

# Diagnosticar y ajustar

## Representatividad y ajustes en encuestas con

Máster Análisis Político y Electoral (UC3M)

Pablo Cabrera Álvarez (USAL)

pablocal@usal.es |  |  @pablocalv

Nov. 2019

Tenemos un problema

# Reino Unido, 2015

The screenshot shows the BBC News homepage. At the top, there is a navigation bar with the BBC logo, a sign-in link, and categories for News, Sport, Weather, Shop, Reel, Travel, and a menu icon (M). Below this is a large red banner with the word "NEWS" in white capital letters. Underneath the banner is a secondary navigation bar with links for Home, Video, World, UK (which is underlined), Business, Tech, Science, Stories, Entertainment & Arts, Politics, Parliaments, and Brexit. The main content area features a large, bold title: "Election 2015: How the opinion polls got it wrong". Below the title, it says "By David Cowling, Editor, BBC Political Research Unit". At the bottom of the page, there is a footer with a timestamp ("© 17 May 2015"), social media sharing icons (Facebook, Messenger, Twitter, Email, and a green "Share" button), and a page number ("3 / 65").

BBC | Sign in

News | Sport | Weather | Shop | Reel | Travel | M

# NEWS

Home | Video | World | **UK** | Business | Tech | Science | Stories | Entertainment & Arts

Politics | Parliaments | Brexit

## Election 2015: How the opinion polls got it wrong

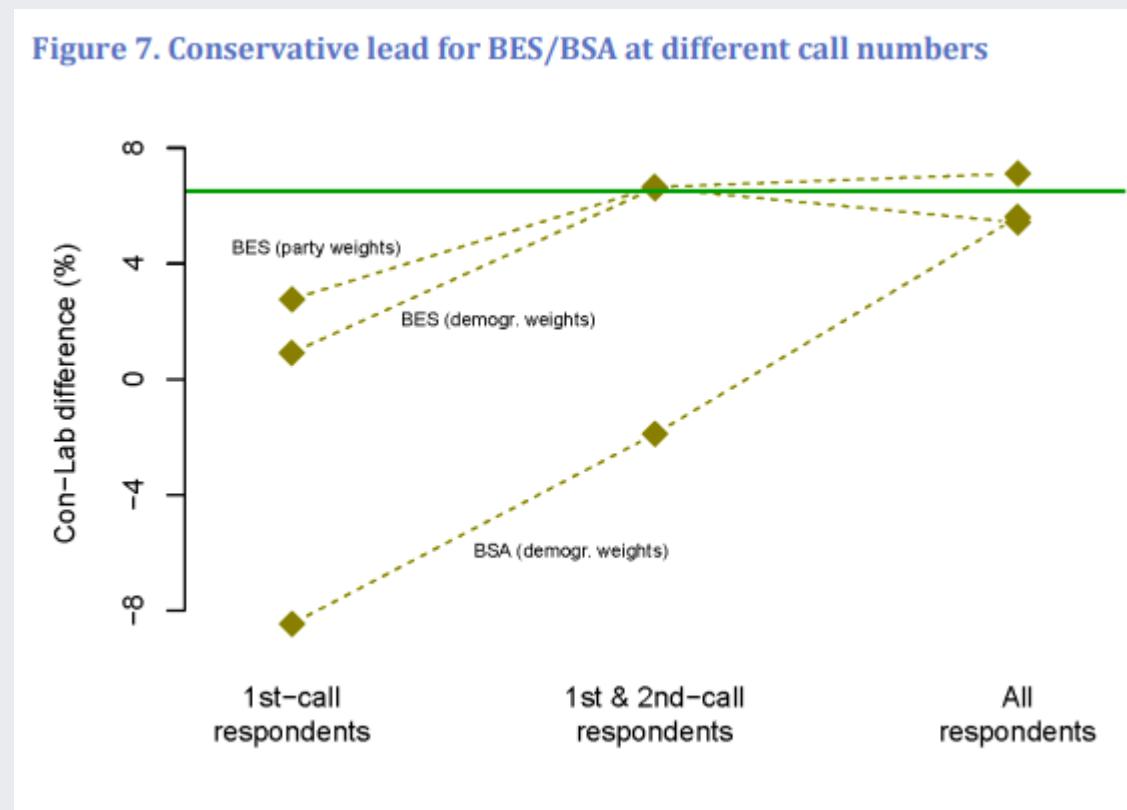
By David Cowling  
Editor, BBC Political Research Unit

© 17 May 2015

f Share

3 / 65

# La clave estaba en la representatividad



Sturgis *et al.*, 2016

# Diagnosticar y ajustar

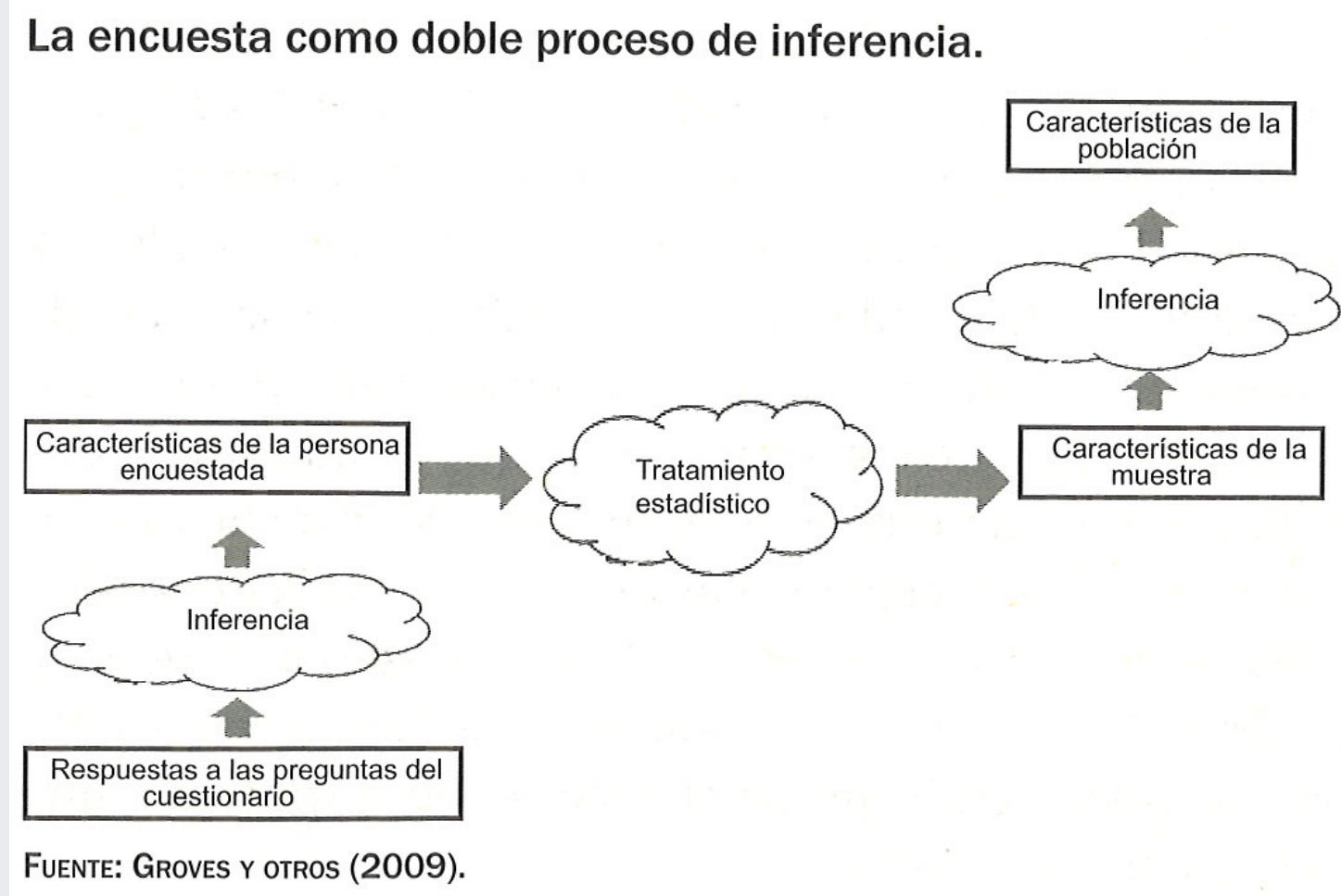
- Encuestas, muestreo y otros problemas
- Encuestas electorales y muestreo
- Análisis de representatividad
  - **Práctica:** análisis de representatividad
- Ajustes en las encuestas
- Generar un peso paso a paso con R
  - **Práctica:** ponderar una muestra

