# Trabajo final Muestreo

*Arturo Gonzalez Moya*

*20 marzo, 2021*

## Contents



# 1 Enunciado

**La Encuesta sobre equipamiento y uso de tecnologías de información y comunicación en los hogares (INE) es una encuesta anual que elabora el Instituto Nacional de Estadística y tiene por objeto obtener datos del desarrollo de lo que se ha denominado Sociedad de la Información.1 En la encuesta se pide información sobre los hogares (factor de elevación asociado FACTOR_H) y sobre individuos de 16 a 74 años (factor de elevación asociado FACTOR_P). Los microdatos de las encuestas realizadas durante los años 2019 y 2020 se encuentran en los ficheros TIC_2019.csv y TIC_2020.csv. Los ficheros Diseño de Registro anonimizado TIC-H19.xlsx y Diseño de Registro TIC_H2020_ANONIMIZADO.xlsx contienen el diseño del registro de la encuesta, que recoge, entre otros aspectos, los descriptores de las variables así como los valores válidos de las mismas. Hay que tener en cuenta que el diseño de registro ha cambiado de un año para el otro y algunas variables han cambiado de nombre, otras han desaparecido, yse han incorporado de nuevas. El objetivo de este trabajo de prácticas será cuantificar algunos de los cambios en el uso de la tecnología por parte de individuos presumiblemente a raíz del confinamiento por la pandemia. Se pide:**

## 1.1 Comparar para ambos años y para los grandes grupos de edad 16-29, 30-44, 45-59 y "más de 60 años":

### a) Número y porcentaje de individuos que han usado de internet alguna vez

Lo primero que haremos será leer los datos para las años 2019 y 2020.

```
datos_2019 <- read.csv2("TIC_2019.csv")
datos_2020 <- read.csv2("TIC_2020.csv")
```

Veamos que contienen estos ficheros y si hay valores perdidos.

```
head(datos_2019)[1:3,1:15]
```

```
##   X CPRO SERIAL TR N_INF N_REF TOT_ENC N_ENC NPERS SEXO EDAD PNACTO
## 1 1    1  20001  1     1     1       1     1     1    1   32      1
## 2 2    1  20002  1     4     4       2     1     1    6   40      1
## 3 3    1  20003  1     1     1       1     1     1    6   36      1
```

```
##   NACIONALIDAD PNALDAD ESTC
## 1            1      NA    1
## 2            1      NA    2
## 3            1      NA    1
```

```
head(datos_2020)[1:3,1:15]
```

```
##   X CPRO SERIAL TR N_INF N_REF TOT_ENC N_ENC NPERS SEXO EDAD PNACTO
## 1 1    1  25224  1     1     1       1     1     1    1   33      1
## 2 2    1  25225  1     4     4       2     1     1    6   42      1
## 3 3    1  25226  1     1     1       1     1     1    6   37      1
##   NACIONALIDAD PNALDAD ESTC
## 1            1      NA    1
## 2            1      NA    2
## 3            1      NA    1
```

Podemos ver que en ambos conjuntos de datos,por ejemplo, la variable `PNALDAD` tiene muchos valores perdidos. Mientras no sea necesario, no la eliminaremos.

Veamos ahora el número y porcentaje de individuos que han usado internet alguna vez separando por grupos de edad. En ambos ficheros, la variable que dice si un encuestado ha utilizado o no internet es `USO_INT`. Para calcular el número y el porcentaje de individuos en cada apartado utilizaremos la siguiente función.

```
num_porcentaje_encuesta<-function(variable,valor,factorElev=1){
        n<-length(variable)
        if (missing(factorElev) || factorElev==1) fe<-rep(1,n)
        else fe<-factorElev
        a<-cbind(variable,fe)
        num<-sum(a[,2]*(a[,1]==valor))
        return(list("Numero" = round(num,0), "Porcentaje" = num/sum(a[,2])))
}
```

Comenzaremos con los individuos entre 16 y 29 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=16 & EDAD <=29)

p2 <-datos_2020 %>%
  filter(EDAD >=16 & EDAD <=29)

d_a_16_2019 <- num_porcentaje_encuesta(p1$USO_INT,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_a_16_2020 <- num_porcentaje_encuesta(p2$USO_INT,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos en este rango de edad que han utilizado internet alguna vez para el año 6600875 y para 2020 son 6788961. Vemos que en el año 2020, el número de individuos entre 16 y 29 años que han utilizado internet alguna vez ha aumentado con respecto a 2019. si miramos el porcentaje, observamos que en 2019 el 99.1442635% de los individuos con edades entre 16 y 29 años ha utilizado internet alguna vez, mientras que en 2020 tenemos un 99.8010786%.

