Análisis de encuestas de hogares con R

Módulo 1: Análisis de variables continuas

CEPAL - División de Estadísticas Sociales

Motivación

Motivación

Los desarrollos estadísticos están en permanente evolución, surgiendo nuevas metodologías y desarrollando nuevos enfoquen el análisis de encuestas. Estos desarrollos parten de la academia, luego son adoptados por las empresas (privadas o estatales) y entidades estatales. Las cuales crean la necesidad que estos desarrollos sean incluidos en software estadísticos licenciados. Proceso que puede llevar mucho tiempo.

Motivación

Algunos investigadores para acortar los tiempos y poner al servicio de la comunidad sus descubrimientos y desarrollos, hacen la implementasión de sus metodología en paquetes estadísticos de código abierto como **R** o **Python**. Teniendo **R** un mayor número de desarrollos en el procesamiento de las encuestas.

Motivación

Dentro del software *R* se disponen de múltiples librería para el prcesamiento de encuestas, estas varian dependiendo el enfoque de programación desarrollado por el autor o la necesidad que se busque suplir. En esta presentación nos centraremos en las libreria `survey` y `srvyr`. Se incluiran más librerías de acuerdo a las necesidad se presente.

Lectura y procesamientos de encuestas con R

Lectura de la base

La base de datos (tablas de datos) puede estar disponible en una variedad de formatos (`.xlsx`, `.dat`, `.cvs`, `.sav`, `.txt`, ...), sin embargo, por experiencia es recomendable realizar la lectura de cualesquiera de estos formatos y proceder inmediatamente a guardarlo en un archivo de extensión **.rds**, la cual es nativa de `R`. El hacer esta acción reduce considerablemente los tiempo de cargue de la base de datos.

Sintaxis


```
library(haven)
encuestaDOM2020 <- read_dta("../Data/DOM_2020N1.dta")
saveRDS(object = encuestaDOM2020,
        file =  "../Data/encuesta.rds")
```

```
encuesta <- readRDS("../Data/encuesta.rds") %>%
  transmute(estrato  = as.character(`_estrato`),
            upm   = as.character(`_upm`),
            fep   = `_fep`,
            Zone = as_factor(areageo2),
            Sex = as_factor(sexo),
            Age = edad,
            lp, li,  # Linea de pobreza e indigencia
            Income = ingcorte,
            anoest, # años de estudio
            empleo = condact3)
```

Definir diseño de la muestra con `srvyr`

La librería `srvyr` surge como un complemento para `survey`. Estas librerías permiten definir objetos tipo "**survey.design**" a los que se aplican los métodos "**survey.design**" complementados con la programación de tubería (%>%) del paquete `tidyverse`.

Cómo definir un objeto *survey.design*

Para el desarrollo de la presentación se define el diseño muestral con la función `as_survey_design`.

```
# En caso de tener estratos con una muestra.
# Calcula la varianza centrada en la media de la pob.
options(survey.lonely.psu = "adjust")
library(srvyr)

diseno <- encuesta %>% # Base de datos.
  as_survey_design(
    strata = estrato,  # Id de los estratos.
    ids = upm,         # Id para las observaciones.
    weights = fep,      # Factores de expansión.
    nest = T           # Valida el anidado dentro
                       #  del estrato
  )
```

