(df_cis$recPREFPTE), column.width=c("30%", "70%"),  
       col.names = c("Líder", "Repeticiones"), caption="Frecuencia absoluta")
```

Tabla 2: Frecuencia absoluta

Líder	Repeticiones
Pedro Sánchez	2631
Alberto Núñez Feijóo	2576
Santiago Abascal	658
Yolanda Díaz	1300
Ione Belarra	0
Íñigo Errejón	0
Isabel Díaz Ayuso	8

Frecuencias relativas.

La frecuencia relativa indica el porcentaje respecto al total de casos de los valores del apartado anterior. Es decir, el porcentaje de casos que se prefieren a cada candidato.

```
kable(round(prop.table(table(df_cis$recPREFPTE,useNA="no"))*100,digits=1),
      col.names = c("Líder","% de repeticiones"),caption="Frecuencia relativa")
```

Tabla 3: Frecuencia relativa

Líder	% de repeticiones
Pedro Sánchez	36.7
Alberto Núñez Feijóo	35.9
Santiago Abascal	9.2
Yolanda Díaz	18.1
Ione Belarra	0.0
Íñigo Errejón	0.0
Isabel Díaz Ayuso	0.1

- La función *table* de R devolverá el total de casos de cada valor. *Ejemplo: Pedro Sánchez: 2731*
- La función *prop* devolverá el tanto x 1 del valor anterior. *Ejemplo: Pedro Sánchez: 0,361*
- Se multiplicará x 100 para obtener el porcentaje.
- Con la función aritmética *round* se redondeará a 1 decimal como es habitual en Ciencias Sociales.

6 Gráficos

Para variables nominales, se puede optar por:

- Gráfico de barras
- Gráfico de sectores
- Gráfico de Pareto

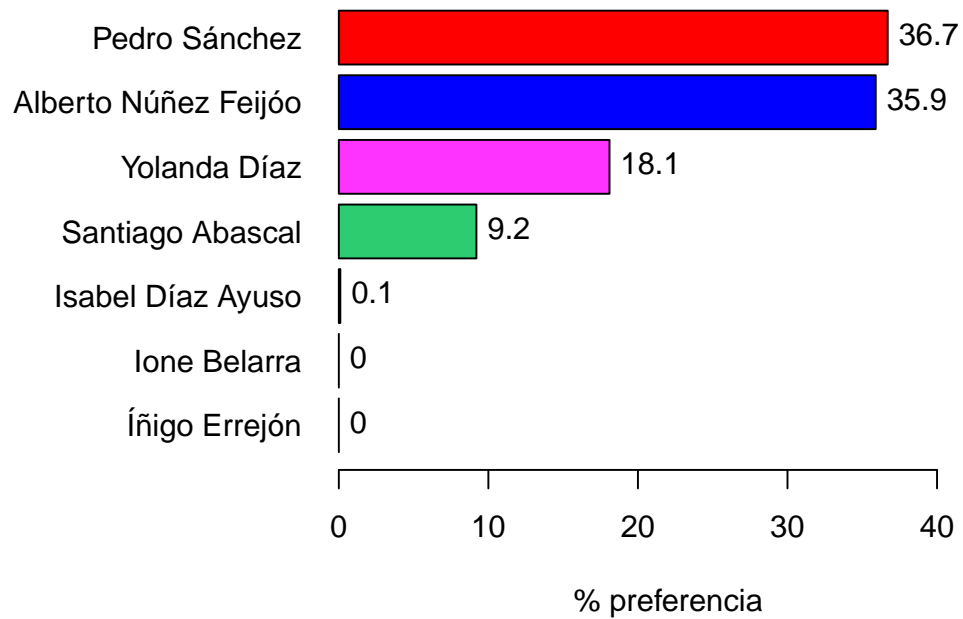
En los siguientes subapartados se verán los resultados en las diferentes opciones.

6.1 Gráfico de barras

Para una variable nominal como es el caso, el gráfico más adecuado es el diagrama de barras.

