OPSO79-1-UCSH2021

Inferencia desde muestras complejas

26/11/2021

Muestras complejas en R

Inferencia a la población

El desafío de la inferencia

La reducción de costos y esfuerzos que implica estudiar una población mediante una muestra, se compensa con el costo de la "incertidumbre" o "imprecisión".

La estadística nos da elementos para conocer y manejar esta incertidumbre.

Desde nuestra muestra vamos a estimar o inferir un valor aproximado del parámetro de la población.

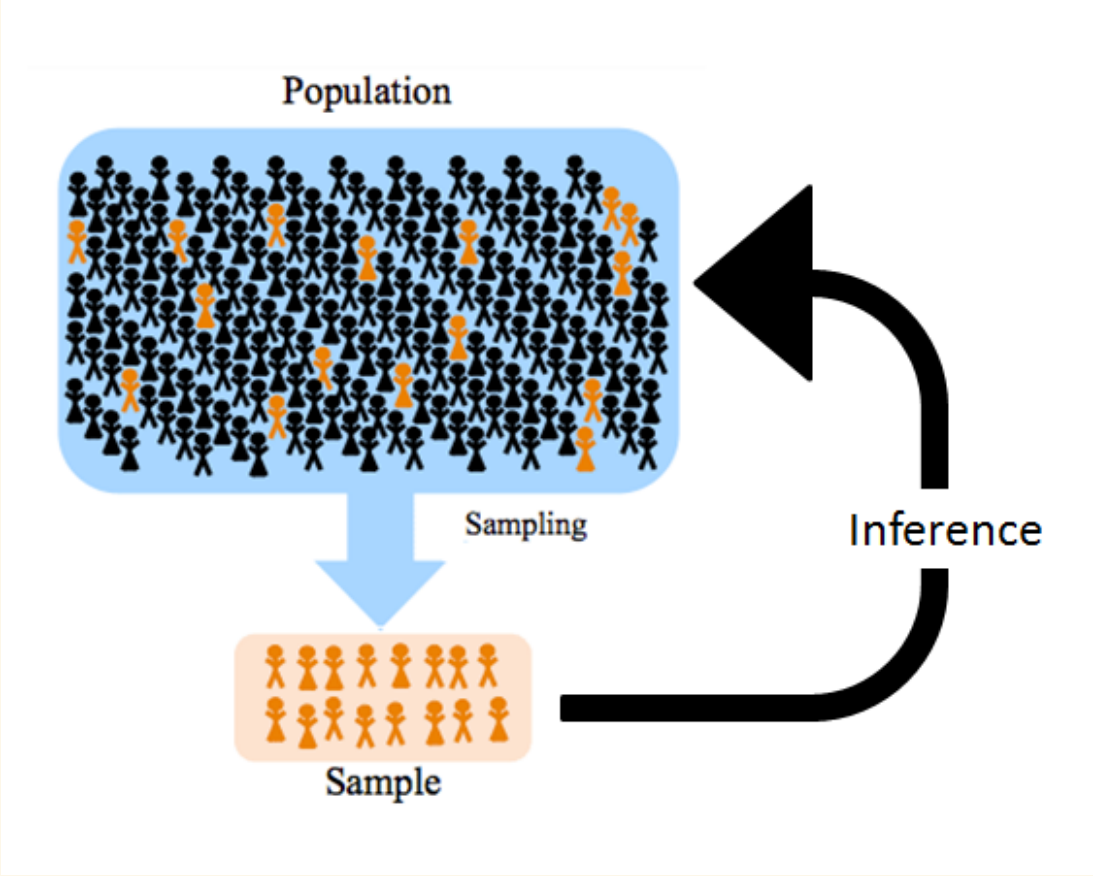
Al hacer este proceso, no solo ocuparemos estimaciones puntuales (como medias, quntiles, medianas, etc.)

También tendremos que calcular la precisión de estas estimaciones

Todo estimador está compuesto por dos elementos

- Estimación puntual
- Precisión (ci, se, var, cv)

El desafío de la inferencia



La forma incorrecta

Para estimar el valor del parámetro poblacional existen dos alternativas:

(i) definir un estadístico como estimación del parámetro poblacional (estimación puntual)

(ii) establecer en torno a un estadístico un intervalo de confianza para estimar en términos probabilísticos el parámetro.

La segunda alternativa es la más apropiada. Sin embargo, comprendamos la simplicidad de la primera.