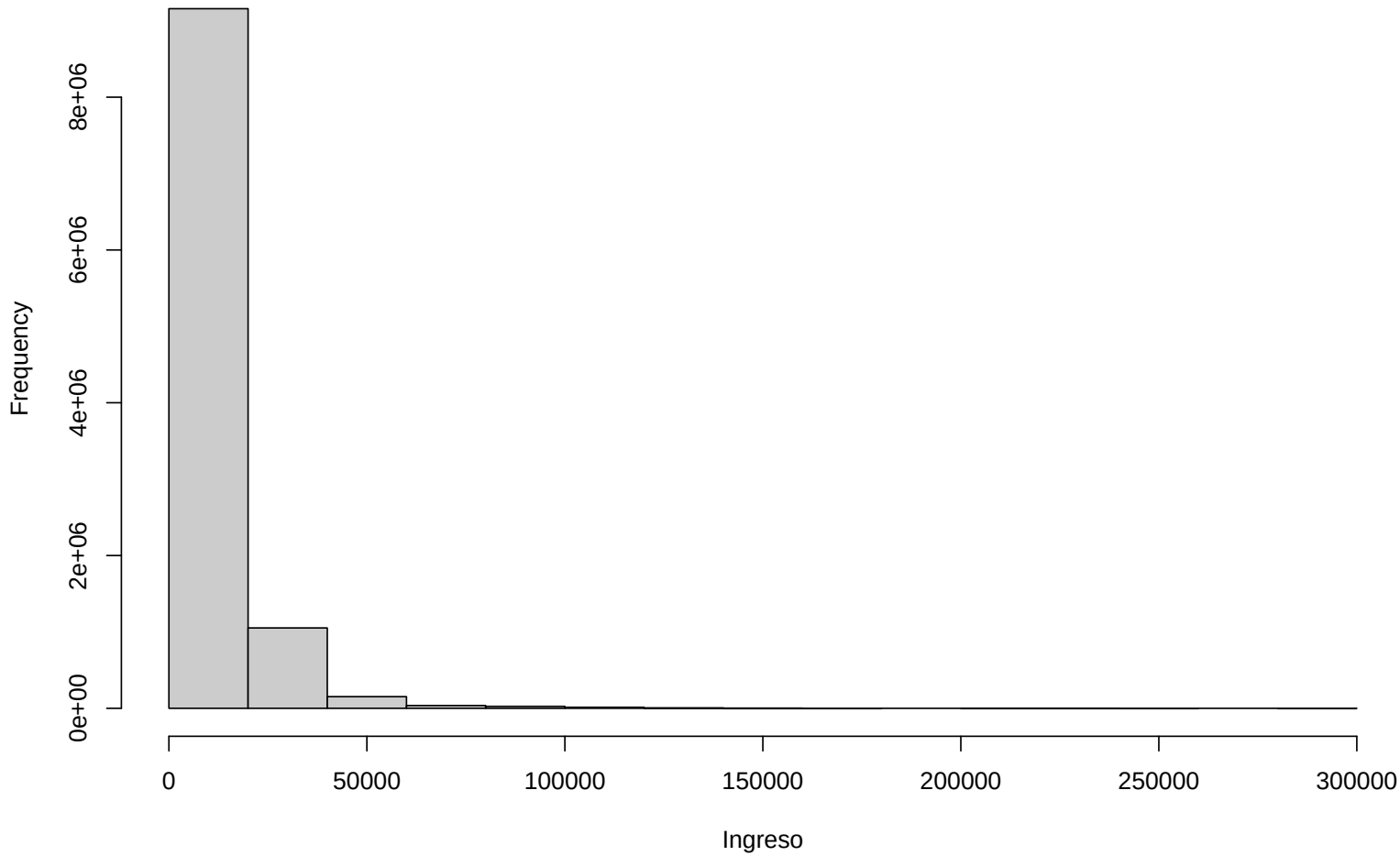
Análisis gráfico

Histograma ponderado para la variable ingreso

A continuación observan la sintaxis para crear una histograma de la variable ingreso haciendo uso la función `svyhist` de la librería `survey`

```
svyhist(
  ~ Income ,
  diseno,
  main = "",
  col = "grey80",
  xlab = "Ingreso",
  probability = FALSE
)
```

Histograma ponderado para la variable ingreso



Dividiendo la muestra en Sub-grupos

En ocasiones se desea realizar estimaciones por sub-grupos de la población, en este caso se extraer 4 sub-grupos de la encuesta.

```
sub_Urbano <- diseno %>%  filter(Zone == "Urbana")
sub_Rural  <- diseno %>%  filter(Zone == "Rural")
sub_Mujer  <- diseno %>%  filter(Sex == "Mujer")
sub_Hombre <- diseno %>%  filter(Sex == "Hombre")
```

