# Trabajo final Muestreo II

Fiorella Lúngaro, Emanuelle Marsella y Maximiliano Saldaña

Diciembre 2021

# Parte 1

```
# Carga la muestra
muestra <- read_xlsx("datos/muestra grupo 2.xlsx")

# Convertir las variables categóricas a su formato correspondiente

muestra <- muestra %>%
  mutate(across(where(is.double) & !c(ingreso, w0, edad, R), as.factor))
```

Se calculan las estimaciones puntuales de la tasa de desempleo, la proporción de personas pobres y del ingreso promedio, haciendo uso de los ponderadores originales $w_0$, es decir, sin ajustar por no respuesta. Esta estrategia de cómputo resulta correcta si el esquema de no respuesta que se considera es *Missing Completely at Random* (MCAR), bajo el cual la probabilidad de responder no depende de las variables de interés ni auxiliares y todas las unidad del marco tienen la misma probabilidad de responder (Ferreira y Zoppolo, 2017).

```
# Diseño usando los ponderadores originales, MCAR.
##FPC?
design1 <- muestra %>%
  filter(R==1) %>%
  as_survey_design(ids = id_hogar, strata = estrato, weights = w0)
```

```
## Tasa de desempleo (desempleados/activos)

#REVISAR ESTO
#Se piensa como un problema de estimación en dominios, nos interesan los desempleados considerando el g
design1 %>%
  filter(activo == 1) %>%
  group_by(desocupado) %>%
  summarise(tasa_desempleo = survey_mean(deff = TRUE, vartype = c('se','cv')))
```

```
## # A tibble: 2 x 5
##   desocupado tasa_desempleo tasa_desempleo_se tasa_desempleo_cv tasa_desempleo_~
##   <fct>               <dbl>             <dbl>             <dbl>            <dbl>
## 1 0                  0.918           0.00330           0.00360             1.07
## 2 1                  0.0824          0.00330           0.0400              1.07
```

La estimación puntual de la proporción de desempleados es 0,0824; mientras que el error estándar (la medida que empleamos para medir la variación del estimador entre muestra y muestra) es 0,0033. Otra medida de la calidad de un estimador $\hat{\theta}$ es su coeficiente de variación, que mide su dispersi'on relativa. Se define como (Zoppollo, x):

$$CV(\hat{\theta}) = \frac{\sqrt{\hat{V}(\hat{\theta})}}{\left|E(\hat{\theta})\right|}$$

Y en el caso del estimador de la proporci'on de desempleados su estimaci'on es 0,04. El efecto dise~no es una medida que permite comparar la eficiencia en t'erminos de variabilidad del estimador para el dise~no utilizado, respecto al dise~no aleatorio simple sin reposici'on que. Siendo $p(s)$ el dise~no medible considerado, se define como:

$$Deff(p(s), \hat{\theta}) = \frac{V_{p(s)}(\hat{\theta})}{V_{SI}(\hat{\theta})}$$

En el caso del estimador de la proporci'on de desempleados su valor es 1,07; lo que indica que en este caso el diseño SI es un 7 % m'as eficiente que el empleado.

```
## Proporción de personas pobres
design1 %>%
  group_by(pobreza) %>%
  summarise(prop_pobres = survey_mean(deff = TRUE, vartype = c('se','cv')))
```

```
## # A tibble: 2 x 5
##   pobreza prop_pobres prop_pobres_se prop_pobres_cv prop_pobres_deff
##   <fct>         <dbl>          <dbl>          <dbl>            <dbl>
## 1 0            0.919         0.00377        0.00410             2.84
## 2 1            0.0811        0.00377        0.0465              2.84
```