Para la estimación puntual, solo necesitaremos el peso de cada unidad de nuestros datos (`weight`)

Este **ponderador** o **factor de expansión (FE)** indica a cuántas unidades representa cada elemento de la muestra.

Los ponderadores suman 1, mientras que los FE suman el tamaño de la población.

Abramos R

Una vez más, trabajaremos con la Encuesta Nacional de Empleo.

¿Cuántes personas ocupadas existían en Chile en trimestre Enero-Marzo 2021? Consultar acá

8.148.210 ocupados: 4.826.060 hombres y 3.322.150 mujeres.

¿Como podemos reproducir este resultado desde la base de datos?

```
# Descargar la base de datos
ene <- read.csv(file = "https://www.ine.cl/docs/default-source/ocupacion
```

```
ene %>% filter(activ==1) %>%
  group_by(sexo) %>% summarise(ocupados=n())
```

```
## # A tibble: 2 x 2
##    sexo ocupados
##   <int>    <int>
## 1     1    19443
## 2     2    14716
```

Abramos R

No nos da lo mismo, ya que solo estamos considerando a los elementos de la muestra.

Para estimar el total, tendremos que utilizar la variable factor de expansión.

¿Como se comporta esta variable?

```
summary(ene$fact_cal)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    8.106   85.573  136.782  216.799  239.963 4107.542
```

```
#sum(ene$fact_cal) ¿Cuánto debería sumar aproximadamente?
```

```
## [1] 19595837
```

Abramos R

Ahora filtramos para dejar solo los ocupados y agrupamos por sexo.

```
ene %>% filter(activ==1) %>%
  group_by(sexo) %>%
  summarise(ocupados=sum(fact_cal))
```

```
## # A tibble: 2 x 2
##    sexo ocupados
##   <int>    <dbl>
## 1     1 4826056.
## 2     2 3322150.
```

Y el total de ocupados

```
sum(ene[ene$activ==1,]$fact_cal,na.rm = TRUE)
```

```
## [1] 8148206
```

Sobre el uso de estimaciones puntuales

Si bien es una mala práctica, tiene buen rendimiento y se usa en estudios descriptivos.

El mismo INE presenta las estimaciones puntuales sin advertencia de su precisión*.

Medir la precisión de la estimación

Si solo nos interesa la estimación puntual, podríamos simplemente usar el peso de cada caso y olvidarnos del resto.

Sin embargo, debemos ser capaces de conocer la precisión de nuestras estimaciones y poder determinar, al menos, si son significativamente diferentes de cero.

Para esto debemos suponer cosas, conocer la error estándar de nuestra variable, el nivel de confianza con el que estamos trabajando y otros elementos del diseño muestral.

R tiene unos paquetes que nos simplificarán la vida.

La forma correcta: `survey` y `srvyr`

Para trabajar con muestras complejas en R son necesarios dos paquetes:

`survey`

`srvyr`

El primero fue creado por Thomas Lumley.

El segundo es su adaptación por terceros para que dialogue con la gramática de `dplyr` y los `pipes`.

```
## Opciones generales
options(survey.lonely.psu = "certainty" )

## Crear objeto tbl_svy
ds <- data %>% as_survey_design(ids = conglomerados,  ## ids=1 (no hay c
                                strata = estratos,
                                weights = pesos)
```

La forma correcta: `survey` y `srvyr`

Tenemos que poner atención a las siguientes variables, que hasta ahora dejábamos de lado:

```
ene %>% select(ano_trimestre,mes_central,idrph,estrato,conglomerado,fact
```

```
## 'data.frame':    90387 obs. of  6 variables:
##  $ ano_trimestre: int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ..
##  $ mes_central  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ idrph        : int  544011 72737011 736011 719011 15805111 15804111 1580
##  $ estrato      : int  5121 5121 5121 5121 10100111 10100111 10100111 10100
##  $ conglomerado : num  5.80e+12 5.80e+12 5.80e+12 5.80e+12 1.01e+09 ...
##  $ fact_cal     : num  295 261 270 295 126 ...
```

```
library(survey)
library(srvyr)
```

La forma correcta: `survey` y `srvyr`

Crear objeto survey con la ENE

Todas las recodificaciones y ediciones hacerlas antes de crear el objeto `survey`.