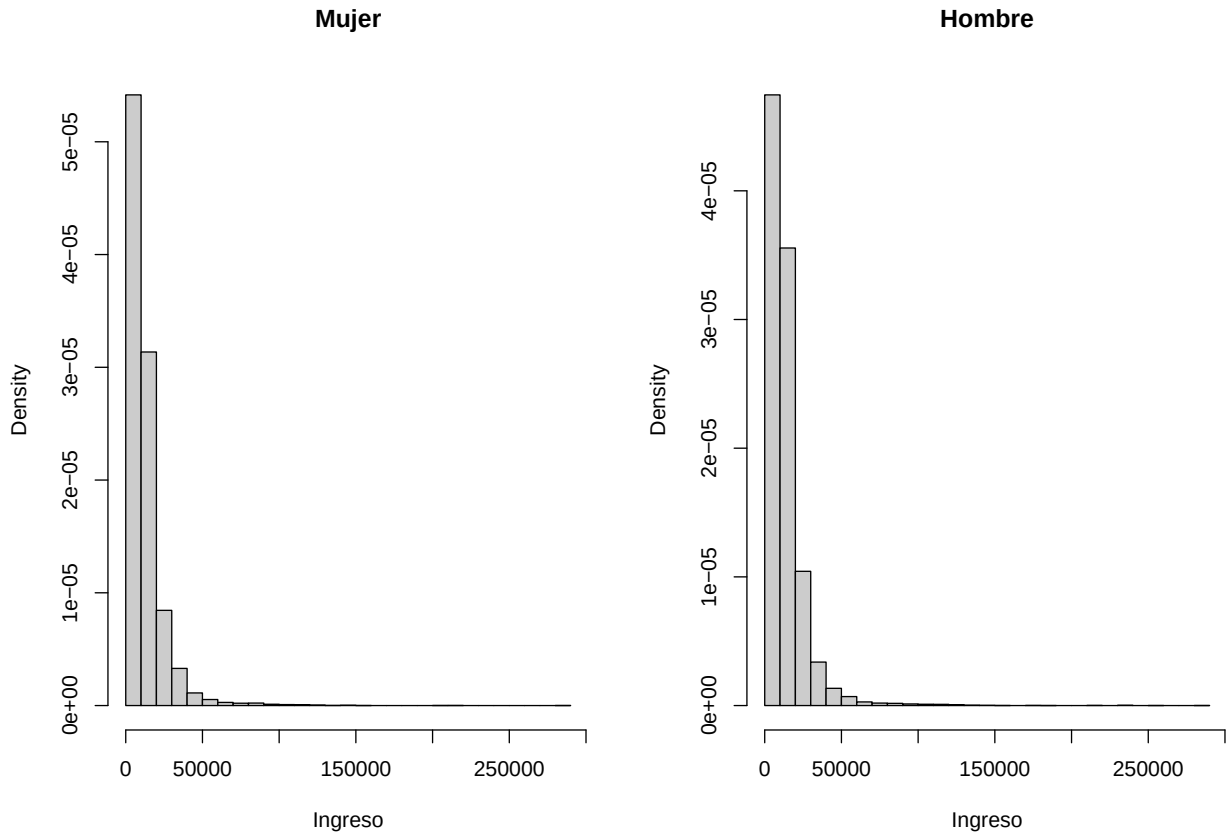
Histograma ponderado en sub-grupos

La sintaxis incluye un filtro de las personas mayores a 18 años

```
par(mfrow = c(1,2))
svyhist(
  ~ Income ,
  design = subset(sub_Mujer, Age >= 18),
  main = "Mujer",
  breaks = 30,
  col = "grey80",
  xlab = "Ingreso"
)

svyhist(
  ~ Income ,
  design = subset(sub_Hombre, Age >= 18),
  main = "Hombre",
  breaks = 30,
  col = "grey80",
  xlab = "Ingreso"
)
```