```
# Se adoptan los colores identificativos de la opción política
colores<-c("red","blue","#FF33FF","#2ecc71","#2980b9","#ab47bc","#c8e6c9")
coloresInvertidos<-rev(colores)
grafico <- tab1(df_cis$recPREFPTE,cum.percent = TRUE,
               sort.group="decreasing",xlab="% preferencia",
               decimal=1,bar.values = "percent",
               main="Gráfico 1. Preferencia Presidencia de Gobierno",
               col=coloresInvertidos, missing = FALSE)
```

Gráfico 1. Preferencia Presidencia de Gobierno

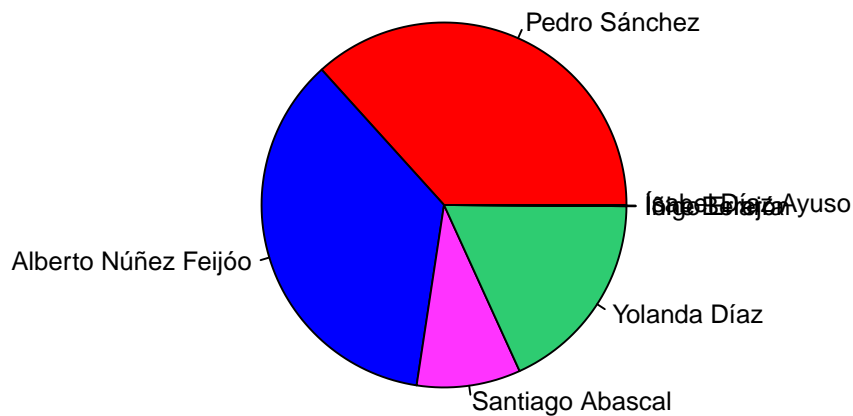


6.2 Gráfico de sectores

Alternativamente se podría usar el gráfico de sectores.

```
valores<-as.numeric(round(prop.table(table(df_cis$recPREFPTE))*100,
                                digits=1))
nombres<-names(round(prop.table(table(df_cis$recPREFPTE))*100,digits=1))
#vector con los índices ordenados según valor decreciente
orden<-order(valores,decreasing=TRUE)
valoresOrdenados<-valores[orden]
nombresOrdenados<-nombres[orden]
pie(valores, labels = nombres,
    main = "Gráfico 2. Preferencia para Presidencia de Gobierno",
    col = colores, cex=0.8)
```

Gráfico 2. Preferencia para Presidencia de Gobierno

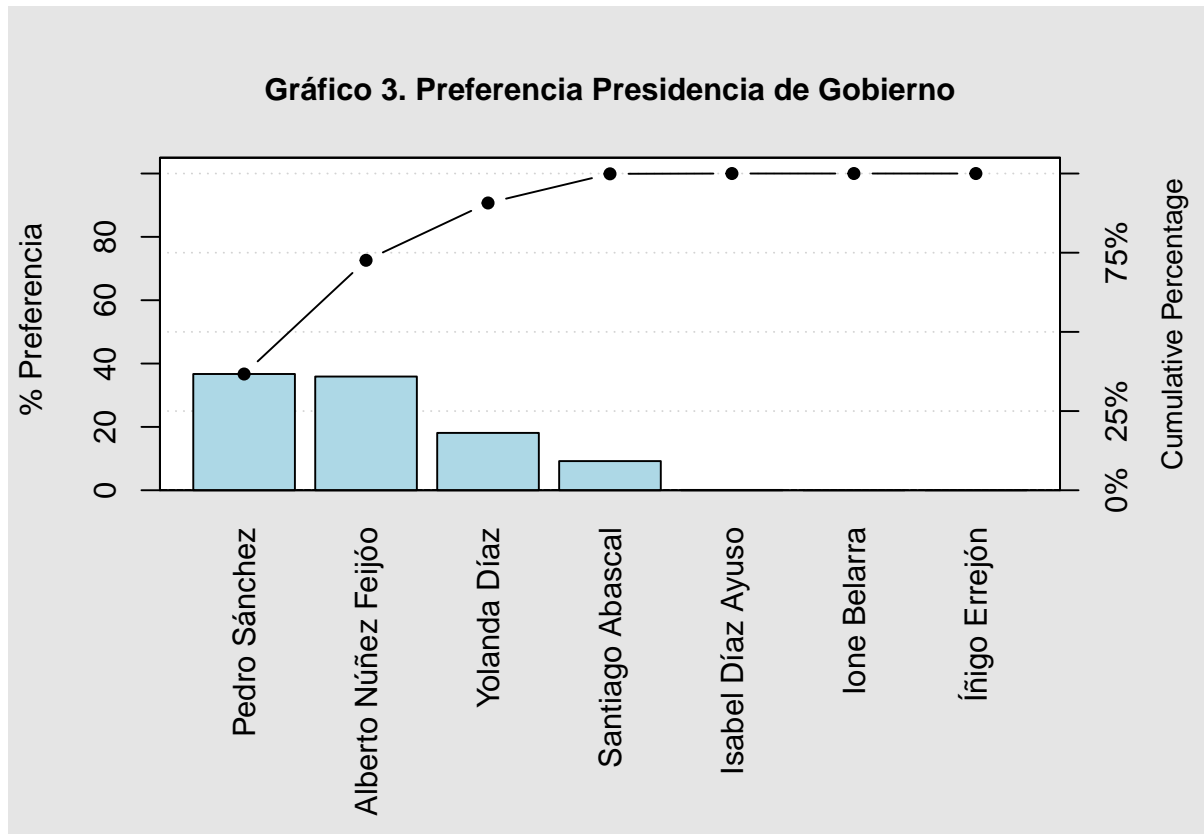


Inconvenientes del gráfico de sectores

- No muestra con la misma claridad los porcentajes que el gráfico de barras.
- Presenta el problema de solapar etiquetas continuas cuando los valores son muy pequeños.

6.3 Gráfico de Pareto

Esta otra alternativa ordena los datos y las frecuencias en sentido decreciente y muestra qué porcentaje acumulan en total las opciones representadas.



7 GUARDAR RESULTADOS EN FICHERO EXTERNO

Los resultados de la tabla de frecuencia se van a guardar en un fichero externo en la misma carpeta donde hemos guardado los datos fuente (DATOS).