- Las variables con las que se harán agrupamientos deben mutarse a formato `factor`.

```
ene$activ<-as.factor(ene$activ) ## ojo
ene$sexo<-as.factor(ene$sexo) ## ojo
ene$categoria_ocupacion<-as.factor(ene$categoria_ocupacion)  ## ojo
```

Crear el objeto `survey`

```
ds<- ene %>% as_survey_design(ids = conglomerado,
                              strata = estrato,
                              weights = fact_cal)
```

Opciones generales, quedan definidas para toda la sesión de R (como cargar un paquete)

```
options(survey.lonely.psu = "certainty" )
```

La forma correcta: `survey` y `srvyr`

```
class(ds) ## Consultar tipo de objeto
```

```
## [1] "tbl_svy"        "survey.design2" "survey.design"
```

Contar casos por categoría de respuesta

```
# Ocupados - Desocupados - Fuera de la FT
ds %>% group_by(activ) %>% summarise(trabajadores=survey_total(na.rm = T
```

```
## # A tibble: 4 x 3
##   activ trabajadores trabajadores_se
##   <fct>        <dbl>           <dbl>
## 1 1         8148206.          91044.
## 2 2          941088.          26958.
## 3 3         6763752.          70928.
## 4 <NA>      3742791.          62359.
```

La forma correcta: `survey` y `srvyr`

Error estandar indica la variabilidad de las medias muestrales.

Tiende a disminuir cuando aumenta el tamaño de las muestras.

$$se = sd/\sqrt{n}$$

Como la desviación estándar de la población rara vez se conoce, el error estándar de la media suele estimarse como la desviación estándar de la muestra dividida por la raíz cuadrada del tamaño de la muestra.

Con el error estandar podemos obtener los intervalos de confianza

$$[\overline{x} + z_{a/2}\frac{sd}{\sqrt{n}}, \overline{x} - z_{a/2}\frac{sd}{\sqrt{n}}]$$

Survey lo hace por nosotros...

La forma correcta: survey y srvyr

```
ds %>% group_by(activ) %>%
  summarise(trabajadores=survey_total(na.rm = TRUE,
                                      vartype=c("ci")))
```

```
## # A tibble: 4 x 4
##   activ trabajadores trabajadores_low trabajadores_upp
##   <fct>        <dbl>            <dbl>            <dbl>
## 1 1         8148206.         7969723.         8326688.
## 2 2          941088.          888240.          993936.
## 3 3         6763752.         6624705.         6902799.
## 4 <NA>      3742791.         3620543.         3865039.
```

Por defecto survey trabaja con nivel de confianza del 95% (z=1,96). Se puede cambiar:

```
ds %>% group_by(activ) %>%
  summarise(trabajadores=survey_total(na.rm = TRUE,
                                      vartype=c("ci"),
                                      level=c(0.90)))
```

```
## # A tibble: 4 x 4
##   activ trabajadores trabajadores_low trabajadores_upp
```

La forma correcta: `survey` y `srvyr`

Con esto, la tasa de desocupación publicada de 10,4% en EFM 2021 tendrá un nivel de precisión:

```
tasa<-ds %>% group_by(activ) %>%
  summarise(trabajadores=survey_total(na.rm = TRUE,
                          vartype=c("ci"))) %>%
  filter(activ==1|activ==2) %>% # seleccionar solo 2 filas
  janitor::adorn_totals("row")  # total por columna
```

Tasas de desocupación:

```
tasa[2,2:4]/tasa[3,2:4]
```

```
##   trabajadores trabajadores_low trabajadores_upp
## 1    0.1035381        0.1002759        0.1066384
```

La forma correcta: `survey` y `srvyr`

Proporciones por categoría de respuesta

```
ds %>% filter(categoria_ocupacion!=0) %>% group_by(categoria_ocupacion)
  summarise(proportion = survey_mean(vartype = c("ci"),na.rm = TRUE))
```

```
## # A tibble: 7 x 4
##   categoria_ocupacion proportion proportion_low proportion_upp
##   <fct>                    <dbl>          <dbl>          <dbl>
## 1 1                       0.0303         0.0276         0.0330
## 2 2                       0.202          0.195          0.209
## 3 3                       0.598          0.590          0.607
## 4 4                       0.135          0.129          0.141
## 5 5                       0.0211         0.0187         0.0235
## 6 6                       0.00396        0.00278        0.00515
## 7 7                       0.00885        0.00732        0.0104
```

