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

Técnicas de Investigación en Ciencias Políticas. UBU
Práctica 1. Apartado 3

Tomàs Ferrandis Moscardó

2023-03-18

Contents

1. INTRODUCCIÓN

En esta primera parte de la actividad se hará un *analisis univariado* sobre la variable PREFPTE (Pregunta P8) del Barómetros del CIS de julio de 2023: *De los/las principales líderes políticos/as, ¿quién preferiría que fuese el/la presidente/a del Gobierno tras las elecciones?* .

En la segunda parte se tratará 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*. Aunque se le dará un tratamiento distinto como veremos.

Evidentemente al tratarse trabajos de análisi univariados solo pueden ser descriptivos.

2. OBTENCIÓN DE DATOS

2.1 Fichero y Datos en Abierto

En esta práctica se realizará a partir del *barómetro de julio 2023* correspondiente al *estudio 3415*. Se optará por descargar los datos estadísticos en un formato de fichero *.sav* de la web del CIS.

Se usará una variable con la ruta absoluta del directorio donde esté el fichero *.csv* y otra donde se inddicará el nombre en si del fichero. *Variables*

```
# El fichero .sav lo tenemos en una subcarpeta DATOS
library(knitr)
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<-haven::read_sav(rutaFichero)
```

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

2.3 Consultar el cuestionario y el diccionario de datos

Se observará el cuestionario del CIS para ver qué tipo de variable es. En el caso de ser cualitativa, en R podemos comprobar que todos los valores existentes en la base de datos coincidan coeherentemente con las respuestas válidas del cuestionario. La función *unique* que obtiene los valores existentes en la columna será de gran ayuda.

```
unique(df_cisjulio$PREFPTE)
```

```
## <labelled<double>[11]>: Preferencia personal como presidente del Gobierno central
##  [1]  2  4  3 99  1 97  8 98 96  7 10
##
## Labels:
##  value                          label
##      1                Pedro Sánchez
##      2           Alberto Núñez Feijóo
##      3               Santiago Abascal
##      4                   Yolanda Díaz
```

```
##     10                    Ione Belarra
##      7                   Íñigo Errejón
##      8               Isabel Díaz Ayuso
##     96                (NO LEER) Otro/a
##     97 (NO LEER) Ninguno/a de ellos/as
##     98                            N.S.
##     99                            N.C.
```

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* típica: con escasos valores (11) y, donde cada uno de ellos tiene una etiqueta asociada.

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

3. PREPARACIÓN DE LA MATRIZ DE DATOS

A partir de la matriz obtenida se deberá 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, se puede prescindir de los demás datos usando así, un *datra frame* de menor tamaño y más simple.

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

Para combrobar el resultado se usará, por ejemplo, una instrucción de R similar al *head* pero que leerá las filas últimas. En este caso los 4 últimos casos.

```
tail(df_cis,4)
```

```
## # A tibble: 4 x 1
##   PREFPTE
##   <dbl+lbl>
## 1  4 [Yolanda Díaz]
## 2  2 [Alberto Núñez Feijóo]
## 3 97 [(NO LEER) Ninguno/a de ellos/as]
## 4 97 [(NO LEER) Ninguno/a de ellos/as]
```

Otra comprobación interesante consistiría en comparar las dimensiones de ambos *data frames* con funciones de R como:

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

```
dim(df_cis)
```

```
## [1] 8798    1
```

```
dim(df_cisjulio)
```

```
## [1] 8798  141
```

```
ncasos<-nrow(df_cis)
ncasos1<-nrow(df_cisjulio)
nvars<-ncol(df_cisjulio)
```

Se puede ver, por ejemplo que el número de casos es 8798 en ambos *data frames* .

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 duplicaremos la columna del *data frame*

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

Valores válidos y no válidos

A partir de la observación que se ha hecho en el punto 1.2 anterior, ya se puede decidir qué valores de la variable son válidos y cuales no interesan a efectos de 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 97, 96, 98 y 99. Antes de proceder se deberá duplicar esta columna, es decir insertar 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_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. Se puede usar la función *summary* especificándole que no borre los valores NA (remove missing values = FALSE)

```
summary(df_cis$recPREFPTE, na.rm=FALSE)
```

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

```
numero_na<-sum(is.na(df_cis$recPREFPTE))
porcentaje_na<-round(mean(is.na(df_cis$recPREFPTE))*100,1)
```

Importancia de los valores perdidos NA

Hay un total de 1625 lo que supone un 18.5 de casos no válidos sobre el total de 8798 que se exluirán del análisis

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* carecen de sentido.

4.1 Moda

Para el cálculo de la moda, R no dispone de ninguna función en sus librerías, se deberá crear.

Función Moda

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

Una vez creada la función e incorporado el código en nuestro script r (o R-markdown) se podrá llamar. Se puede guardar el resutado en una variable para poder consultarlo.

```
varModa<-moda(df_cis$recPREFPTE)
```

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 o qué valores representan una preferencia por los candidatos.