En cuanto a la proporci'on de personas pobres, la estimaci'on puntual es de 0,0811. El error est'andar se estima que es 0,004 aproximadamente, mientras que el coeficiente de variaci'on se estima que es 0,05 aproximadamente. La estimaci'on del efecto dise~no es 2,84; un elevado valor que indica que el dise'no empleado es altamente ineficiente en comparaci'on con el SI, en particular casi tres veces m'as.

```
## Ingreso promedio
design1 %>%
  summarise(ingreso_prom = survey_mean(ingreso, deff = TRUE, vartype = c('se','cv')))
```

```
## # A tibble: 1 x 4
##   ingreso_prom ingreso_prom_se ingreso_prom_cv ingreso_prom_deff
##          <dbl>           <dbl>           <dbl>             <dbl>
## 1       21799.            240.          0.0110             0.935
```

La estimaci'on puntual del ingreso promedio es 21799, siendo la estimaci'on de su error est'andar 240. Por otro lado, el coeficiente de variaci'on toma el valor 0,011. La estimaci'on del efecto dise~no es 0,94 aproximadamente, por lo que en este caso el dise~no empelado resulta m'as eficiente que el SI, un 6 % m'as.

```
muestra %>%
  summarise(
    # tasa de no respuesta no ponderada
    nr_np = 1 - mean(R),
    # tasa de no respuesta ponderada
    nr_p = 1 - weighted.mean(R, w0)
  )
```

```
## # A tibble: 1 x 2
##   nr_np  nr_p
##   <dbl> <dbl>
## 1 0.474 0.476
```

```
summary(muestra$w0)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.4   110.8   124.8   125.6   140.3   162.0
```

```r
#considerando por estratos

muestra %>%
  group_by(estrato) %>%
  summarise(
    # tasa de no respuesta no ponderada
    nr_np = mean(R),
    # tasa de no respuesta ponderada
    nr_p = weighted.mean(R, w0)
    )
```

```
## # A tibble: 12 x 3
##    estrato nr_np  nr_p
##    <fct>   <dbl> <dbl>
##  1 1       0.455 0.455
##  2 2       0.530 0.530
##  3 3       0.533 0.533
##  4 4       0.547 0.547
##  5 5       0.538 0.538
##  6 6       0.487 0.487
##  7 7       0.543 0.543
##  8 8       0.543 0.543
##  9 9       0.526 0.526
## 10 10      0.508 0.508
## 11 11      0.552 0.552
## 12 12      0.564 0.564
```

```r
#considerando por departamento

muestra %>%
  group_by(dpto) %>%
  summarise(
   # tasa de no respuesta no ponderada
    nr_np = mean(R),
    # tasa de no respuesta ponderada
    nr_p = weighted.mean(R, w0)
    )
```

```
## # A tibble: 19 x 3
##    dpto  nr_np  nr_p
##    <fct> <dbl> <dbl>
##  1 1     0.523 0.521
##  2 2     0.545 0.545
##  3 3     0.515 0.512
##  4 4     0.542 0.542
##  5 5     0.511 0.511
##  6 6     0.570 0.570
##  7 7     0.552 0.552
##  8 8     0.576 0.576
##  9 9     0.556 0.556
## 10 10    0.534 0.534
## 11 11    0.517 0.517
## 12 12    0.498 0.498
```

```
## 13 13    0.535 0.535
## 14 14    0.533 0.533
## 15 15    0.545 0.545
## 16 16    0.498 0.494
## 17 17    0.499 0.499
## 18 18    0.540 0.540
## 19 19    0.556 0.556
```

La tasa de no respuesta no ponderada es del 47,4 % mientras que la ponderada es del 47,6 %. El hecho de que ambas tasas de no respuesta sean similares puede deberse a que los pesos $w_0$ no son muy disímiles entre sí, siendo su mínimo 104,4; su media 125.6 y su máximo 162. Al considerar la proporción de no respondentes por estrato se puede apreciar que ocurre lo mismo. En este caso se puede observar que la tasa de no respuesta varía según el estrato considerado, siendo el primero (Montevideo bajo) el que cuenta con la mayor tasa, del 56 %, y el doceavo el que cuenta con la menor, del 46 %. Estas diferencias se ven reflejadas también al considerar la tasa de no respondentes por departamento.