La forma correcta: `survey` y `srvyr`

Media

Similar a como se programan las proporciones, pero incluyendo una variable numérica dentro de `survey_mean`

```
ds %>% filter(categoria_ocupacion!=0) %>% group_by(categoria_ocupacion)
  summarise(media_edad = survey_mean(edad,vartype = c("ci"),na.rm = TRUE
```

```
## # A tibble: 7 x 4
##   categoria_ocupacion media_edad media_edad_low media_edad_upp
##   <fct>                     <dbl>          <dbl>          <dbl>
## 1 1                          50.4           49.2           51.5
## 2 2                          45.8           45.3           46.3
## 3 3                          40.2           40.0           40.5
## 4 4                          42.2           41.7           42.7
## 5 5                          48.4           47.1           49.6
## 6 6                          50.1           45.6           54.6
## 7 7                          40.8           38.4           43.1
```

Mediana (2do cuartil)

```
ds %>% filter(categoria_ocupacion!=0) %>% group_by(categoria_ocupacion)
  summarise(mediana_edad = survey_median(edad,vartype = c("ci"),na.rm =
```

```
## # A tibble: 7 x 4
##   categoria_ocupacion mediana_edad mediana_edad_low mediana_edad_upp
##   <fct>                      <dbl>             <dbl>             <dbl>
## 1 1                             51                49                52
## 2 2                             45                44                46
## 3 3                             39                38                39
## 4 4                             40                40                41
## 5 5                             49                48                51
## 6 6                             54                43                57
## 7 7                             39                36                42
```

La forma correcta: `survey` y `srvyr`

Cuartiles (y otros percentiles)

```
ds %>% filter(!is.na(activ)) %>% group_by(activ) %>%
  summarise(edad=survey_quantile(edad,c(0.25, 0.5, 0.75),na.rm = TRUE))
```

```
## # A tibble: 3 x 7
##   activ edad_q25 edad_q50 edad_q75 edad_q25_se edad_q50_se edad_q75_se
##   <fct>    <dbl>    <dbl>    <dbl>       <dbl>       <dbl>       <dbl>
## 1 1           31       41       52       0.255       0.255       0
## 2 2           26       34       47       0.255       0.510       0.765
## 3 3           23       46       67       0.255       0.510       0.255
```

Desviación estándar y varianza

```
ds %>% filter(categoria_ocupacion!=0) %>% group_by(categoria_ocupacion)
  summarise(sd_edad=survey_sd(edad,na.rm = TRUE),
            varianza_edad=survey_var(edad,vartype = c("ci"),na.rm = TRUE
```

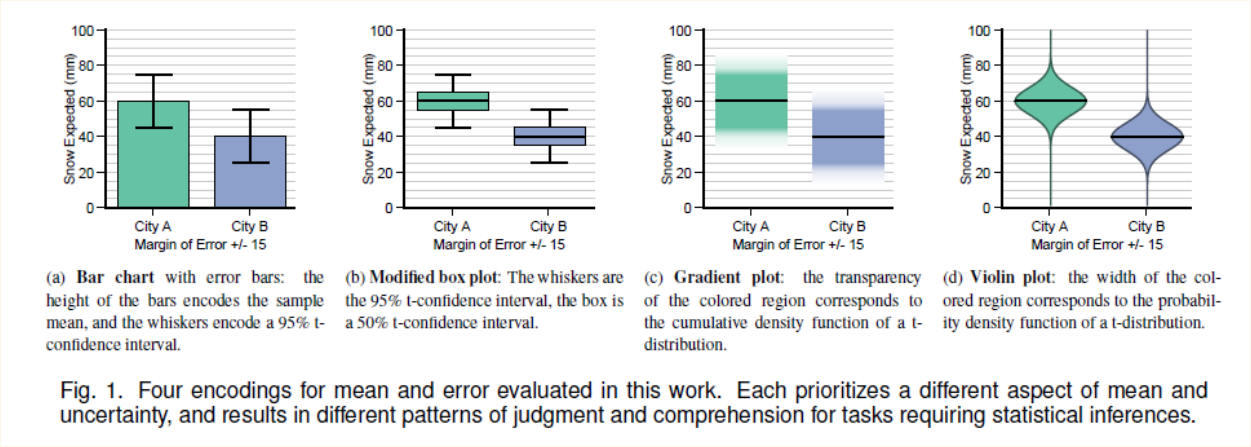
```
## # A tibble: 7 x 5
##   categoria_ocupacion sd_edad varianza_edad varianza_edad_low varianza_edad
##   <fct>                <dbl>         <dbl>             <dbl>             <
## 1 1                     12.8          164.              148.
## 2 2                     14.0          196.              189.
## 3 3                     12.6          159.              155.
## 4 4                     11.9          143.              135.
## 5 5                     11.2          126.              108.
## 6 6                     13.8          191.              130.
## 7 7                     16.4          268.              231.
```

Muestras complejas en R

Visualización de la incertidumbre

Visualizar la incertidumbre

La forma más común es el gráfico de barras o de líneas con barras de error (A).