Pasemos ahora al grupo de individuos con edad entre 30 y 44 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=30 & EDAD <=44)

p2 <-datos_2020 %>%
  filter(EDAD >=30 & EDAD <=44)
```

```
d_a_30_2019 <- num_porcentaje_encuesta(p1$USO_INT,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_a_30_2020 <- num_porcentaje_encuesta(p2$USO_INT,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos en este rango de edad que han utilizado internet alguna vez para el año 2019 son 9897718 y para 2020 son 9858062. Vemos que en el año 2020, el número de individuos entre 16 y 29 años que han utilizado internet alguna vez ha aumentado con respecto a 2019. si miramos el porcentaje, observamos que en 2019 el 97.6667597% de los individuos con edades entre 16 y 29 años ha utilizado internet alguna vez, mientras que en 2020 tenemos un 99.4336209%. Vemos que este porcentaje ha aumentado en un 2% entre ambos años.

Miremos ahora las personas con una edad comprendida entre los 45 y 59 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=45 & EDAD <=50)

p2 <-datos_2020 %>%
  filter(EDAD >=45 & EDAD <=50)

d_a_45_2019 <- num_porcentaje_encuesta(p1$USO_INT,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_a_45_2020 <- num_porcentaje_encuesta(p2$USO_INT,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos en este rango de edad que han utilizado internet alguna vez para el año 2019 son 4287494 y para 2020 son 4491396. Vemos que en el año 2020, el número de individuos entre 16 y 29 años que han utilizado internet alguna vez ha aumentado con respecto a 2019. si miramos el porcentaje, observamos que en 2019 el 95.9825069% de los individuos con edades entre 16 y 29 años ha utilizado internet alguna vez, mientras que en 2020 tenemos un 97.6490999%. Vemos que este porcentaje ha aumentado en un 2% entre ambos años, igual que en el caso anterior. También recalcar que el porcentaje ha disminuido con respecto a los grupos de edad anteriores.

Por ultimo en este apartado, veamos que ocurre con los individuos con edad mayor a 60 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=60)

p2 <-datos_2020 %>%
  filter(EDAD >=60)

d_a_60_2019 <- num_porcentaje_encuesta(p1$USO_INT,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_a_60_2020 <- num_porcentaje_encuesta(p2$USO_INT,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos en este rango de edad que han utilizado internet alguna vez para el año 2019 son 6442231 y para 2020 son 7293160. Vemos que en el año 2020, el número de individuos entre 16 y 29 años que han utilizado internet alguna vez ha aumentado con respecto a 2019. si miramos el porcentaje, observamos que en 2019 el 55.3873348% de los individuos con edades entre 16 y 29 años ha utilizado internet alguna vez, mientras que en 2020 tenemos un 61.2262863%. Vemos que este porcentaje ha aumentado casi un 6% entre ambos años. Además podemos observar que el porcentaje de ambos años ha disminuido de forma notable con respecto a los otros grupos de edades.

**b) Número y porcentaje de individuos que usan internet varias veces al día.**

La columna que representa esta pregunta en nuestro conjunto de datos es la que se llama `VINTD`. Esta columna

contiene valores perdidos que corresponden a los que han respondido que no utilizan internet diariamente y no queremos eliminarlos porque variariamos la población total. Lo que haremos entonces será cambiar los valores `NA` al valor 6 (que corresponde a los que han respondido *NO* a la encuesta).

```
datos_2019$VINTD[is.na(datos_2019$VINTD)] <- 6
datos_2020$VINTD[is.na(datos_2020$VINTD)] <- 6
```

Comenzamos ahora con el grupo de individuos con edades entre 16 y 29 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=16 & EDAD <=29)

p2 <-datos_2020 %>%
  filter(EDAD >=16 & EDAD <=29)

d_b_16_2019 <- num_porcentaje_encuesta(p1$VINTD,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_b_16_2020 <- num_porcentaje_encuesta(p2$VINTD,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

En el año 2019, el número de individuos con edad entre 16 y 29 años que utilizan internet más de una vez al día fue de 6245266, mientras que en 2020, este número fue de 6564331. Vemos que ha aumentado notablemente. Si lo miramos en porcentajes, en el año 2019 tenemos un porcentaje de 93.8030603% , mientras que en 2020 tenemos un porcentaje de 96.4988948%. Podemos observar que ha aumentado casi un 3% de un año para otro.