# Encuestas, muestreo y otros problemas

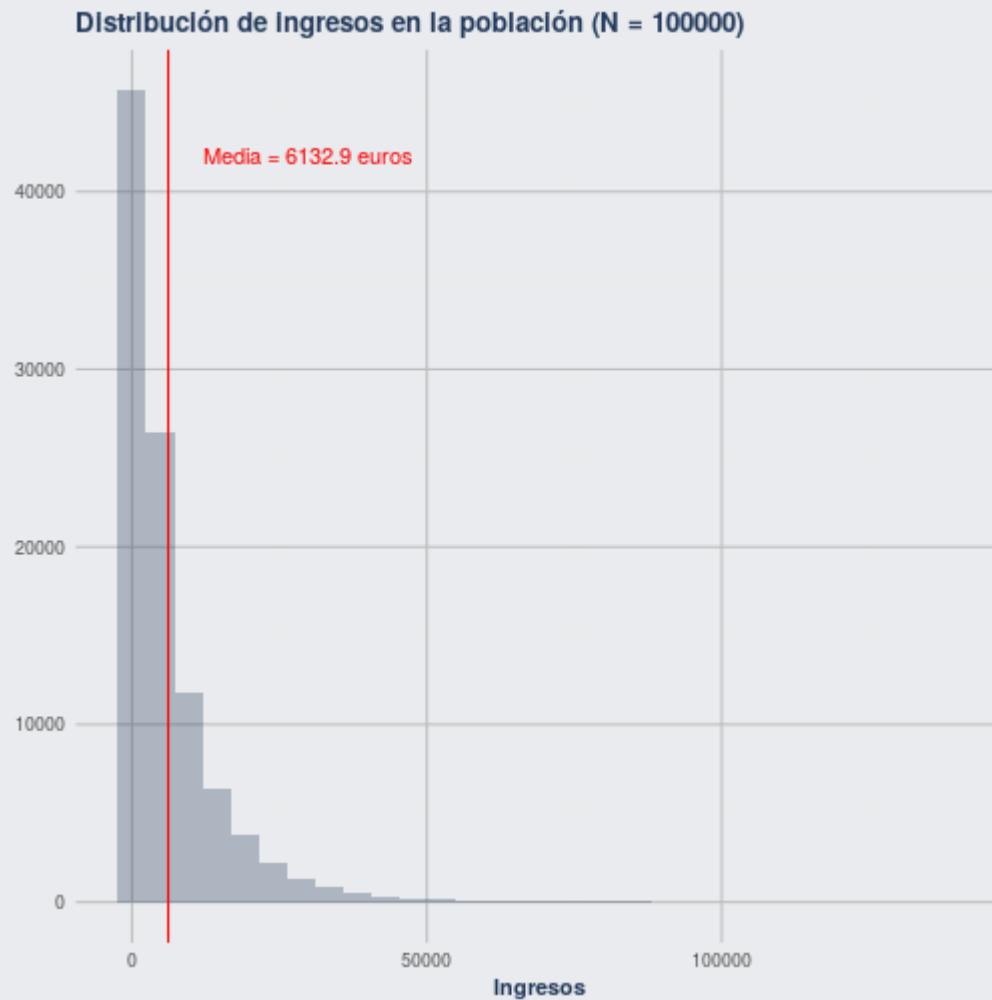
# Doble inferencia

La encuesta como doble proceso de inferencia.