(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.

(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.

(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.

(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

Fig. 1. Four encodings for mean and error evaluated in this work. Each prioritizes a different aspect of mean and uncertainty, and results in different patterns of judgment and comprehension for tasks requiring statistical inferences.

La visualización puede aportar más de un intervalo (B)

O incluso se puede ir más allá, visualizando la incertidumbre de forma continua (C y D)

La visualización también puede confundir, presentándose errores estándar como si fuesen intervalos.

Revisar Correl, M. Error bars considered harmful para conocer la discusión.

Visualizar la incertidumbre

geom_bar

```
ds %>% filter(categoria_ocupacion!=0) %>% group_by(categoria_ocupacion)
  summarise(trabajadores=survey_total(na.rm = TRUE, vartype=c("ci"),leve
  ggplot(aes(x=categoria_ocupacion,y=trabajadores,fill=categoria_ocupaci
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin=trabajadores_low, ymax=trabajadores_upp),
                width=0.2, position=position_dodge(.9))
```

Visualizar la incertidumbre

Gráfico de líneas

Data con la evolución de los ocupados (en miles) en Chile entre 2010 y 2021.

```
## # A tibble: 133 x 4
##    ocupados periodo ocupados_low ocupados_upp
##       <dbl>   <int>        <dbl>        <dbl>
##  1    7156.       1        6798.        7514.
##  2    7199.       2        6839.        7559.
##  3    7182.       3        6823.        7541.
##  4    7222.       4        6860.        7583.
##  5    7257.       5        6894.        7619.
##  6    7289.       6        6925.        7654.
##  7    7389.       7        7020.        7759.
##  8    7414.       8        7044.        7785.
##  9    7503.       9        7128.        7878.
## 10    7572.      10        7194.        7951.
## # ... with 123 more rows
```

Visualizar la incertidumbre

Gráfico de líneas, solo estimaciones puntuales

```
serie %>%  ggplot(aes(x=periodo, y=ocupados)) +
  geom_line() +
  geom_point() +
  theme_bw()
```

Visualizar la incertidumbre

Gráfico de líneas, con medidas de precisión

```
serie %>%  ggplot(aes(x=periodo, y=ocupados)) +
  geom_line() +
  geom_point() +
  theme_bw() +
  geom_errorbar(aes(ymin=ocupados_low, ymax=ocupados_upp), width=.01)
```

Gráficos de cajas, estimando a la población

```
svyboxplot(edad~categoria_ocupacion, design=ds,all.outliers=TRUE)
```