```
# El atributo *output.table* permite exportar los datos a una tabla.
df_tabla<-grafico$output.table

# Se redefinen los nombres de campos
colnames(df_tabla)<-c("Frecuencia", "%", "%_acumulado", "%_válido",
                      "%_acumulado válido")

# La función *knitr* mejora el formato de la tabla a mostrar en Rmd
kable(df_tabla, column.width=c("10%", "20%", "20%", "20%", "20%"),
      caption="Tabla de distribución de frecuencias")
```

Tabla 4: Tabla de distribución de frecuencias

		%_acumulado			
	Frecuencia	%	%_acumulado	%_válido	válido
Pedro Sánchez	2631	29.9	29.9	36.7	36.7
Alberto Núñez Feijóo	2576	29.3	59.2	35.9	72.6
NA's	1625	18.5	100.0	0.0	100.0
Yolanda Díaz	1300	14.8	81.4	18.1	99.9
Santiago Abascal	658	7.5	66.7	9.2	81.8
Isabel Díaz Ayuso	8	0.1	81.5	0.1	100.0
Ione Belarra	0	0.0	81.4	0.0	99.9
Íñigo Errejón	0	0.0	81.4	0.0	99.9
Total	8798	100.0	100.0	100.0	100.0

7.1 Fichero CSV (*Comma-Separated Values*)

Una opción sería crear un fichero *csv*, se trata de un fichero de texto plano que usa el “,” o “;” como carácter delimitador de campos mediante la función *write.csv* e indicando que la

primera fila contenga el nombre de los campos (row.names = TRUE)

```
# Se guarda la tabla de frecuencias en un fichero csv  
write.csv(grafico, file = "DATOS/tablafrecResultado.csv", row.names=TRUE)
```

7.2 Hoja de cálculo MS Excel (xlsx)

Una segunda opción sería crear una hoja de cálculo de MS Excel.

```
# file= nombre del fichero hoja de cálculo de MSOffice  
write.xlsx(grafico, file = "DATOS/tablafrecResultado.xlsx",  
           sheetName = "prefpte", append = FALSE)
```







Alternativamente se puede añadir una nueva página en un fichero excel ya existente.

```
# file= nombre del fichero hoja de cálculo de MSOffice  
write.xlsx(grafico, file = "DATOS/tablafrecResultado.xlsx",  
           sheetName = "prefpte2", append = TRUE)
```

8 EJECUCIÓN DEL Rmd Y REPOSITORIO DE FICHEROS

Para la ejecución del código R, está disponible el fichero ejecutable **preferenciaPte.Rmd** y el fichero de datos **datosjulio2023.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	preferenciaPte.Rmd
	Fichero HTML	preferenciaPte.html
	Fichero PDF	preferenciaPte.pdf
	Repositorio tofermos	