5.1 Tabla de frecuencias de Preferencia de presidente/a

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

Frecuencias absolutas

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

```
table(df_cis$recPREFPTE, useNA="no")
```

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

La función *table*, por defecto, ignora los valores NA (useNA="no"/"ifany")

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 prefienren a cada candidato.

```
round(prop.table(table(df_cis$recPREFPTE,useNA="no"))*100,digits=1)
```

```
##
##        Pedro Sánchez Alberto Núñez Feijóo     Santiago Abascal
##                 36.7                 35.9                  9.2
##         Yolanda Díaz         Ione Belarra        Íñigo Errejón
##                 18.1                  0.0                  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
- Grafico de Pareto

Vemos las diferentes opciones.

6.1 Gráfico de barras

```
require(epiDisplay)
```

```
## Loading required package: epiDisplay
```

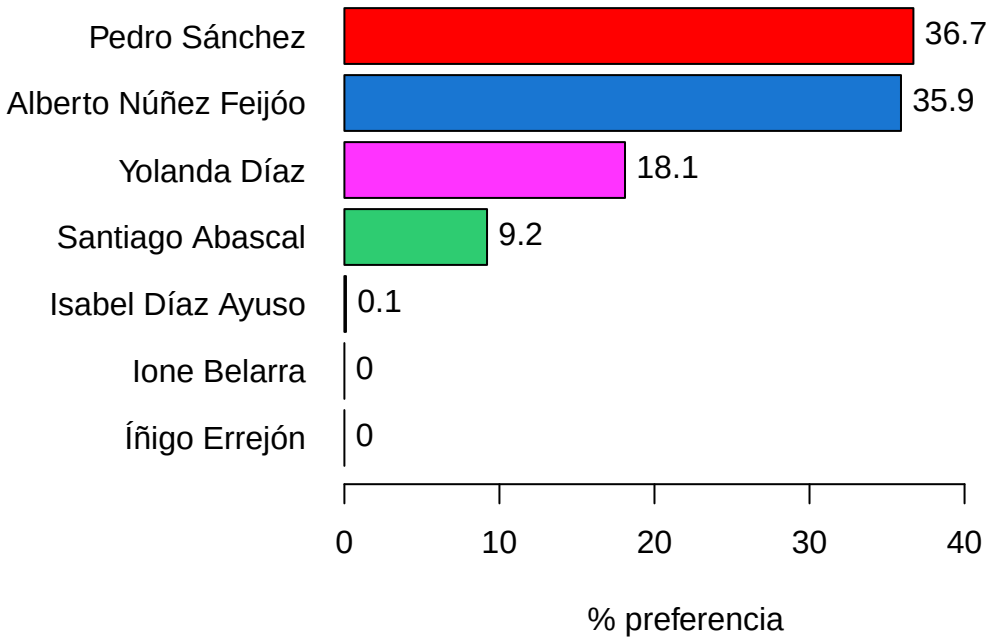
```
## Loading required package: foreign
```