Font y Pasadas, 2016

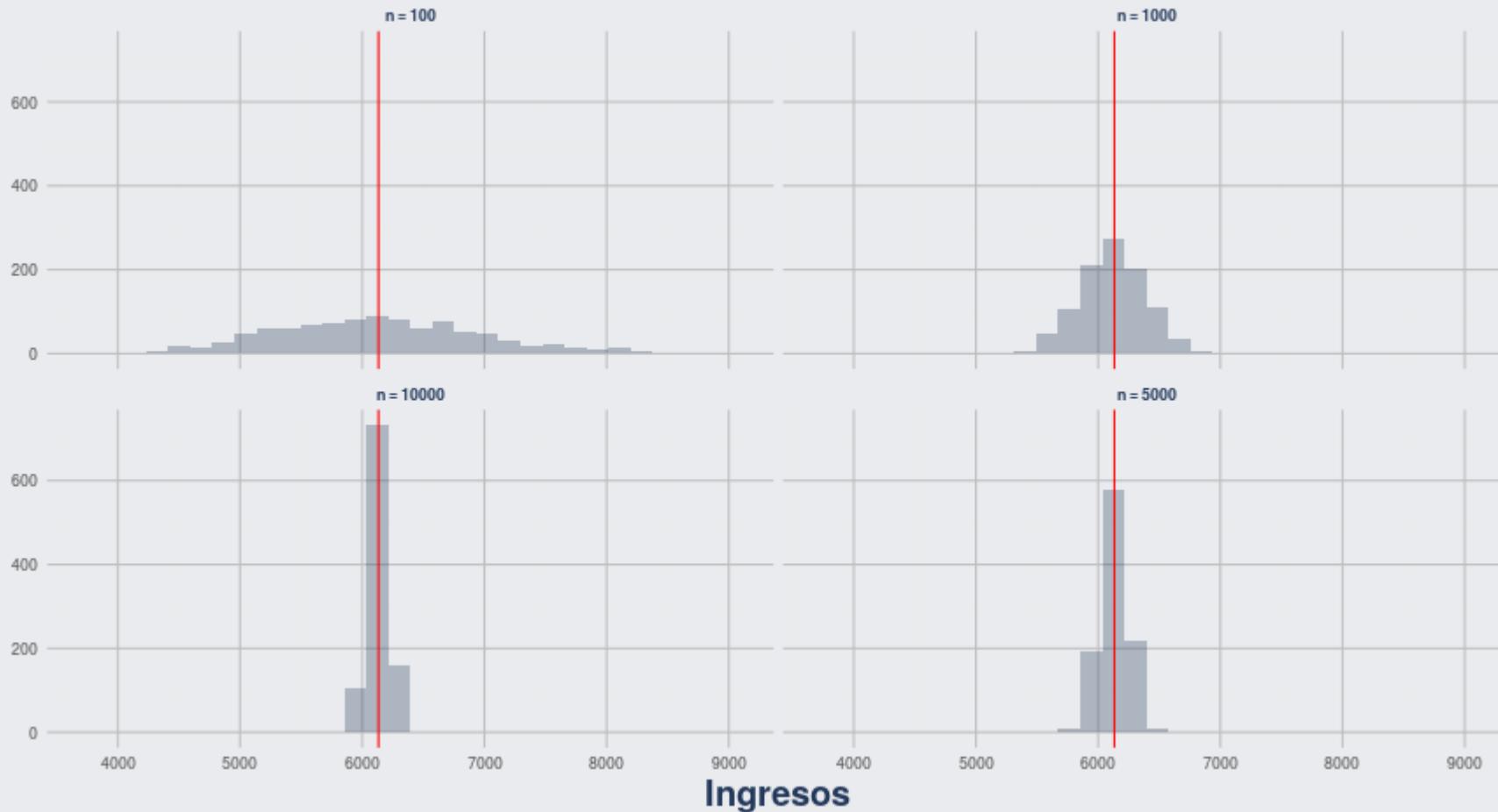


# Muestreo e inferencia



# Muestreo e inferencia

Distribución de las medias de 1000 muestras



# Varianza y sesgo

# Muestreo ideal (probabilístico)

- Existe un **marco muestral** en el que están listados *todos* los elementos de la población.
- El marco muestral contiene **información auxiliar** como sexo, edad o ingresos. Estas variables son útiles para estratificar la muestra.
- Todos los elementos seleccionados en la muestra **responden** a la encuesta.
- **Procedimiento:** todos los casos tienen una probabilidad  $\frac{n}{N}$  de ser elegidos. Se seleccionan  $n$  casos a partir de números aleatorios.

Nombre	Sexo	Edad	Población	Teléfono
Fernández Fernández, Antonio	H	25	Madrid	666534
Pérez Pérez, Clara	M	26	Madrid	6663245
González González, María	M	47	Fuenlabrada	9234534

# Muestreo real

¿Existe en España un **marco muestral** de la población general al que tengan acceso las organizaciones dedicadas a la investigación social?

# Muestreo real

¿Existe en España un **marco muestral** de la población general al que tengan acceso las organizaciones dedicadas a la investigación social?

- Existen registros como el **padrón de población** (INE), pero las empresas de investigación sólo pueden acceder a los datos anonimizados.

# Muestreo real

¿Existe en España un **marco muestral** de la población general al que tengan acceso las organizaciones dedicadas a la investigación social?

- Existen registros como el **padrón de población** (INE), pero las empresas de investigación sólo pueden acceder a los datos anonimizados.

¿Todos los elementos de la muestra **responden** a la encuesta?

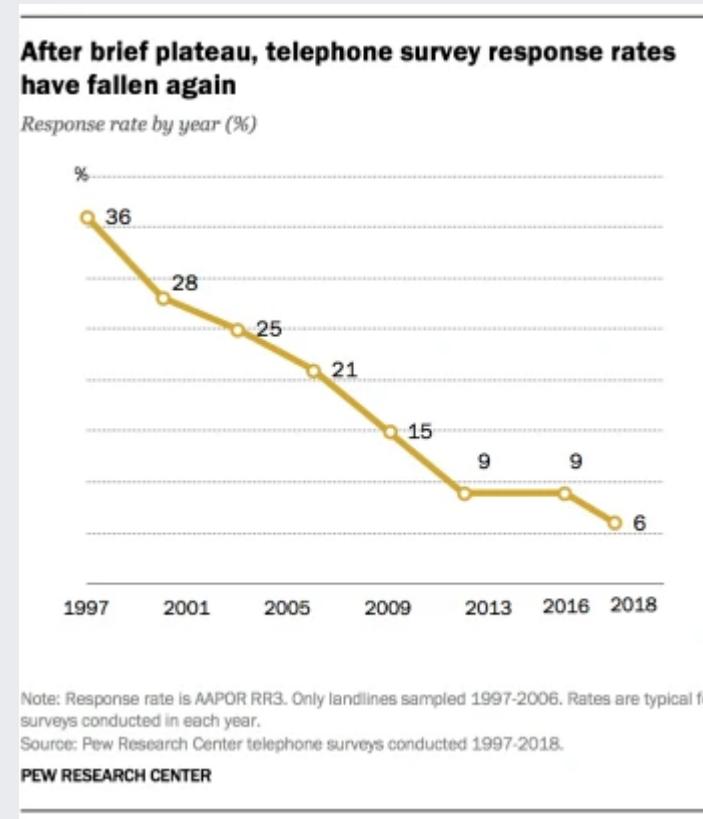
# Muestreo real

¿Existe en España un **marco muestral** de la población general al que tengan acceso las organizaciones dedicadas a la investigación social?

- Existen registros como el **padrón de población** (INE), pero las empresas de investigación sólo pueden acceder a los datos anonimizados.

¿Todos los elementos de la muestra **responden** a la encuesta?

- No, las tasas de respuestas en encuestas telefónicas pueden llegar a estar **por debajo del 10%** (Kennedy y Hartig, 2019).



# Los problemas del muestreo probabilístico

Problemas de las **encuestas probabilísticas**:

- Impedimentos **técnicos**. No existe marco muestral, está incompleto o no es posible acceder. Posible sesgo provocado por la no respuesta.
- Mayor **coste** que un muestreo no probabilístico. Necesidad de extender el trabajo de campo en encuestas telefónicas y personales.
- Necesita más **tiempo** para recoger los datos. Los hogares que no responden necesitan ser contactados en diferentes horarios por un período de tiempo suficiente.