Gráficos de cajas, estimando a la población

Gráficos de cajas modificados

o The Modern Box Plot

Para visulizar precisión (IC) en vez cuartiles

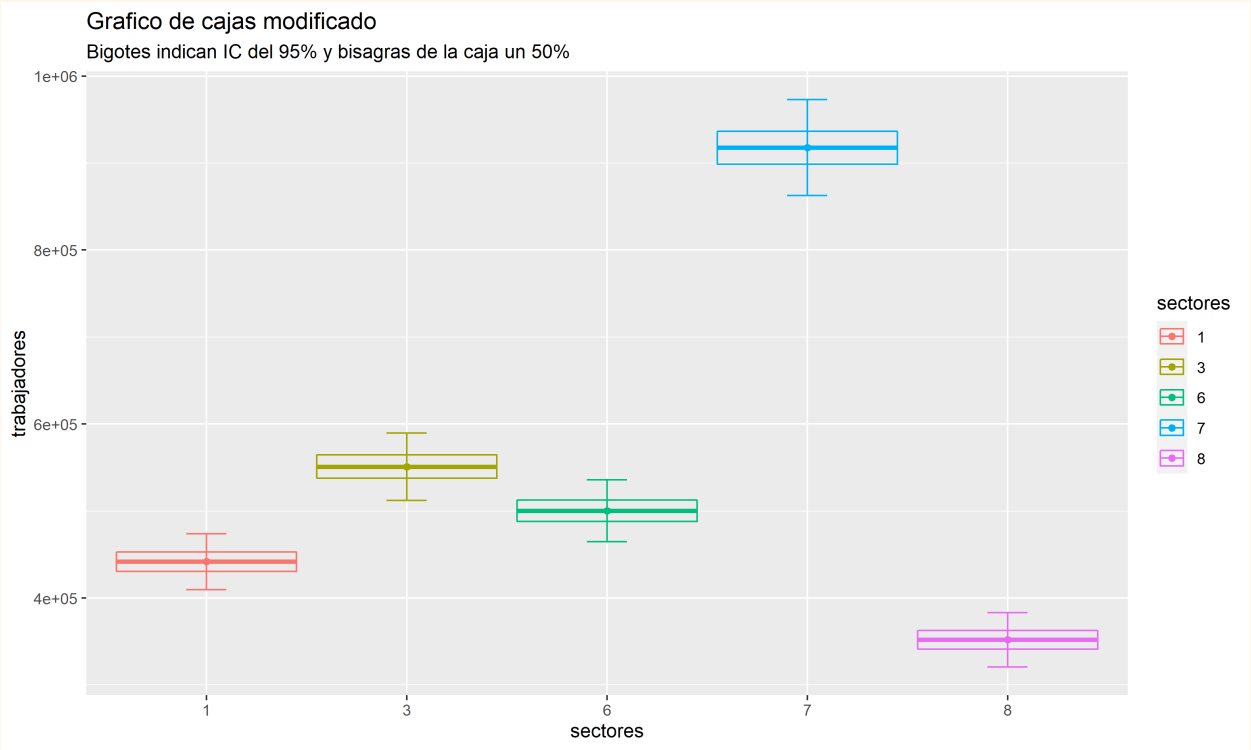
- Box center: sample mean (o total)
- Box edges: standard error of the mean (IC 50% en este caso)
- Box whiskers: 95% confidence interval

```
## Crear tabla de ocupados por sector económico
base<-ds %>% filter(categoria_ocupacion==3&r_p_rev4cl_caenes%in%c(1,3,6,
  group_by(r_p_rev4cl_caenes) %>%
  summarise(trabajadores=survey_total(na.rm = TRUE, vartype=c("ci"),leve
  rename(sectores=r_p_rev4cl_caenes) %>% mutate(sectores=as.factor(secto

base
```

Gráficos de cajas modificados

```
base %>% ggplot(aes(y = trabajadores, x = sectores,color=sectores)) +
  geom_point() +
  geom_crossbar(aes(ymin = trabajadores_low50, ymax = trabajadores_upp5(
  geom_errorbar(aes(ymin = trabajadores_low95, ymax =  trabajadores_upp9
```



Correlación con pesos

```
## Correlación considerando pesos
library(weights)
ene2<-ene %>% select(edad,c2_1_1,c2_2_1)
weighted_corr<-wtd.cor(ene2, weight = ene$fact_cal)
weighted_corr$correlation
```

```
##               edad      c2_1_1    c2_2_1
## edad    1.00000000 -0.03761003 0.1405247
## c2_1_1 -0.03761003  1.00000000 0.2348693
## c2_2_1  0.14052466  0.23486931 1.0000000
```

```
# Correlación considerando sin considerar pesos
cor(ene2, use = "complete.obs")
```

```
##               edad      c2_1_1     c2_2_1
## edad    1.00000000 -0.05798551 0.04832614
## c2_1_1 -0.05798551  1.00000000 0.24264390
## c2_2_1  0.04832614  0.24264390 1.00000000
```

Bibliografía y elementos consultados

Heiss, A. Uncertainty. En curso "Data Visualization".

INE. Boletín Mensual DEF 2021 Encuesta Nacional de Empleo.

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Lehmann et al (2021) Presentación paquete "calidad" en LatinR

Lohr, S. L. (2000). *Muestreo: Diseño y Análisis*. 519.52 L6. International Thomson Editores.

Vivanco, M. (2006). "Diseño de Muestras En Investigación Social". In: *Metodolog\'ias de La Investigación Social. Introducción a Los Oficios.* Santiago: LOM, pp. 141-168.