

Práctica: análisis de representatividad

Máster de Análisis Político y Electoral (UC3M)

Nov. 2019

La práctica

En esta práctica vas a realizar un análisis de representatividad de una encuesta realizada por la empresa GESOP.

La **encuesta** (n=1000) fue realizada por la empresa GESOP. El trabajo de campo se hizo mediante llamadas, sin especificar si a fijos o a móviles, entre los días 13 y 15 de febrero. El muestreo de los hogares fue estratificado por comunidad y tamaño del municipio. En cada hogar la persona entrevistada fue elegida utilizando cuotas de sexo y edad. La ficha técnica y los microdatos están disponibles en abierto.

Planificar la tarea

Antes de empezar con los datos dedica unos cinco minutos a planificar la tarea. En la carpeta **datos** está el `cuestionario_GESOP_feb2019.pdf`. Abre el cuestionario (pp. 4-7) y responde a las siguientes preguntas:

¿Qué variables contenidas en el cuestionario pueden servir para realizar un análisis de representatividad?

¿De qué fuente (estudio/organización) podrías recabar los datos poblacionales de las variables que has señalado en la pregunta anterior?

Aquí tienes una lista de las variables de la encuesta para las que existen referentes poblacionales y la fuente de esos datos:

Variable	Descripción	Fuente
caut	Comunidad autónoma de residencia	Padrón continuo de habitantes (INE)
edad	Edad	Padrón continuo de habitantes (INE)
estud	Estudios acabados	Encuesta Social Europea (R8)
ocupa	Actividad económica	Encuesta Población Activa (INE)
sexo	Sexo	Padrón continuo de habitantes (INE)
tamuni	Tamaño del municipio	Padrón continuo de habitantes (INE)

1. Datos y paquetes

Para este ejercicio vas a utilizar la encuesta `encuesta_gesop.RDS` y los datos poblacionales `datos_poblacionales.RDS` que se encuentran en la carpeta `datos`. Además harás uso de los paquetes:

- `tidyverse`: Contiene a su vez un conjunto de paquetes que facilitan la gestión y el análisis de los datos.
- `expss`: Permite crear tablas personalizadas.

1) Cargar los paquetes:

```
# install.packages("tidyverse")
# install.packages("expss")
library(tidyverse) # tidyverse for data management
library(expss)
```

2) Cargar los datos:

```
encuesta <- read_rds("datos/encuesta_gesop.RDS")
datos_poblacionales <- read_rds("datos/datos_poblacionales.RDS")
```

El objeto `encuesta` es un *data frame* en el que están contenido los datos de la encuesta. El objeto `datos_poblacionales` es también un *data frame* en el que están las estimaciones (porcentajes) poblacionales de las variables que vas a utilizar en el análisis de representatividad. Los datos poblacionales están en formato largo:

```
head(datos_poblacionales)
```

```
##                                variable_valor variable
## 1 Comunidad autónoma de residencia (AUT)|Andalucía    caut
## 2   Comunidad autónoma de residencia (AUT)|Aragón    caut
## 3   Comunidad autónoma de residencia (AUT)|Asturias    caut
## 4   Comunidad autónoma de residencia (AUT)|Balears    caut
## 5   Comunidad autónoma de residencia (AUT)|Canarias    caut
## 6 Comunidad autónoma de residencia (AUT)|Cantabria    caut
##                                variable_desc          fuente
## 1 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
## 2 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
## 3 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
## 4 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
## 5 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
```

```
## 6 Comunidad autónoma de residencia Padrón continuo de habitantes (INE)
##      valor valor_orden pobla
## 1 Andalucía          1  18.1
## 2  Aragón            2   2.8
## 3 Asturias          3   2.5
## 4  Balears          4   2.2
## 5 Canarias          5   4.5
## 6 Cantabria         6   1.3
```