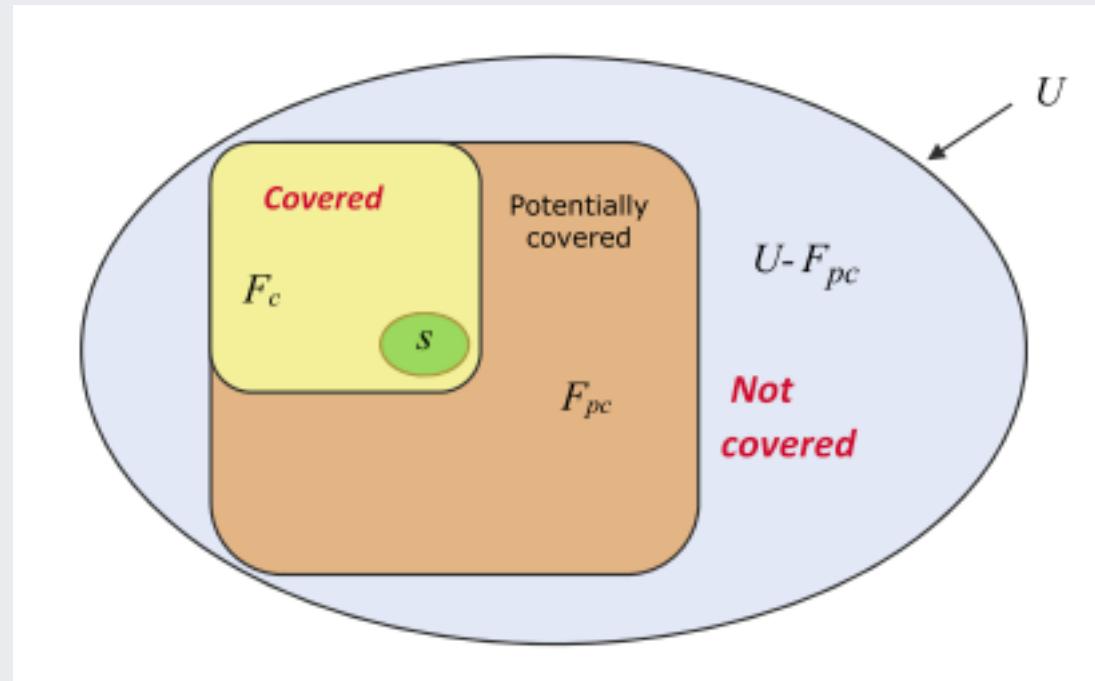
# Alternativas

- **Muestreo probabilístico combinado con no probabilístico.** Una parte de la selección de las unidades muestrales se realiza utilizando procedimientos probabilísticos. En el último paso se utilizan cuotas para elegir a la persona entrevistada.
  - **Muestro de hogares** se realiza a partir de rutas aleatorias o de número de teléfono (fijos y móviles) generados de forma aleatoria (RDD).
  - Las **cuotas** se establecen a partir de los datos poblacionales. Se utilizan variables como sexo y edad. Cada zona geográfica tiene asignadas unas cuotas. Al finalizar el trabajo de campo el perfil de sexo y edad de la muestra coincidirá con el de la población.
- **Muestreo no probabilístico.** No se conoce la probabilidad de selección de los elementos de la población (Baker *et al.*, 2013).
  - Encuestas por **cuotas** o *river sampling*. Paneles de internautas reclutados a partir de captación activa o pasiva.
  - **Inferir a partir de muestras no probabilísticas.** Procedimientos estadísticos o ajustes para mejorar la representatividad de la muestra e inferir.

# Problemas asociados a las muestras no probabilísticas

- Encuesta a partir de un **panel online** de voluntarios para conocer **intención de voto** de la población general.
  - Sesgo de **cobertura**. Algunos elementos de la población no tienen ninguna posibilidad de ser elegidos. Por ejemplo, una parte importante de las personas mayores de 65 años no acceden con regularidad a internet. Este problema también puede afectar a las muestras probabilísticas.
  - Sesgo de **autoselección**. La entrada en el panel está determinada por los propios usuarios, que deciden unirse, en ocasiones porque reciben una compensación económica a cambio de cada encuesta completada.
  - Sesgo de **no respuesta**. Todos los panelistas invitados no completarán la encuesta. Esto puede ser un problema si aquellos que contestan son diferentes que los que no responden. Este problema también afecta a las muestras probabilísticas.

# Problemas asociados a las muestras no probabilísticas



# Muestreo en encuestas electorales

# Muestreo en encuestas electorales

Las encuestas se pueden diferenciar según el modo de muestreo elegido.

- Las encuestas realizadas para el análisis político y electoral suelen partir de **muestras probabilísticas/no probabilísticas o no probabilísticas**. Se ha demostrado que algunas muestras no probabilísticas han sido más precisas al estimar el resultado electoral (Baker *et al.*, 2013).
- En España las encuesta **telefónicas utilizan cuotas de sexo y edad**. Desde hace unos años algunas empresas han dejado el sistema de cuotas. Algunas encuestas de panel online utilizan **cuotas más exhaustivas**.
- En algunas encuestas telefónicas se especifica el **uso de teléfonos fijos y móviles**. En otros casos no se conoce.
- El muestreo suele estar **estratificado por comunidad y tamaño de hábitat**. En algunas encuestas telefónicas se utiliza el municipio como punto muestral.

# Encuestas electorales en España

Organización	Muestreo	Descripción
CIS (Barómetros)	Prob./No prob.	Presencial. Selección de secciones censales de forma probabilística. Hogares con rutas aleatorias. Entrevistados por cuotas sexo y edad.
CIS (ESE)	Probabilístico	Presencial. Selección de personas directamente del padrón.
Metroscopia	Prob./No prob.	Telefónica (móviles). Muestra aleatoria de teléfonos. Sin cuotas.
GAD3	Prob./No prob.	Telefónica (fijos y móviles). Cuotas sexo, edad y región.

# Encuestas electorales en España

Organización	Muestreo	Descripción
IMOP insights	Prob./No prob.	Telefónica (fijos 55% y móviles 45%). Muestra aleatoria de teléfonos sin cuotas.
40dB	Prob./No prob.	Panel de internautas. Cuotas de sexo, edad, clase social, región y tamaño de hábitat.
GESOP	Prob./No prob.	Telefónica. Cuotas de sexo y edad.
NC Report	Prob./No prob.	Telefónica. Cuotas de sexo y edad.
Sociométrica	Prob./No prob.	Telefónica y panel de internautas. Cuotas sexo, edad y provincia.

# Análisis de representatividad

# La información de la ficha técnica

Estudiar **ficha técnica** de la encuesta:

- **Modo de administración** → Problemas de cobertura o respuesta.
  - Encuesta **telefónica**: móviles o fijos, generación aleatoria de números.
  - Encuesta **panel de internautas**: tipo de captación.
- Detalles del **muestreo** → Tipo de muestreo (variables que han tenido en cuenta en el diseño).
- **Fechas** del trabajo de campo → Procedimientos de campo (llamadas o visitas).
- **Tamaño** de la muestra y margen de error → Precisión de las estimaciones (probabilística o cuasiprobabilística).

# Análisis de representatividad por pasos

Comparar la distribución poblacional de ciertas variables con la muestra:

1. Identificar las variables del cuestionario que son susceptibles de tener equivalentes poblacionales.
2. Encontrar la fuentes de datos poblacionales:
  - Registros poblacionales (INE).
  - Encuestas con altos estándares metodológicos (ESE, CIS).
3. Determinar la equivalencia de los datos poblacionales. Tener en cuenta posibles errores de medición y comparabilidad de las variables (*p. ej.* estudios).
4. Calcular la diferencia entre las distribuciones de los datos poblaciones y las variables de la encuesta.
5. Calcular el error absoluto medio (*MAE*) de cada variable. En la que  $\hat{y}$  es el porcentaje de la categoría en la encuesta e  $y$  el porcentaje en la población.  $k$  es el número de categorías de la variable.