Histograma ponderado en sub-grupos



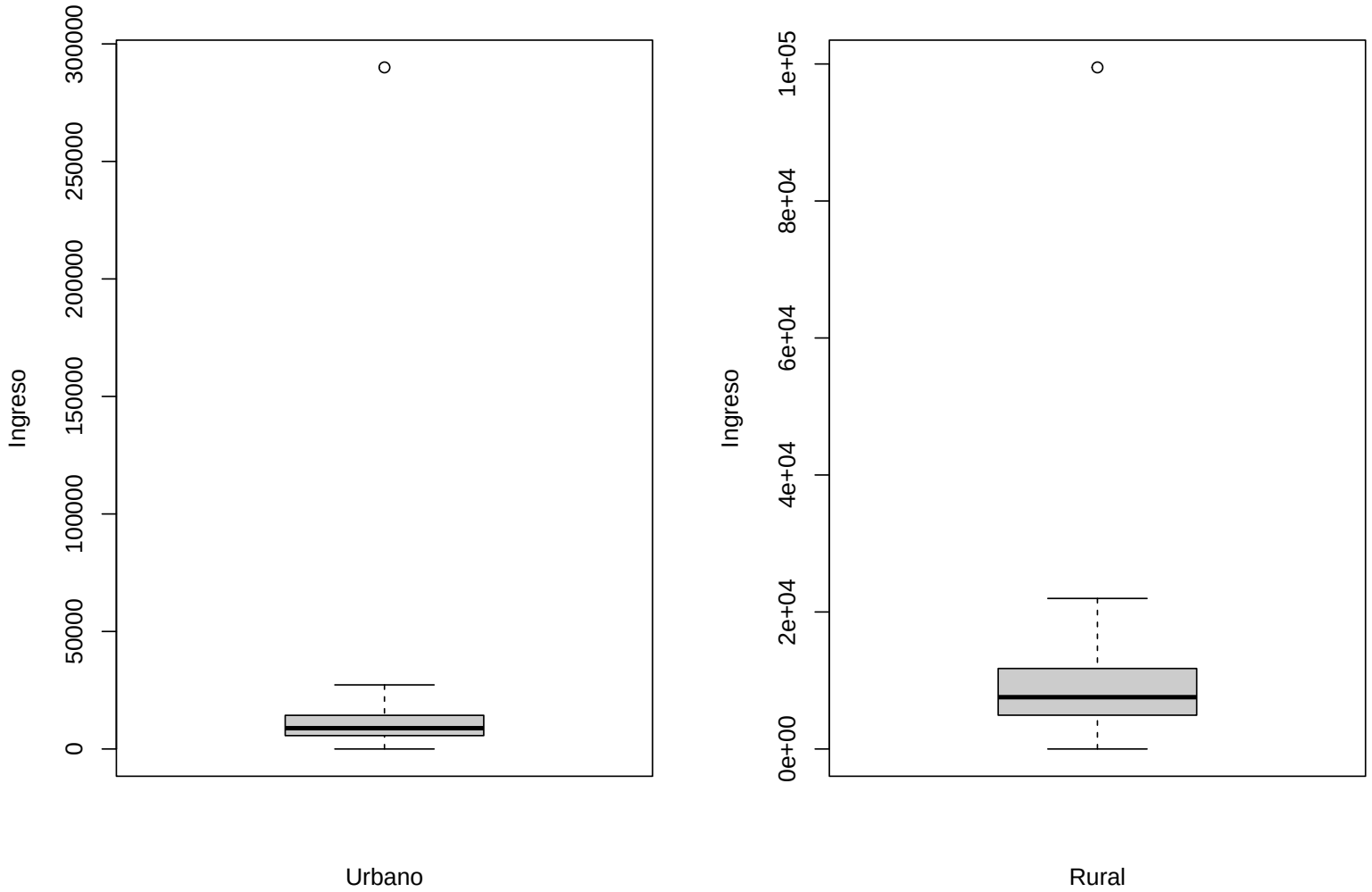
Observe que hay una mayor proporción de hombres en el rango de los 1000 a 3000 que mujeres.

Boxplot ponderado del ingreso por sub-grupos


```
par(mfrow = c(1,2))
svyboxplot(
  Income ~1 ,
  sub_Urbano,
  col = "grey80",
  ylab = "Ingreso",
  xlab = "Urbano")

svyboxplot(
  Income ~ 1 ,
  sub_Rural,
  col = "grey80",
  ylab = "Ingreso",
  xlab = "Rural"
)
```

Boxplot ponderado del ingreso por sub-grupos



Estimaciones puntuales.

Estimaciones puntuales.

Después de realizar el análisis gráfico de las tendencias de las variables continuas, es necesarios obtener las estimaciones puntuales de la variables. Los cuales son obtenidos de forma general o desagregado por niveles, de acuerdo con las necesidades de la investigación.