Veamos ahora que ocurre con las personas con edad entre 30 y 44 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=30 & EDAD <=44)

p2 <-datos_2020 %>%
  filter(EDAD >=30 & EDAD <=44)

d_b_30_2019 <- num_porcentaje_encuesta(p1$VINTD,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_b_30_2020 <- num_porcentaje_encuesta(p2$VINTD,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

En el año 2019, el número de individuos con edad entre 30 y 44 años que utilizan internet más de una vez al día fue de 8792078, mientras que en 2020, este número fue de 9085112. Vemos que ha aumentado de forma considerable. Si lo miramos en porcentajes, en el año 2019 tenemos un porcentaje de 86.7567432% , mientras que en 2020 tenemos un porcentaje de 91.6372424%. Podemos observar que ha aumentado casi un 5% de un año para otro.

Proseguimos con las personas con edad entre 45 y 59 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=45 & EDAD <=59)

p2 <-datos_2020 %>%
  filter(EDAD >=45 & EDAD <=59)

d_b_45_2019 <- num_porcentaje_encuesta(p1$VINTD,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_b_45_2020 <- num_porcentaje_encuesta(p2$VINTD,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

En el año 2019, el número de individuos con edad entre 45 y 59 años que utilizan internet más de una vez al día fue de 7846550, mientras que en 2020, este número fue de 8858731. De un año para otro este número ha crecido en un millón, lo que es impresionante. Si lo miramos en porcentajes, en el año 2019 tenemos un porcentaje de 73.1179194% , mientras que en 2020 tenemos un porcentaje de 80.9813961%. Observemos que ha aumentado un 7% de un año para otro.

Por ultimo, veamos que ocurre con las personas mayores de 60 años.

```
p1 <- datos_2019 %>%
  filter(EDAD >=60)

p2 <-datos_2020 %>%
  filter(EDAD >=60)

d_b_60_2019 <- num_porcentaje_encuesta(p1$VINTD,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_b_60_2020 <- num_porcentaje_encuesta(p2$VINTD,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

En el año 2019, el número de individuos mayores de 60 años que utilizan internet más de una vez al día fue 3687303, mientras que es 2020 fue 4685783. En este caso, podemos observar que también hay un millón más de estos individuos en 2020 que en 2019. Si lo miramos en porcentaje, en 2019 tenemos un 31.7017313%, mientras que en 2020 tenemos un 39.3372836%. Vemos que el porcetaje en este rango de edad es notablemente menor a los porcentajes de los otros grupos de edad. De un año para otro, este porcentaje ha crecido casi un 8%.

**c) Número y porcentaje de individuos que, habiendo usado internet, han realizado llamadas telefónicas o videoconferencias por internet.**

En este caso, las variables asociadas a la pregunta que nos piden son diferentes en los años 2019 y 2020. En el año 2019 la variable es `SERV16_2` y en 2020 la variable es `SERV14_2`. Además, la población total que queremos estudiar en este apartado son los individuos que han utilizado internet alguna vez, por lo que hemos de filtrar los datos.

```
apartado_c_2019 <- datos_2019 %>%
  filter(USO_INT == 1)
apartado_c_2020 <- datos_2020 %>%
  filter(USO_INT == 1)

anyNA(apartado_c_2019$SERV16_2)
```

```
## [1] TRUE
```

```
anyNA(apartado_c_2020$SERV14_2)
```

```
## [1] TRUE
```

Podemos observar que hay valores perdidos, que pertenecen a los individuos que hace mas de 3 meses que utilizaron internet, por lo tanto pondremos estos `NAs` como valores 6 que corresponde a los que han respondido que no en esa pregunta de la encuesta.

```
apartado_c_2019$SERV16_2[is.na(apartado_c_2019$SERV16_2)] <- 6
apartado_c_2020$SERV14_2[is.na(apartado_c_2020$SERV14_2)] <- 6
```

Ahora comenzamos con el primer grupo de edad a estudiar (individuos entre 16 y 29 años).