$$MAE = \frac{\sum |\hat{y} - y|}{k}$$

# Cálculo del MAE

Variable	Categoría	Muestra	Población	Dif.
Grupo de edad	18-29	22	18	4
Grupo de edad	30-44	25	24	1
Grupo de edad	45-64	23	23	0
Grupo de edad	65+	30	35	5

$$MAE = \frac{10}{4} = 2.5$$

# Práctica: análisis de representatividad

# Ajustes en las encuestas

# Una muestra de usuarios de Xbox

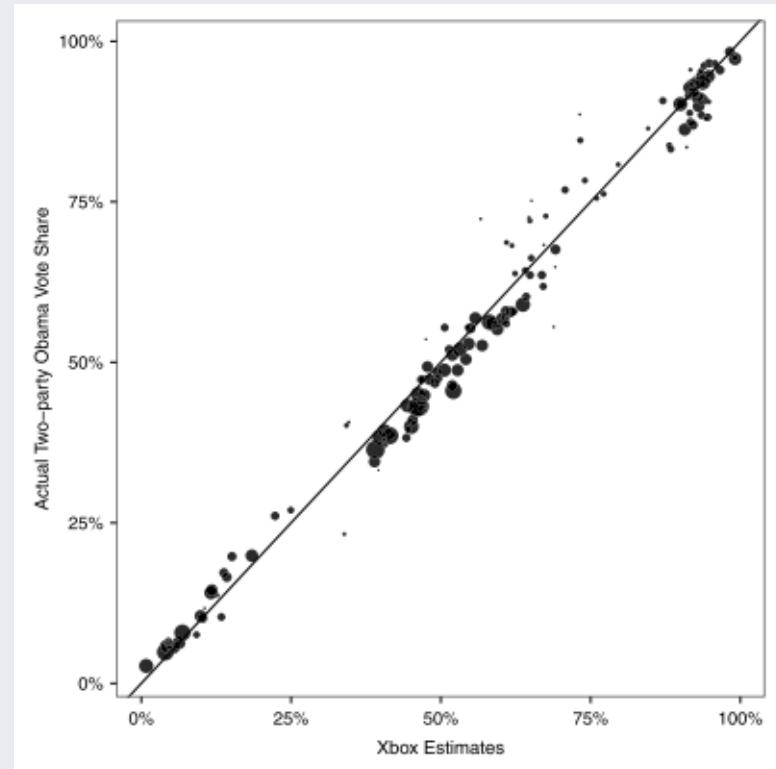
- En el marco de las elecciones presidenciales de 2012 en EE.UU Wang *et al.* realizaron una **encuesta (no probabilística)** a través de la plataforma de **usuarios de la Xbox**. La encuesta tuvo 345000 respuestas.

# Una muestra de usuarios de Xbox

- En el marco de las elecciones presidenciales de 2012 en EE.UU Wang *et al.* realizaron una **encuesta (no probabilística)** a través de la plataforma de **usuarios de la Xbox**. La encuesta tuvo 345000 respuestas.
- El nivel de **desajuste era notable**: 65% de los encuestados entre 18-29 años (19% en la población). Los hombres representaban el 93% (47% en la población). Mitt Romney sobrerepresentado.

# Una muestra de usuarios de Xbox

- En el marco de las elecciones presidenciales de 2012 en EE.UU Wang *et al.* realizaron una **encuesta (no probabilística)** a través de la plataforma de **usuarios de la Xbox**. La encuesta tuvo 345000 respuestas.
- El nivel de **desajuste era notable**: 65% de los encuestados entre 18-29 años (19% en la población). Los hombres representaban el 93% (47% en la población). Mitt Romney sobrerepresentado.
- Ajustaron la muestra utilizando **técnicas estadísticas**. El resultado fue similar al del promedio de encuestas tradicionales.



Wang *et al.*, 2015

# Pesos o ponderaciones

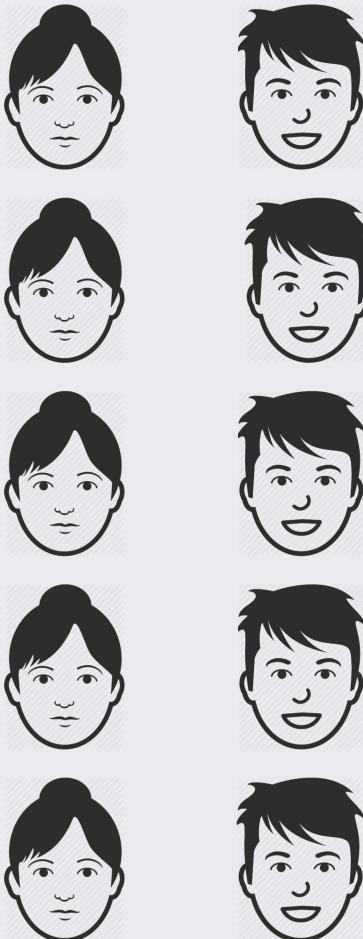
- Son **ajustes estadísticos** para preservar la representatividad de la muestra. Sirven para corregir desviaciones de la muestra debidas a la incidencia de los sesgos de cobertura, selección o no respuesta.
- A cada **caso le corresponde un peso** según sus características. El peso será más alto que la media si las características del elemento muestral están subrepresentadas en la muestra.
- Los pesos parten de un **cálculo general**:

$$w_k = \frac{N_k}{n_k}$$

- Los pesos tienen como misión corregir desviaciones de la muestra, pero también pueden tener un **efecto negativo en la varianza de las estimaciones**, que serán menos precisas.
- Algunas encuestas como la ESE **ofrecen pesos** en los datos que hacen públicos para que sean utilizados en los análisis.

# Población

Población de cinco  
**hombres** y cinco **mujeres**  
(censo)

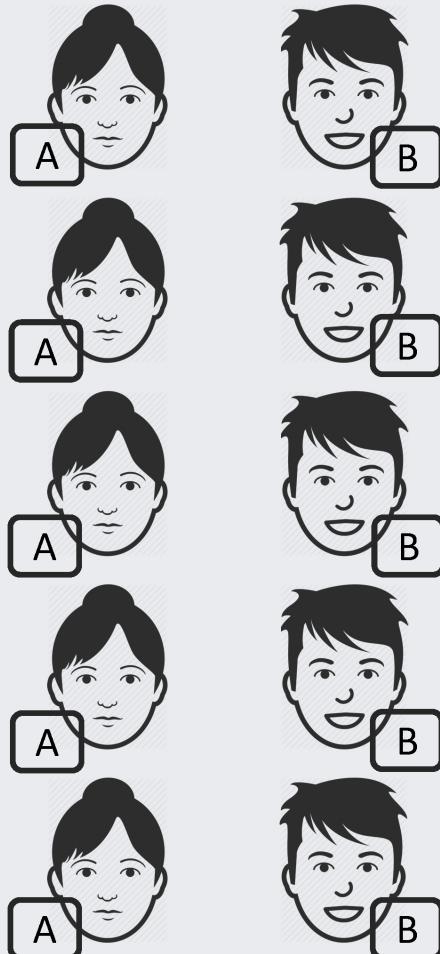


Investigar su **intención de voto**

# Población

Población (censo):