Estimación de totales e intervalos de confianza del ingreso

La estimación del total se mediante la función `svytotal` y el intervalos de confianza con la función `confint` de la librería `survey`.

```
svytotal(~Income, diseno, deff=T) %>%
  data.frame()
```

	total	Income	deff
Income	1.218e+11	2.225e+09	24.96

```
confint(svytotal (~Income, diseno, deff=T))
```

	2.5 %	97.5 %
Income	1.174e+11	1.262e+11

Estimación de totales por sub-grupos

En esta oportunidad se hace uso de la función cascadede la libraría srvyr, la cual permite agregar la suma de las categorías al final tabla. La función group_by permite obtener resultados agrupados por los niveles de interés.

```
diseno %>% group_by(Sex) %>%
  cascade(Total = survey_total(
    Income, level = 0.95,
    vartype =  c("se", "ci")),
          .fill = "Nacional")
```

Sex	Total	Total_se	Total_low	Total_upp
Hombre	6.094e+10	1.177e+09	5.863e+10	6.325e+10
Mujer	6.086e+10	1.291e+09	5.832e+10	6.339e+10
Nacional	1.218e+11	2.225e+09	1.174e+11	1.262e+11

Estimación de la media e intervalo de confianza del ingreso

Un resultado más interesante para las variables ingreso y gasto es el promedio de la variable.

```
svymean(~Income, diseno, deff=T) %>%
  data.frame()
```

	mean	Income	deff
Income	11658	216.5	25.8

```
confint(svymean (~Income, diseno, deff=T))
```

	2.5 %	97.5 %
Income	11234	12082

Estimación de la media por sub-grupos

La función cascade regresa el resultado promedio ignorando los niveles.

```
diseno %>% group_by(Zone, Sex) %>%
  cascade(
    Media = survey_mean(
      Income, level = 0.95,
       vartype =  c("se", "ci")),
        .fill = "Nacional") %>%
  data.frame()
```

Zone	Sex	Media	Media_se	Media_low	Media_upp
Urbana	Hombre	12458	267.0	11935	12982
Urbana	Mujer	11826	277.8	11281	12371
Urbana	Nacional	12131	260.4	11620	12642
Rural	Hombre	9929	225.3	9487	10371
Rural	Mujer	8987	199.6	8595	9378
Rural	Nacional	9465	192.6	9087	9843
Nacional	Nacional	11658	216.5	11233	12083

Estimación de la varianza de los ingresos por sub-grupo

La estimación de la varianza se obtiene con `survey_var`

```
(tab_var <- diseno %>% group_by(Zone) %>%
   summarise(Var =
  survey_var(
    Income,
    level = 0.95,
    vartype =  c("se", "ci"),
   )))
```

Zone	Var	Var_se	Var_low	Var_upp
Urbana	146690165	17974185	111419843	181960486
Rural	49769232	3677926	42552125	56986340

Estimación de la desviación estándar de los ingresos por sub-grupo


```
(tab_sd <- diseno %>% group_by(Zone) %>%
   summarise(Sd =
  survey_sd(
    Income,
    level = 0.95,
    vartype =  c("se", "ci"),
   )
)) %>% data.frame()
```

Zone	Sd
Urbana	12112
Rural	7055

Estimación de la mediana para el ingreso

La estimación de la median se obtiene con `survey_median`

```
diseno %>% summarise(Mediana =
  survey_median(
    Income,
    level = 0.95,
    vartype =  c("se", "ci"),
   ))
```

Mediana	Mediana_se	Mediana_low	Mediana_upp
8600	115.5	8380	8833

Estimación de la mediana por sub-grupo


```
diseno %>% group_by(Zone) %>%
  summarise(Mediana =
  survey_median(
    Income,
    level = 0.95,
    vartype =  c("se", "ci"),
   ))
```

Zone	Mediana	Mediana_se	Mediana_low	Mediana_upp
Urbana	8851	143.4	8600	9163
Rural	7560	171.7	7217	7894

Estimación de la mediana por sub-grupo


```
diseno %>% group_by(Sex) %>%
  summarise(Mediana =
  survey_median(
    Income,
    level = 0.95,
    vartype =  c("se", "ci"),
   ))
```

Sex	Mediana	Mediana_se	Mediana_low	Mediana_upp
Hombre	9000	136.1	8732	9267
Mujer	8303	117.3	8073	8533