```
p1 <- apartado_c_2019 %>%
  filter(EDAD >=16 & EDAD <=29)
```

```
p2 <-apartado_c_2020 %>%
  filter(EDAD >=16 & EDAD <=29)


d_c_16_2019 <- num_porcentaje_encuesta(p1$SERV16_2,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_c_16_2020 <- num_porcentaje_encuesta(p2$SERV14_2,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 16 y 29 años que han utilizado internet alguna vez y que han realizado llamadas telefónicas o videoconferencias por internet en 2019 fue de 5031290, mientras que en 2020 fue de 6312394. Vemos que el número de individuos ha aumentado en un millon trecientos mil entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 76.2215633%, mientras que en 2020 fue de un 92.9802559%. El porcentaje ha aumentado de un año para otro en un 16%.

Sigamos estudiando el grupo de individuos con edad entre 30 y 44 años.

```
p1 <- apartado_c_2019 %>%
  filter(EDAD >=30 & EDAD <=44)

p2 <-apartado_c_2020 %>%
  filter(EDAD >=30 & EDAD <=44)

d_c_30_2019 <- num_porcentaje_encuesta(p1$SERV16_2,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_c_30_2020 <- num_porcentaje_encuesta(p2$SERV14_2,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 30 y 44 años que han utilizado internet alguna vez y que han realizado llamadas telefónicas o videoconferencias por internet en 2019 fue de 6326842, mientras que en 2020 fue de 8631864. Vemos que el número de individuos ha aumentado más de dos millones entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 63.9222242%, mientras que en 2020 fue de un 87.5614681%. El porcentaje ha aumentado de un año para otro en más de un 20%.

Prosigamos con los inviduos con edad entre 45 y 59 años.

```
p1 <- apartado_c_2019 %>%
  filter(EDAD >=45 & EDAD <=59)

p2 <-apartado_c_2020 %>%
  filter(EDAD >=45 & EDAD <=59)

d_c_45_2019 <- num_porcentaje_encuesta(p1$SERV16_2,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_c_45_2020 <- num_porcentaje_encuesta(p2$SERV14_2,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 45 y 59 años que han utilizado internet alguna vez y que han realizado llamadas telefónicas o videoconferencias por internet en 2019 fue de 5470292, mientras que en 2020 fue de 8337968. Vemos que el número de individuos ha aumentado casi tres millones entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 54.1701422%, mientras que en 2020 fue de un 79.4341668%. El porcentaje ha aumentado de un año para otro un 25%.

Por ultimo, estudiaremos los individuos con edad mayor a 60 años.

```
p1 <- apartado_c_2019 %>%
  filter(EDAD >=60)
```

```r
p2 <-apartado_c_2020 %>%
  filter(EDAD >=60)

d_c_60_2019 <- num_porcentaje_encuesta(p1$SERV16_2,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_c_60_2020 <- num_porcentaje_encuesta(p2$SERV14_2,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos mayores de 60 años que han utilizado internet alguna vez y que han realizado llamadas telefónicas o videoconferencias por internet en 2019 fue de 2769035, mientras que en 2020 fue de 4830037. Vemos que el número de individuos casi se ha duplicado entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 42.9825482%, mientras que en 2020 fue de un 66.226937%. El porcentaje ha aumentado de un año para otro un 24%.

Podemos concluir que el número de individuos que han utilizado alguna vez internet y que han realizado llamadas telefónicas o videoconferencias por internet ha aumentado de manera soprendente en 2020 por culpa de la pandemia.

**d) Número y porcentaje de individuos que han realizado compras por internet.**

En este apartado, la variable que necesitamos utilizar es la variable `COMPRAS`. Veamos primero si contiene valores perdidos.

```r
anyNA(datos_2019$COMPRAS)
```

```
## [1] TRUE
```

```r
anyNA(datos_2020$COMPRAS)
```

```
## [1] TRUE
```

Estos valores provienen de los individuos que utilizaron internet hace más de un año o que no lo han utilizado nunca. Por lo tanto lo que haremos será tranformar estos valores perdidos como individuos que no ha comprado por intenet.

```r
datos_2019$COMPRAS[is.na(datos_2019$COMPRAS)] <- 6
datos_2020$COMPRAS[is.na(datos_2020$COMPRAS)] <- 6
```

Comenzamos con los individuos con edades entre 16 y 29 años.