5 (50%) **Mujeres**  
5 (50%) **Hombres**



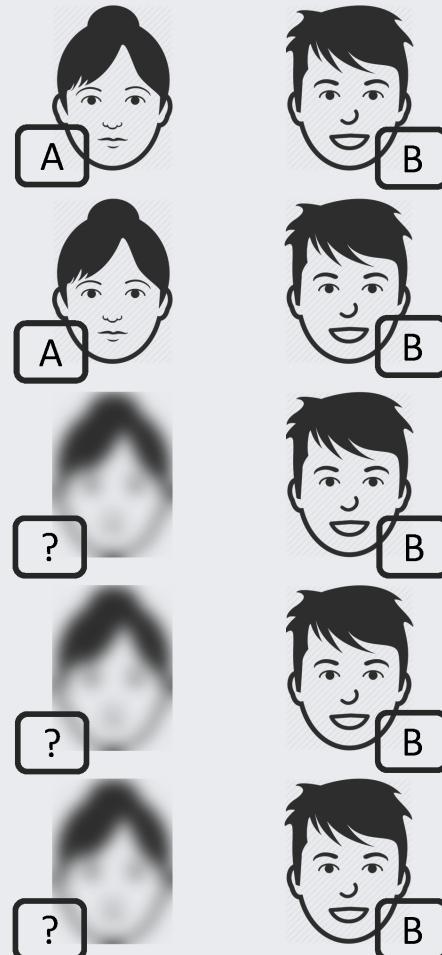
Población (modo Dios):

5 (50%) **Partido A**  
5 (50%) **Partido B**

# Incidencia de la no respuesta

Muestra final:

2 (29%) Mujeres  
5 (71%) Hombres



Muestra final:

2 (29%) Partido A  
5 (71%) Partido B

# Ponderación

Ponderación como:

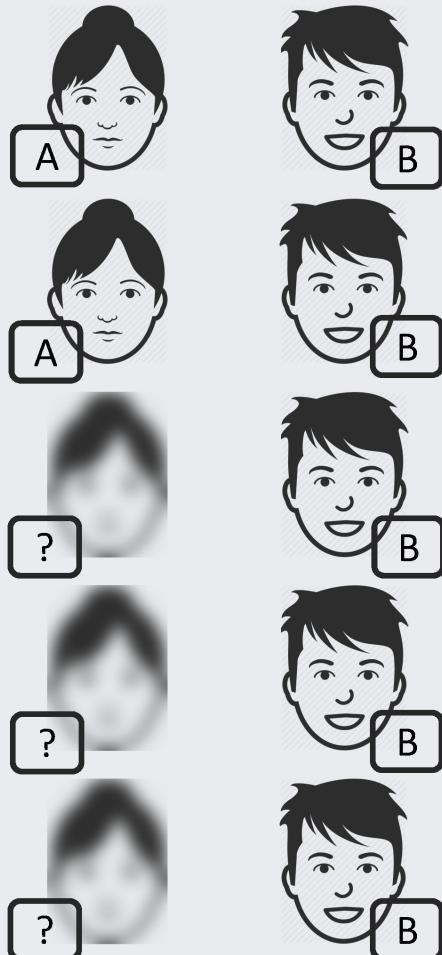
$$w_k = \frac{N_k}{n_k}$$

Para las **mujeres**:

$$w_m = \frac{5}{2} = 2.5$$

Para los **hombres**:

$$w_h = \frac{5}{5} = 1.0$$



Estimar voto con pesos:

Mujeres - Partido A:  $2 * 2.5 = 5$  (50%)

Mujeres - Partido B:  $0 * 2.5 = 0$  (0%)

Hombres - Partido A:  $0 * 1.0 = 0$  (0%)

Hombres - Partido B:  $5 * 1.0 = 5$  (50%)

# ¿Qué es necesario para que un peso funcione?

- Un peso funciona cuando la **variable auxiliar** que se utiliza para generarlo está **correlacionada con la probabilidad de responder y con la variable de interés**. En el ejemplo:
  - Sexo (variable auxiliar) estaba relacionado con la probabilidad de responder ( $H=100\%$ ;  $M=40\%$ ).
  - Sexo (variable auxiliar) correlacionaba de manera perfecta con el partido a votar (variable de interés).
- Las **variables auxiliares**. Estas variables son información que existe para los que participan en la encuesta y para los que no.
  - Puede ser a nivel **individual** (marco muestral) o a nivel **agregado** (totales poblacionales).
  - La información auxiliar disponible es **limitada** (INE, otras encuestas), lo que condiciona la capacidad de los ajustes.
  - Pensar **teóricamente** qué variables pueden explicar mejor la **probabilidad de responder** y la **variable de interés**.

# Tipos de ajustes

Según la función y la manera de generarlos hay diferentes tipos de pesos:

- Pesos de **selección**.
- Pesos de **no respuesta**.
- Pseudopesos a partir de una **encuesta de referencia**.
- Ajuste por modelos **MRP** o **superpoblación**.
- **Postestratificación y calibración.**

# Pesos de selección

- Se utilizan en encuestas cuando las **probabilidades de selección de los grupos son diferentes** debido al muestreo. También se computan para ajustar las probabilidades de selección dentro de hogares o conglomerados de tamaño desigual.
  - Por ejemplo, cuando en una encuesta a población general en España una comunidad autónoma está sobrerepresentada.
- **Necesario** → Este peso se calcula a través de las probabilidades de selección. Es necesario conocer, al menos, el total de casos de cada grupo de la población.

Grupo	N	n	Prob_sel	Peso_sel
Grupo1	1000	500	0.5	2
Grupo2	1000	100	0.1	10
Grupo3	1000	100	0.1	10

# Pesos de no respuesta

- Ajustan ante la pérdida de representatividad debido a la **no respuesta o la falta de cobertura**. Dos formas de calcularlos **determinista** o **probabilística**.
- En la forma **determinista** o *cell-weighting* se vuelve a realizar en cálculo de  $\frac{N_k}{n_k}$ .
- En el método **probabilístico** se realiza un **modelo de regresión logística** en el que la variable dependiente es la respuesta y las variables independientes son las variables auxiliares. A partir del modelo se predice la probabilidad de responder a la encuesta de cada caso. El inverso de esa probabilidad es el peso.
- **Necesario** → En el determinista es suficiente con tener los totales poblacionales de cada grupo. En el probabilístico hay que tener **información a nivel individual de los que no responden**.

Grupo	N	n	Prob_sel	Peso_sel	Resp	RespXPeso_sel	Peso_nr	Peso_final
Grupo1	1000	500	0.5	2	250	500	2.00	4.0
Grupo2	1000	100	0.1	10	90	900	1.11	11.1
Grupo3	1000	100	0.1	10	70	900	1.43	14.3

# Pseudopesos a partir de una encuesta de referencia

- Esta técnica se usa para **equilibrar muestras no probabilísticas**. Se trata de calcular la *pseudoprobabilidad* de que un elemento de la población se haya unido al panel de internautas y haya respondido a la encuesta.
- Para llevarla a cabo se utiliza una **encuesta de referencia** con estándares metodológicos más altos, que comparta algunas variables con la encuesta no probabilística. Estas dos encuestas se unen para **calcular la probabilidad de participar** en la encuesta mediante el uso de una regresión logística. El inverso de esa probabilidad es el peso de ajuste.
- **Necesario** → Una encuesta de referencia que sea de más calidad y que comparta parte de las variables con la encuesta no probabilística. Las variables compartidas deben explicar el mecanismo de participación.

