Análisis de encuestas de hogares con R

Módulo 1: Análisis de variables continuas

CEPAL - Unidad de Estadísticas Sociales

Tabla de contenidos I

Introducción

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

```
encuesta <- readRDS("../Data/encuesta.rds")
```

Definir diseño de la muestra con `srvyr`

La libreria `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 = Stratum,  # Id de los estratos.
    ids = PSU,         # Id para las observaciones.
    weights = wk,      # Factores de expansión.
    nest = T           # Valida el anidado dentro del estrato
  )
```

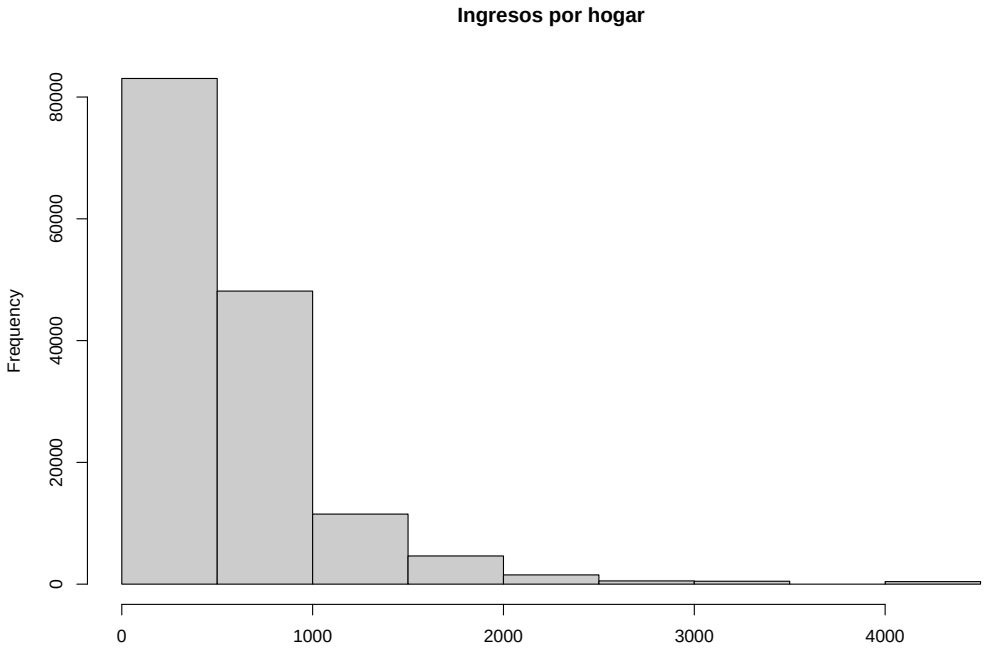
Análisis gráfico

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 = "Ingresos por hogar",
  col = "grey80",
  xlab = "Ingreso",
  probability = FALSE
)
```

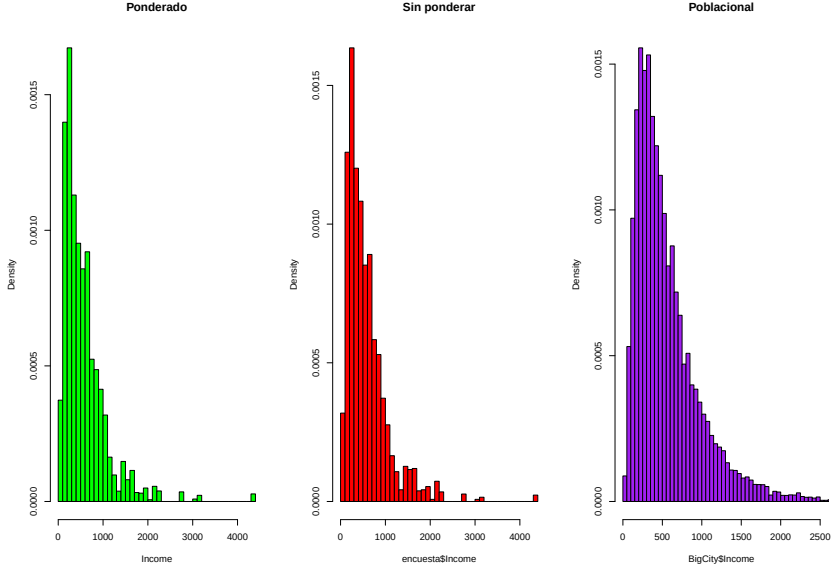
Histograma ponderado para la variable ingreso



Comparación de histogramas

```
data("BigCity", package = "TeachingSampling")
par(mfrow = c(1,3))
svyhist(   ~ Income,
  diseno, main = "Ponderado",
  col = "green", breaks = 50
)
hist( encuesta$Income,
  main = "Sin ponderar",
  col = "red", prob = TRUE, breaks = 50
)
hist(  BigCity$Income,
  main = "Poblacional",
  col = "purple", prob = TRUE,
  xlim = c(0, 2500), breaks = 500
)
```