```
## Loading required package: survival
```

```
## Loading required package: MASS
```

```
## Loading required package: nnet
```

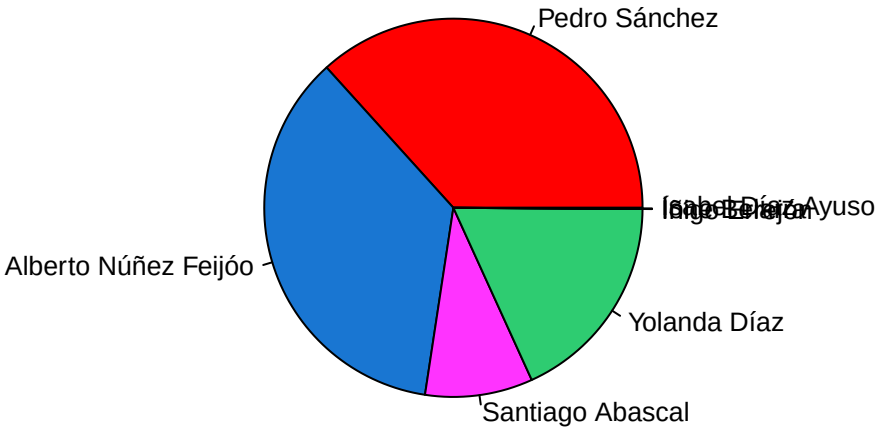
```
# Este vector tendrá los colores para cada candidato (orden decreciente ) que
#los indentifica con su opción política. Se usará en todos los gráficos.
colores<-c("red","#1976d2","#FF33FF","#2ecc71","#2980b9","#ab47bc","#c8e6c9")
coloresInvertidos<-rev(colores)
grafico <- tab1(df_cis$recPREFPTE,cum.percent = TRUE,sort.group="decreasing",
                xlab="% preferencia",decimal=1,bar.values = "percent",
                main="Gráfico 1. Preferencia para Presidencia de Gobierno",
                col=coloresInvertidos, missing = FALSE)
```



6.2 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



Como se puede observar, el gráfico de sectores no muestra con la misma claridad los porcentajes que el gráfico de barras.

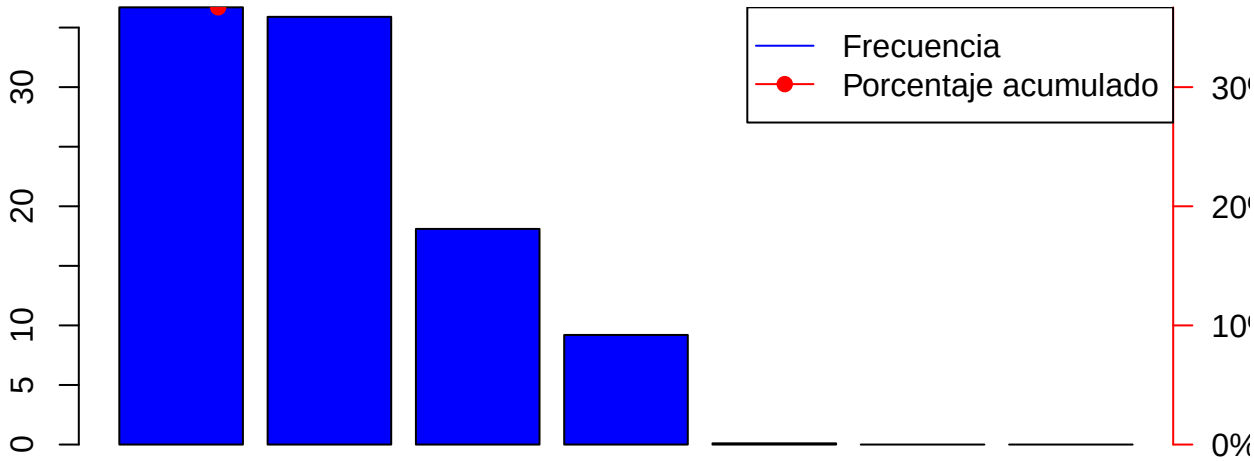
6.3 Gráfico de Pareto

Se ordenarán los datos y las frecuencias en sentido decreaciente

```
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]
colores=c("red","#1976d2","#FF33FF","#2ecc71","#2980b9","#ab47bc","#c8e6c9")

# Calcular el porcentaje acumulado
porcentajeAcumulado<-cumsum(valoresOrdenados/sum(as.numeric(valoresOrdenados)))*100

# Crear el gráfico de Pareto
par(mar = c(1, 2, 10, 2))  # Ajustar los márgenes del gráfico
ylim <- range(c(valoresOrdenados, porcentajeAcumulado))
ylim[2] <- ylim[2] + 10  # Ajustar el límite superior para dejar espacio adicional

barplot(valoresOrdenados, main = "Gráfico 3. Preferencia Presidencia de Gobierno.",
        xlab = "Candidatos", ylab = "Frecuencia", col = "blue")
        lines(porcentajeAcumulado,
                type = "b", col = "red", pch = 21, bg = "red", yaxt = "n")
```

```
        axis(4, at = seq(0, 100, by = 10), labels = paste0(seq(0, 100, by = 10),"%")
             , col = "red", las = 1)
legend("topright", legend = c("Frecuencia", "Porcentaje acumulado"), col = c("blue", "red")
       , lty = 1, pch = c(NA, 21), pt.bg = c(NA, "red"))
```

Gráfico 3. Preferencia Presidencia de Gobierno.



7 GUARDAR RESULTADO EN FICHERO EXTERNO

7.1 Fichero CSV (*Comma-Separated Values*)

El objeto *gráfico* de R tiene el atributo *output.table* para exportar los datos a una tabla. Con la fución *knitr* de R mejoramos el formato de la tabla.

```
# grafico$output.table # Esta dentro de los atributos, es un exportable que se genera
df_tabla<-grafico$output.table
colnames(df_tabla)<-c("Frecuencia","%","%_acumulado", "%_válido","%_acumulado válido")
knitr::kable(df_tabla,column.width=c("10%","20%","20%","20%","20%"))
```

	Frecuencia	%	%_acumulado	%_válido	%_acumulado 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

Esta tabla la guardaremos en un ficheros externo de tipo *csv* mediante el la función *write.csv* indicando que la primera fila contenga el nombre de los campos (row.names = TRUE)

```
# Guardamos la tabla de frecuencias en un fichero csv
write.csv(grafico, file = "tablafrec.csv", row.names = TRUE)
```

7.2 Hoja de cálculo MS-Excel (xlsx)

Mediante esta opción se puede crear una página nueva en un fichero existente de MS Excel.

```
if (!require(xlsx)){install.packages("xlsx")}
```

```
## Loading required package: xlsx
```

```
#install.packages("xlsx")
#require(xlsx)

write.xlsx(grafico, file = "resultados.xlsx",                # file= nombre del fichero
           sheetName = "prefpte", append = FALSE)            # file= nombre de la página
```