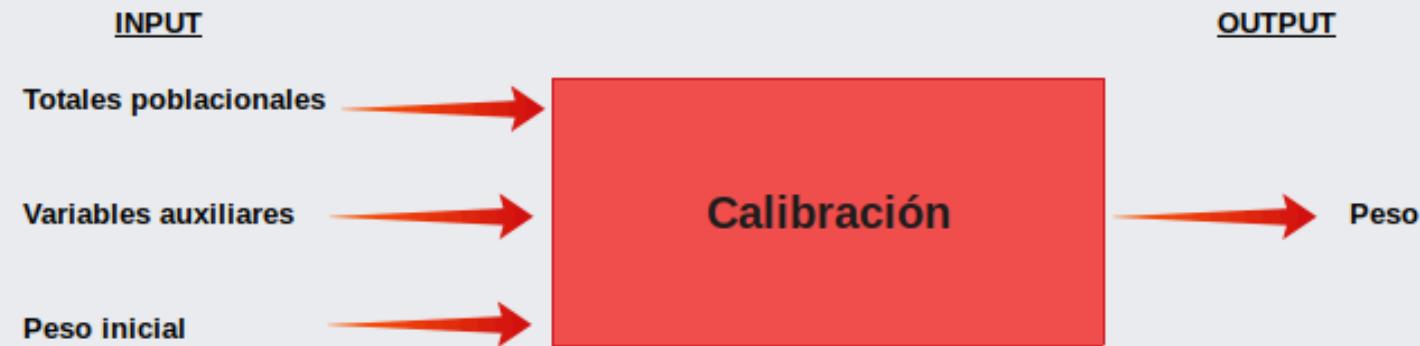
# Ajuste por modelos MRP y superpoblación

- En este caso no se trata de estimar un peso, sino de utilizar **un modelo para ajustar la muestra y predecir el valor de la variable de interés sin sesgo**.
- La técnica **MRP** (*multilevel regression and poststratification*) consiste en un modelo multinivel en el que para cada subgrupo se predice la variable dependiente en base a las características disponibles en la muestra y en los datos auxiliares. Las predicciones son ponderadas según las características del subgrupo. Se utiliza sobre todo para realizar estimaciones de voto para subgrupos o regiones con muestras relativamente pequeñas.
- En el **modelo de superpoblación** se predice la variable de interés para cada caso en la población a partir de un modelo ajustado en la muestra no probabilística.
- **Necesario** → Totales poblacionales a nivel de subgrupo para **MRP**. Datos poblacionales a nivel individual para los **modelos de superpoblación**.

# Postestratificación y calibración

- Se trata de **ajustes a través de modelos**. En ambos métodos los *inputs* son los datos de la encuesta y los totales poblacionales. El modelo se encarga de generar unos pesos que fuerzan a la distribución de la muestra a ser idéntica a la de la población.
- En la **postestratificación** se incluyen todas las interacciones entre las variables auxiliares.
- En la **calibración** el ajuste se produce sin tener en cuenta las interacciones entre las variables. Es una técnica más flexible que la postestratificación.
- **Necesario** → Los totales poblacionales de las variables por separado en la **calibración** y de los cruces de variables en la **postestratificación**.

# Esquema de la calibración



# La diferencia entre postestratificación y calibración

- Se realiza una **calibración** y **postestratificación** utilizando las variables sexo y edad en grupos.

	En %		
	Población	Calibración	Postestrat.
Sexo			
Hombre	48	48	48
Mujer	52	52	52
Edad			
18-29	20	20	20
30-44	25	25	25
45-59	24	24	24
60+	31	31	31

# La diferencia entre postestratificación y calibración

En %

## Población

	18-29	30-44	45-59	60+
Hombre	11	12	11	14
Mujer	9	13	13	17

## Calibración

	18-29	30-44	45-59	60+
Hombre	9	14	8	17
Mujer	11	11	16	14

## Postestratificación

	18-29	30-44	45-59	60+
Hombre	11	12	11	14
Mujer	9	13	13	17

Generar un peso paso a paso con 

# Un ejemplo paso a paso para construir un peso en

Para calcular el peso por calibración vamos a utilizar el paquete `survey`. Para la manipulación de datos `tidyverse`.

1. Generar un vector con nombres de **totales poblacionales**.
2. Preparar las **variables auxiliares** en la encuesta y el **peso inicial**.
3. Calcular los pesos mediante **calibración**.
4. **Evaluar** la calibración.
5. **Escalar** los pesos para que tengan media 1.

# Datos

- Utilizo los datos gss\_cat de tidyverse para ponderar la muestra utilizando la variable race y ajustar la estimación de partyid.

```
gss <- gss_cat %>%
  filter(year == 2014) %>%
  select(race, partyid)

glimpse(gss)
```

```
## Observations: 2,538
## Variables: 2
## $ race      <fct> White, White, White, White, White, White, White, ...
## $ partyid   <fct> "Not str republican", "Not str republican", "Strong repu...
```

# 1. Generar un vector de totales poblacionales

- Los totales poblacionales extraídos de la *American Community Survey*.
- Crear un vector con **nombres de los totales**. En ese vector:
  - El primer elemento del vector siempre lleva el nombre (`Intercept`) y corresponde con el **total poblacional**.
  - De cada variable que se incluya en la calibración se **retira una categoría**.
  - El **nombre de los elementos** es una combinación del nombre de la variable y la etiqueta de la categoría.

```
totales_pobla <- c("(Intercept)" = 244821329, "raceBlack" = 29330377, "raceWhite" = 196766012)

totales_pobla
```

```
## (Intercept) raceBlack raceWhite
## 244821329 29330377 196766012
```

## 2. Preparar las variables auxiliares en la encuesta y el peso inicial

- Revisar que las **categorías de la variable** coinciden con las del vector de totales poblacionales.
- Puede ocurrir que algunas categorías tengan un **número reducido de casos** ( $n < 10$ ). En esos casos cabe la posibilidad de combinar varias categorías para evitar pesos extremos.

```
sjmisc::frq(gss$race)

## 
## x <categorical>
## # total N=2538  valid N=2538  mean=2.64  sd=0.66
## 

##           val   frq raw.prc valid.prc cum.prc
## Other      262   10.32    10.32    10.32
## Black      386   15.21    15.21    25.53
## White     1890   74.47    74.47   100.00
## Not applicable    0    0.00      0.00   100.00
## <NA>        0    0.00      NA       NA
```

## 2. Preparar las variables auxiliares en la encuesta y el peso inicial

- En la calibración se incluye un peso inicial que es el **peso de selección elevado al total poblacional**. En caso de que no haya peso de selección, se incluye un peso que es el total poblacional entre el número de casos.

```
gss$peso_ini <- 244821329/nrow(gss)
```

# 3. Calcular los pesos mediante calibración

- Declarar el **diseño** de la encuesta con `svydesign()`. Esta función sirve para registrar cuál es el diseño de la encuesta:
  - `ids` sirve para declarar la **variable conglomerados** en el caso de que la muestra sea conglomerada. En caso contrario utilizar `~ 0`.
  - `weights` es el nombre de la variable de los pesos de selección o peso inicial (`~ peso_ini`).

```
gss_svy <- svydesign(ids = ~ 0, weights = ~ peso_ini, data = gss)
```