Estimación del quantil 0.5 para el ingreso

La estimación de la median se obtiene con `survey_quantile`

```
diseno %>%
  summarise(
    Q =  survey_quantile(
    Income,
    quantiles = 0.5,
    level = 0.95,
    vartype =  c("se", "ci"),
    interval_type = "score"
   ))
```

Q_q50	Q_q50_se	Q_q50_low	Q_q50_upp
8600	133.6	8806	9330

Estimación del quantil 0.25 para el ingreso por sub-grupo


```
diseno %>% group_by(Sex) %>%
  summarise(
    Q =  survey_quantile(
    Income,
    quantiles = 0.25,
    level = 0.95,
    vartype =  c("se", "ci"),
    interval_type = "score"
   ))
```

Sex	Q_q25	Q_q25_se	Q_q25_low	Q_q25_upp
Hombre	5708	84.43	5789	6120
Mujer	5397	76.05	5500	5798

Estimación de los quantiles 0.25 y 0.75 para el ingreso por sub-grupo


```
diseno %>% group_by(Zone) %>%
  summarise(
    Q =  survey_quantile(
    Income,
    quantiles = c(0.25, 0.75),
    level = 0.95,
    vartype =  c("se"),
    interval_type = "score"
   ))
```

Zone	Q_q25	Q_q75	Q_q25_se	Q_q75_se
Urbana	5692	14325	87.59	300.0
Rural	4932	11750	118.76	336.6

Estimación de la razón entre hombres y mujeres

La estimación de una razón se obtiene con la función `survey_ratio`.

```
diseno %>% summarise(
    Razon =  survey_ratio(
      numerator = (Sex == "Mujer"),# creando dummy.
      denominator = (Sex == "Hombre"),# creando dummy
      level = 0.95,
    vartype =  c("se", "ci")
    ))
```

Razon	Razon_se	Razon_low	Razon_upp
1.056	0.0134	1.029	1.082

Estimación de la razón entre hombres y mujeres en la zona rural


```
sub_Rural %>% summarise(
    Razon =  survey_ratio(
      numerator = (Sex == "Mujer"),
      denominator = (Sex == "Hombre"),
      level = 0.95,
    vartype =  c("se", "ci")
    ))
```

Razon	Razon_se	Razon_low	Razon_upp
0.9708	0.0265	0.9186	1.023

Pruebas de diferencia medias

Pruebas de diferencia medias

Los analistas de los datos de las encuestas suelen estar interesados en hacer inferencias sobre las diferencias de las estadísticas descriptivas de dos subpoblaciones. A continuación se muestra como realizar estas comparaciones haciendo uso de la función `svyttest`

Ingreso promedio por sexo

Sex	Media	Media_se	Media_low	Media_upp
Hombre	11990	221.0	11557	12424
Mujer	11343	232.7	10886	11800

Pruebas de diferencia medias de los ingresos entre hombres y mujeres

La comparación de los ingresos medios entre hombre y mujeres de la muestra se realiza así:

```
svyttest(Income ~ Sex, diseno)
```

```
##
##  Design-based t-test
##
## data:  Income ~ Sex
## t = -4.7, df = 1026, p-value = 3e-06
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -916.6 -378.0
## sample estimates:
## difference in mean
##             -647.3
```

Pruebas de diferencia medias de los ingresos entre hombres y mujeres en la zona urbana

También es posible realizar el procedimiento en sub-grupos de interés.

Sex	Media	Media_se
Hombre	12458	267.0
Mujer	11826	277.8

```
svyttest(Income ~ Sex, sub_Urbano)
```

```
##
##  Design-based t-test
##
## data:  Income ~ Sex
## t = -3.9, df = 767, p-value = 1e-04
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -948.4 -315.7
## sample estimates:
## difference in mean
```

Pruebas de diferencia medias de los ingresos entre hombres y mujeres mayores a 18 años


```
svyttest(Income ~ Sex, diseno %>%
           filter(Age > 18, empleo == 1))
```

```
##
##  Design-based t-test
##
## data:  Income ~ Sex
## t = 1.6, df = 1024, p-value = 0.1
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  -111.2 1133.1
## sample estimates:
## difference in mean
##              510.9
```

¡Gracias!

Email: andres.gutierrez@cepal.org