```r
p1 <- datos_2019 %>%
  filter(EDAD >=16 & EDAD <=29)

p2 <-datos_2020 %>%
  filter(EDAD >=16 & EDAD <=29)

d_d_16_2019 <- num_porcentaje_encuesta(p1$COMPRAS,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_d_16_2020 <- num_porcentaje_encuesta(p2$COMPRAS,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 16 y 29 años que han realizado compras por internet en 2019 fue de 5277689, mientras que en 2020 fue de 5580871. Vemos que el número de individuos ha aumentado entre un año y otro, pero no tanto como en otros apartados. Si miramos el porcentaje, en 2019 fue de un 79.2701767%, mientras que en 2020 fue de un 82.0415588%. El porcentaje ha aumentado de un año para otro en menos de un 3%.

Sigamos con los individuos con edad entre 30 y 44 años.

```r
p1 <- datos_2019 %>%
  filter(EDAD >=30 & EDAD <=44)

p2 <-datos_2020 %>%
  filter(EDAD >=30 & EDAD <=44)

d_d_30_2019 <- num_porcentaje_encuesta(p1$COMPRAS,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_d_30_2020 <- num_porcentaje_encuesta(p2$COMPRAS,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 30 y 44 años que han realizado compras por internet en 2019 fue de 7735664, mientras que en 2020 fue de 8146451. Vemos que el número de individuos ha aumentado entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 76.3324607%, mientras que en 2020 fue de un 82.1694057%. El porcentaje ha aumentado de un año para otro casi un 6%.

Pasemos con los individuos con edad entre 45 y 59 años.

```r
p1 <- datos_2019 %>%
  filter(EDAD >=45 & EDAD <=59)

p2 <-datos_2020 %>%
  filter(EDAD >=45 & EDAD <=59)

d_d_45_2019 <- num_porcentaje_encuesta(p1$COMPRAS,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_d_45_2020 <- num_porcentaje_encuesta(p2$COMPRAS,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos entre 45 y 59 años que han realizado compras por internet en 2019 fue de 6532408, mientras que en 2020 fue de 7174026. Vemos que el número de individuos ha aumentado en medio millón entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 60.8721089%, mientras que en 2020 fue de un 65.5807983%. El porcentaje ha aumentado de un año para otro casi un 5%.

Por ultimo, veamos que ocurre con los individuos de más de 60 años.

```r
p1 <- datos_2019 %>%
  filter(EDAD >=60)

p2 <-datos_2020 %>%
  filter(EDAD >=60)

d_d_60_2019 <- num_porcentaje_encuesta(p1$COMPRAS,
                                       valor = 1, factorElev = p1$FACTOR_P)
d_d_60_2020 <- num_porcentaje_encuesta(p2$COMPRAS,
                                       valor = 1, factorElev = p2$FACTOR_P)
```

El número de individuos con más de 60 años que han realizado compras por internet en 2019 fue de 2365261, mientras que en 2020 fue de 2831730. Vemos que el número de individuos ha aumentado en casi medio millón entre un año y otro. Si miramos el porcentaje, en 2019 fue de un 20.335424%, mientras que en 2020 fue de un 23.7724571%. El porcentaje ha aumentado de un año para otro más de un 3%.

## 1.2 Comparar los resultados del conjunto de España y los de les Illes Balears (Código provincial, CPRO==7) para el año 2020:

**Número y porcentaje de individuos que han visto peliculas o series bajo demanda (Netflix, HBO, Filmin y similares);**

La variable que necesitamos utilizar en este apartado es `SERV15_3`. Lo primero que hacemos será mirar si contiene valores perdidos.

```
anyNA(datos_2020$SERV15_3)
```

```
## [1] TRUE
```

Vemos que esta variable contiene valores perdidos, que si miramos los datos de la encuesta, vienen de los individuos que no han utilizado internet los últimos 3 meses. Por lo tanto lo que haremos será pasar estos valores perdidos al valor 6, que corresponde a que en la encuesta hayan respondido que no.

```
datos_2020$SERV15_3[is.na(datos_2020$SERV15_3)] <- 6
```

Ahora separaremos los datos de las Islas Baleares para el estudio.

```
datos_ib <- filter(datos_2020, CPRO == 7)
```