2. Crear una tabla con las estimaciones de la muestra

Primero, explora a los datos de la encuesta para ver qué variables nos interesan para este análisis.

```
glimpse(encuesta)
```

```
## Observations: 1,000
## Variables: 29
## $ id          <dbl> 4, 5, 6, 27, 28, 31, 33, 36, 38, 39, 42, 43, ...
## $ caut        <fct> Madrid, Madrid, Madrid, Galicia, Madrid, Madr...
## $ tamuni       <fct> Más de 500.000 hab., Más de 500.000 hab., Más...
## $ import_elec  <fct> Bastantes importantes, Muy importantes, NS, M...
## $ urnas        <fct> Probablemente votaría, Probablemente no votar...
## $ idv          <fct> En blanco, No votaría, No votaría, PP, PSOE, ...
## $ simpa        <fct> NA, C's, Ninguno, NA, NA, NA, NS, NA, NA, Nin...
## $ recuerdo     <fct> PP, C's, No votó, PP, No votó, PP, C's, C's, ...
## $ conoce_pc    <fct> Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, No, Sí, N...
## $ valora_pc     <dbl> 3, 1, 5, 7, 5, 7, 1, 9, 0, NA, 6, NA, 5, NA, ...
## $ conoce_ar    <fct> Sí, Sí, Sí, Sí, No, Sí, Sí, Sí, Sí, Sí, Sí, N...
## $ valora_ar     <dbl> 5, 4, 98, 5, NA, 6, 5, 8, 0, 0, 8, NA, 1, 98,...
## $ conoce_ps    <fct> Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, S...
## $ valora_ps     <dbl> 1, 4, 5, 0, 10, 1, 4, 0, 9, 2, 2, 7, 0, 98, 0...
## $ conoce_pi    <fct> Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, Sí, S...
## $ valora_pi     <dbl> 0, 3, 5, 0, 5, 3, 3, 0, 7, 8, 2, 5, 0, 98, 0...
## $ conoce_sa    <fct> Sí, Sí, Sí, Sí, Sí, Sí, No, Sí, Sí, No, Sí, N...
## $ valora_sa     <dbl> 4, 0, 5, 2, 4, 8, NA, 10, 0, NA, 10, NA, 9, N...
## $ pref_pres    <fct> Pablo Casado, Albert Rivera, Ninguno, Pablo C...
## $ conf_cat      <fct> Mediante el diálogo con la fuerzas independen...
## $ responsab_cat <fct> La Generaitat, La Generaitat, NA, El Gobierno...
## $ cesion_gob_cat <fct> Sí, Sí, No, Sí, Sí, Sí, NS, Sí, No, NS, Sí, N...
## $ juicio_indep_cat <fct> Sí, Sí, Sí, Sí, Sí, No, Sí, Sí, No, No, Sí, N...
## $ ideo         <fct> Centro, Centro, NS, Centro derecha, Centro, C...
## $ nacional      <fct> Únicamente de su comunidad, Tan español como ...
## $ sexo         <fct> Mujer, Mujer, Hombre, Mujer, Hombre, Mujer, M...
## $ edad         <fct> 18-29, 30-44, 45-59, 45-59, 45-59, 30-44, 30-...
## $ estud        <fct> Obligatorios, Universitarios, Universitarios,...
## $ ocupa        <fct> Trabajador, Trabajador, Trabajador, Trabajado...
```

Para construir la tabla vas a utilizar el paquete `expss`. En ese paquete se declaran los datos. Posteriormente, las variables en filas, que se incluyen en la función `tab_cells()`. La operación que se pretende realizar, en este caso obtener los porcentajes de columna, se define con `tab_stat_cpct()`. La función `tab_pivot()` se encarga de ensamblar todas las partes.

Para poder unir la tabla con los datos poblacionales vas a cambiar el nombre de las columnas del objeto `tabla`. La primera columna será `variable_valor` y la segunda `muestra`.

```

tabla_muestra <- encuesta %>% # añadir datos
  tab_cells(caut, tamuni, sexo, edad, estud, ocupa) %>% # añadir variables
  tab_stat_cpct() %>% # añadir estadístico: porcentajes de columna
  tab_pivot() # ensamblar tabla

colnames(tabla_muestra) <- c("variable_valor", "muestra") # cambiar los nombres de las columnas

```

3. Unir la tabla de resultados de la encuesta los datos poblacionales y calcular la diferencia

En primer lugar se utiliza la función `left_join()` para unir la `tabla_muestra` con los `datos_poblacionales` en base a la variable `"variable_valor"`, que está presente en ambas tablas.

```

# unir tabla con totales poblacionales
tabla_muestra_pobla <- left_join(tabla_muestra, datos_poblacionales, by = "variable_valor")

```

En segundo término se calcula la **diferencia entre la muestra y la población** y se descartan las filas que contienen valores perdidos NA.

```

tabla_muestra_pobla <- tabla_muestra_pobla %>%
  mutate(dif = round(muestra - pobla, 1)) %>% #diferencia entre la muestra y la población
  filter(!is.na(dif)) # eliminar los casos que son perdidos (NA)