# 3. Calcular los pesos mediante calibración

- Generar los **pesos de calibración** con la función `calibrate()`:
  - `design` es el objeto de tipo `survey` que se genera en el paso anterior.
  - `formula` son las variables incluidas en la calibración antecedidas por `~`.
  - `population` es el vector de totales poblacionales.
  - `calfun` es el tipo de calibración, se utiliza `logit` para evitar que se den pesos negativos, por eso es obligatorio establecer `bounds` de 0 hasta un valor elevado.

```
gss_cal <- calibrate(design = gss_svy,  
                      formula = ~ race,  
                      population = totales_pobla,  
                      calfun = "logit",  
                      bounds = c(0, 999999))
```

# 4. Evaluar la calibración

- Combinar los **pesos con la encuesta**.
- Comprobar los **máximos y los mínimos** del peso, para evaluar si hay que recortar.

```
peso_cal <- weights(gss_cal)
gss$peso_cal <- peso_cal
sjmisc::descr(gss$peso_cal)

## 
## ## Basic descriptive statistics
##
##   var     type label      n NA.prc      mean       sd      se      md    trimmed
##   dd numeric    dd 2538      0 96462.31 13109.73 260.22 104109 98622.07
##                   range skew
##   32639.74 (71469.26-104109) -1.15
```

# 4. Evaluar la calibración

- Comprobar que el peso está funcionando comparando la **variable auxiliar** con los **totales poblacionales**.

```
totales_pobla
```

```
## (Intercept) raceBlack raceWhite
## 244821329 29330377 196766012
```

```
sjmisc::frq(gss$race, weights = gss$peso_cal)
```

```
##
## xw <categorical>
## # total N=244821337 valid N=244821337 mean=2.73 sd=0.59
##
##   val     frq  label raw.prc valid.prc cum.prc
##   Other 18724946 <none>    7.65      7.65    7.65
##   Black 29330379 <none>   11.98     11.98   19.63
##   White 196766012 <none>   80.37     80.37  100.00
##   <NA>      0     NA    0.00       NA      NA
```

# 4. Evaluar la calibración

- Evaluar el efecto sobre una o varias variables relevantes de la encuesta.

```
sjmisc::frq(gss$partyid)
```

```
##  
## x <categorical>  
## # total N=2538  valid N=2538  mean=7.17  sd=2.10  
  
##           val frq raw.prc valid.prc cum.prc  
##       No answer  25    0.99     0.99    0.99  
##       Don't know   1    0.04     0.04    1.02  
##       Other party  62    2.44     2.44    3.47  
##       Strong republican 245    9.65     9.65   13.12  
##       Not str republican 292   11.51    11.51   24.63  
##       Ind,near rep 249    9.81     9.81   34.44  
##       Independent 502   19.78    19.78   54.22  
##       Ind,near dem 337   13.28    13.28   67.49  
##       Not str democrat 406   16.00    16.00   83.49  
##       Strong democrat 419   16.51    16.51  100.00  
##       <NA>      0    0.00      NA      NA
```

# 4. Evaluar la calibración

```
sjmisc::frq(gss$partyid, weights = gss$peso_cal)

## 
## xw <categorical>
## # total N=244821339  valid N=244821339  mean=7.10  sd=2.10
## 

##          val     frq   label raw.prc valid.prc cum.prc
## No answer 2307941 <none>    0.94      0.94    0.94
## Don't know 104109  <none>    0.04      0.04    0.99
## Other party 6150942 <none>    2.51      2.51    3.50
## Strong republican 25081362 <none>   10.24     10.24   13.74
## Not str republican 29628997 <none>   12.10     12.10   25.84
## Ind,near rep 25012718 <none>   10.22     10.22   36.06
## Independent 47993276 <none>   19.60     19.60   55.66
## Ind,near dem 32266014 <none>   13.18     13.18   68.84
## Not str democrat 38264856 <none>   15.63     15.63   84.47
## Strong democrat 38011124 <none>   15.53     15.53 100.00
## <NA>           0       NA     0.00      NA      NA
```

# 5. Escalar el peso

- El último paso consiste en **escalar el peso para que tenga media 1**. En raras ocasiones se mantiene un peso que eleve al total poblacional.

```
gss$peso_cal <- gss$peso_cal/mean(gss$peso_cal)  
sjmisc::descr(gss$peso_cal)
```

```
##  
## ## Basic descriptive statistics  
##  
## var      type   label    n  NA.prc  mean     sd  se    md  trimmed          range  
##   dd numeric    dd 2538       0     1 0.14  0 1.08    1.02 0.34 (0.74-1.08)  
##   skew  
## -1.15
```

# En resumen

- El **diseño y la recogida** de datos pueden afectar a la **representatividad** de la muestra. Falta de cobertura de la población, no repuesta o selección.
- La **falta de representatividad no invalida el uso de los datos**, pero hay que tenerlo en cuenta a la hora de llevar a cabo el análisis.
- El **análisis de representatividad** permite establecer un diagnóstico de los problemas de la muestra. Cuantos más datos auxiliares, más completo será el análisis de representatividad.
- Los **ajustes** son necesarios para garantizar el proceso de inferencia de la muestra a la población. El uso de pesos es necesario, algunas encuestas incluyen sus propios ajustes.
- El funcionamiento de los ajuste es parcial y dependen de lo **adecuadas que sean las variables auxiliares**. Las variables auxiliares deben explicar el mecanismo de respuesta y estar correlacionadas con las variables de interés.

# Práctica: ponderar una muestra

# Pablo Cabrera Álvarez (USAL)

pablocal@usal.es |  |  @pablocalv

Materiales de la sesión en

[https://github.com/pablocal/course\\_2019\\_polling\\_weights](https://github.com/pablocal/course_2019_polling_weights)

# Bibliografía

- Atkeson, L. R., & Alvarez, R. M. (Eds.). (2018). *The Oxford Handbook of Polling and Survey Methods*. Oxford University Press.
- Bethlehem, J., & Cobben, F. (2013, August). Web Panels for Official Statistics?. In - Proceedings 59th ISI World Statistics Congress (Vol. 25).
- Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys* (Vol. 568). John Wiley & Sons.
- De Leeuw, E. D., Hox, J., & Dillman, D. (2012). *International handbook of survey methodology*. Routledge.
- Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.
- Groves, R. M., Biemer, P. P., Lyberg, L. E., Massey, J. T., Nicholls, W. L., & Waksberg, J. (Eds.). (2001). *Telephone survey methodology* (Vol. 328). John Wiley & Sons.
- Lee, S., & Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319-343.
- Leslie Kish. (1995). *Survey sampling* (Vol. 60). Wiley-Interscience.

# Bibliografía

- Kohut, A., Keeter, S., Doherty, C., Dimock, M., & Christian, L. (2012). Assessing the representativeness of public opinion surveys. Washington, DC: Pew Research Center.
- Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4), 375-385.
- Särndal, C. E., & Lundström, S. (2005). Estimation in surveys with nonresponse. John Wiley & Sons.
- Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., . . . & - - Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through
- Sturgis, P., Nick, B., Mario, C., Stephen, F., Jane, G., Jennings, W., . . . & Patten, S. (2016). Report of the inquiry into the 2015 British general election opinion polls.