Pasamos a calcular lo que se nos pide.

```
d_2_a_ib<- num_porcentaje_encuesta(datos_ib$SERV15_3,
                                   valor = 1, factorElev = datos_ib$FACTOR_P)
d_2_a_esp <- num_porcentaje_encuesta(datos_2020$SERV15_3,
                                     valor = 1, factorElev = datos_2020$FACTOR_P)
```

El número de individuos en 2020 en las Islas Baleares que han visto peliculas o series bajo demanda es de 537543, mientras que en España es de 19654919. Si ahora miramos el porcentaje, podemos observar que en las Islas Baleares es de 52.641777% y en España es de 49.6741048%. Vemos que en las Islas Baleares el porcentaje es mayor.

**Número y porcentaje de individuos que han realizado compras en internet de entregas de restaurantes, de comida rápida, etc**

La variable que corresponde a esta pregunta en la encuesta es `PROD11`. Veamos primero si contiene valores perdidos.

```
anyNA(datos_2020$PROD11)
```

```
## [1] TRUE
```

Vemos que esta variable contiene valores perdidos, que si miramos los datos de la encuesta, vienen de los individuos que no han comprado por internet el último año. Por lo tanto lo que haremos será pasar estos valores perdidos al valor 6, que corresponde a que en la encuesta hayan respondido que no.

```
datos_2020$PROD11[is.na(datos_2020$PROD11)] <- 6
```

Seleccionamos los datos de las Islas Baleares.

```
datos_ib <- filter(datos_2020, CPRO == 7)
```

En este caso hemos de modificar un poco la función que calcula el número y porcentaje, ya que en esta variable tenemos 2 valores que representan que un individuo sí ha comprado por internet.

```
num_porcentaje_encuesta<-function(variable,valor_1,valor_2,factorElev=1){
        n<-length(variable)
        if (missing(factorElev) || factorElev==1) fe<-rep(1,n)
        else fe<-factorElev
```

```
        a<-cbind(variable,fe)
        num<-sum(a[,2]*(a[,1]==valor_1))
        num <- num + sum(a[,2]*(a[,1]==valor_2))
        return(list("Numero" = round(num,0), "Porcentaje" = num/sum(a[,2])))
}
```

Pasamos a calcular lo que nos pide. En este caso, los valores de la variable `PROD11` que nos interesan son 1 y 2.

```
d_2_b_ib<- num_porcentaje_encuesta(datos_ib$PROD11,
                                   valor_1 = 1, valor_2 = 2, factorElev = datos_ib$FACTOR_P)
d_2_b_esp <- num_porcentaje_encuesta(datos_2020$PROD11,
                                     valor_1 = 1, valor_2 = 2, factorElev = datos_2020$FACTOR_P)
```

El número de individuos en 2020 en las Islas Baleares que han han realizado compras en internet de entregas de restaurantes, de comida rápida, etc, es de 241294, mientras que en España es de 7762869. Si ahora miramos el porcentaje, podemos observar que en las Islas Baleares es de 23.6300297% y en España es de 19.6191884%. Vemos que en las Islas Baleares el porcentaje es un 4% mayor.

### 1.3 Dado el tamaño de la muestra, suponiendo un muestreo aleatorio simple y máxima incertidumbre (p = q = 0.5), ¿con qué probabilidad el error de estimación de una proporción poblacional será, en valor absoluto, inferior al 1%?

Primero lo haremos para el caso del año 2019. La encuesta realizada es estratificada con reposición, por lo tanto la probabilidad que nos piden es la siguiente.

```
z_cr_2019 <- sqrt((0.01^2*nrow(datos_2019))/0.25)
1-2*pnorm(-z_cr_2019)
```

```
## [1] 0.9912756
```

La probabilidad de que el error cometido con esta muestra sea menor que el 1% es de un 99.12756%.

Ahora lo haremo para el caso del año 2020. La encuesta es, igual que en el caso anterior, estratificada y con reposición. Entonces la probabilidad que nos piden es la siguiente.

```
z_cr_2020 <- sqrt((0.01^2*nrow(datos_2020))/0.25)
1-2*pnorm(-z_cr_2020)
```

```
## [1] 0.9867633
```

La probabilidad de que el error cometido con esta muestra sea menor que el 1% es de un 98.67633%.