```

4. Analizar las desviaciones de la muestra

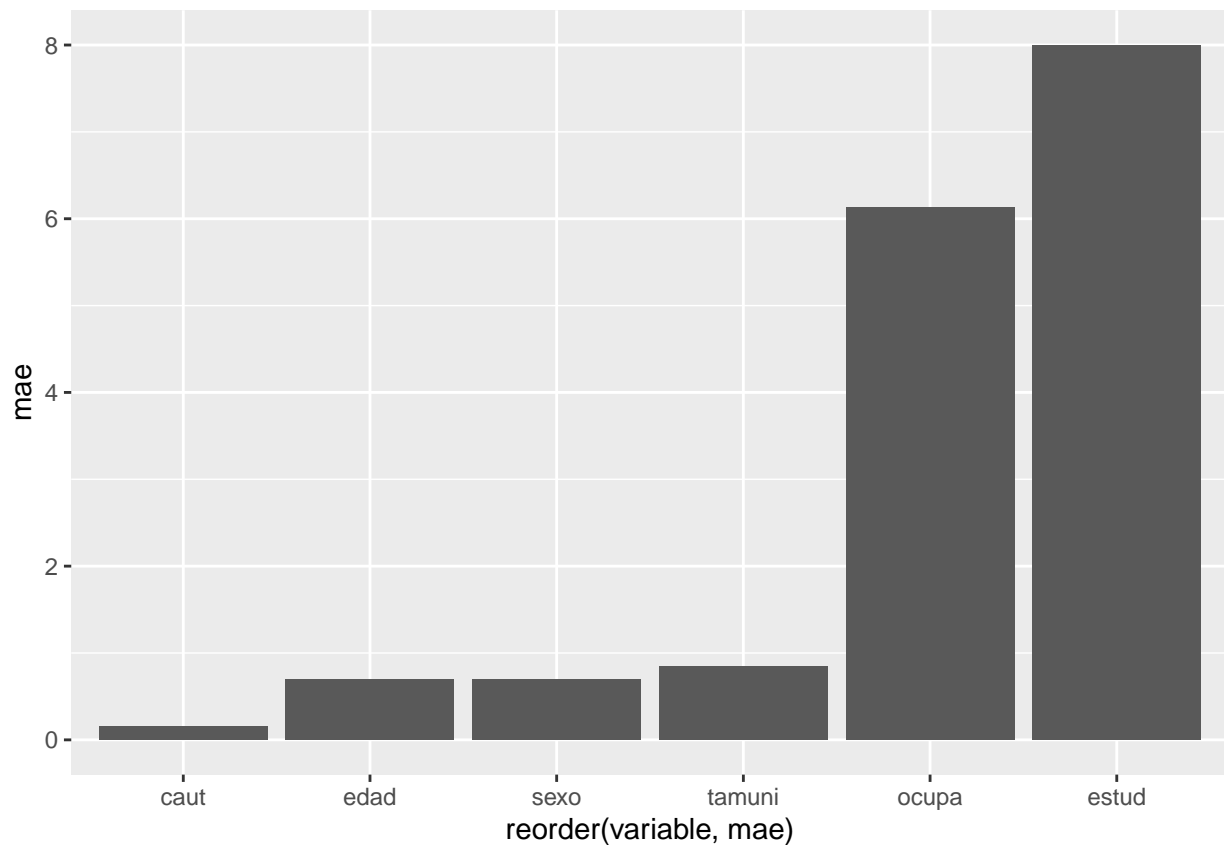
Ahora vas a analizar las desviaciones entre la muestra y la población. Para ello, crea un **gráfico resumen** en el que se visualice el **error medio absoluto** (MAE en inglés) asociado a cada variable.

*El **MAE** (mean absolute error) es una medida resumen empleada para evaluar el error total al comparar las distribuciones de dos variables categóricas. Por ejemplo, si en la población hay un 50% de hombres y un 50% de mujeres, y en la muestra los porcentajes son 45% hombres y 55% mujeres, el MAE será del 5%. En este caso, la diferencia media entre los totales poblacionales y la encuesta es del 5%.*

```

tabla_muestra_pobla %>% # calcular el error medio absoluto para cada variable
  mutate(dif_abs = abs(dif)) %>%
  group_by(variable) %>%
  summarise(mae = mean(dif_abs)) %>%
  ggplot(aes(x = reorder(variable, mae), y = mae)) +
  geom_col()

```



Ahora céntrate en las variables que más desviaciones presentan: `ocupa` y `estud`. Observa cuáles son los grupos sobrerrepresentados y subrepresentados en la muestra.

```
tabla_muestra_pobla %>%
  filter(variable %in% c("ocupa", "estud")) %>%
  select(valor, muestra, pobla, dif)
```

```
tabla_muestra_pobla %>%
  filter(variable %in% c("ocupa", "estud")) %>%
  select(variable, valor, muestra, pobla, dif) %>%
  kableExtra::kable()
```

variable	valor	muestra	pobla	dif
estud	Sin estudios obligatorios acabados	3.7	6.2	-2.5
estud	Obligatorios	22.2	39.0	-16.8
estud	Posobligatorios	18.7	11.7	7.0
estud	Posobligatorios profesionales	18.0	18.7	-0.7
estud	Universitarios	37.4	24.4	13.0
ocupa	Trabajador	54.0	48.1	5.9
ocupa	Parado	10.8	7.5	3.3
ocupa	Inactivo	35.2	44.4	-9.2