Comparación de histogramas



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 == "Urban")
sub_Rural  <- diseno %>%  filter(Zone == "Rural")
sub_Mujer  <- diseno %>%  filter(Sex == "Female")
sub_Hombre <- diseno %>%  filter(Sex == "Male")
```

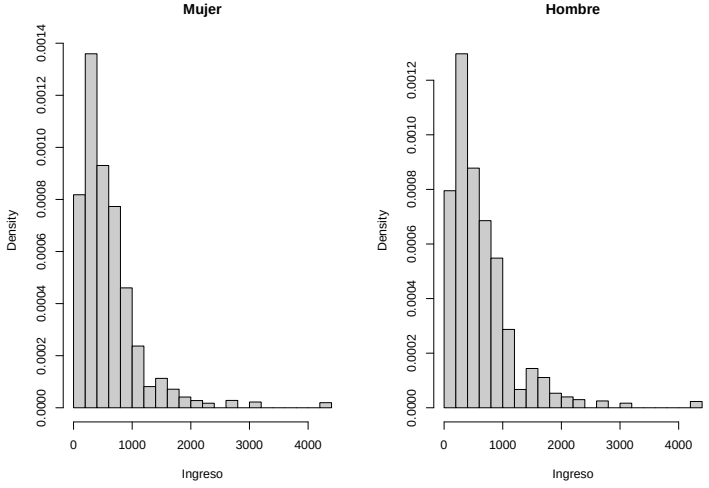
Histograma ponderado en sub-grupos

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



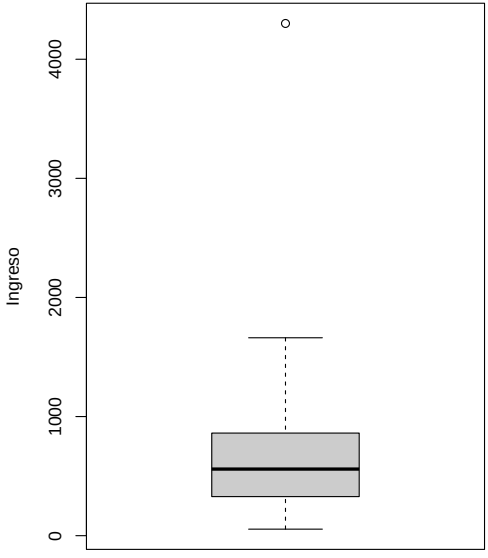
Observe que hay una mayor proporción de hombres en el rango de los 1000 a 3000 que mujeres.

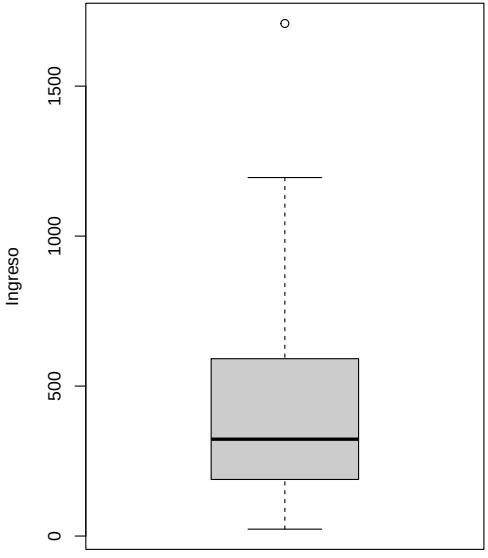
Boxplot ponderado del ingreso por sub-grupos

```
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.

Introducción

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 puntual

- El proceso implica utilizar técnicas avanzadas como los estimadores generales de regresión (GREG) y métodos de calibración.
- **Valiente et al. (2000)** desarrolló una librería en *S-plus* que permite llevar a cabo estos procedimientos en R **(Valliant et al., 2013)**.
- Para estimar el total en un diseño con estratificación y muestreo por conglomerados, se utiliza la fórmula:

$$\hat{Y}_{\omega} = \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}$$

.

Estimación de la varianza

La varianza estimada para este estimador es compleja y se calcula de la siguiente manera:

$$var\left(\widehat{Y}_{\omega}\right) = \sum_{h=1}^{H} \frac{a_h}{(a_h - 1)} \left[\sum_{\alpha=1}^{a_h} \left(\sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i} \right)^2 - \frac{\left(\sum_{\alpha=1}^{a_h} \omega_{h\alpha i} y_{h\alpha i} \right)^2}{a_h} \right]$$

.

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	85793667	4778674	11

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

	2.5 %	97.5 %
Income	76427637	95159697

Estimación de totales e intervalos de confianza del gasto

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

	total	Expenditure	deff
Expenditure	55677504	2604138	10.22

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

	2.5 %	97.5 %
Expenditure	50573486	60781522

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 = "Total ingreso")
```

Sex	Total	Total_se	Total_low	Total_upp
Female	44153820	2324452	39551172	48756467
Male	41639847	2870194	35956576	47323118
Total ingreso	85793667	4778674	76331414	95255920

Estimación de la media e intervalo de confianza

- La estimación de la media poblacional es un parámetro crucial en encuestas de hogares, especialmente en el caso de indicadores como los ingresos medios por hogar.
- Según **Gutiérrez (2016)**, se puede expresar un estimador de la media poblacional como una razón no lineal de dos totales poblacionales finitos estimados:

$$\bar{Y}_{\omega} = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega_{h\alpha i}} = \frac{\hat{Y}}{\hat{N}}$$

.

Estimación de la varianza

- Calcular la varianza de este estimador es complejo, ya que no existe una fórmula cerrada para ello debido a su naturaleza no lineal.
- Una fórmula insesgada para la varianza es:

$$var\left(\bar{Y}_{\omega}\right) \approx \frac{var\left(\hat{Y}\right) + \bar{Y}_{\omega}^{2} \times var\left(\hat{N}\right) - 2 \times \bar{Y}_{\omega} \times cov\left(\hat{Y}, \hat{N}\right)}{\hat{N}^{2}}$$

.

- Estos cálculos pueden realizarse en R utilizando funciones incorporadas, ya que implican componentes complejos como la covarianza entre el total estimado y el tamaño poblacional estimado.

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	570.9	28.48	8.821

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

	2.5 %	97.5 %
Income	515.1	626.8

Estimación de la media e intervalo de confianza del gasto

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

	mean	Expenditure	deff
Expenditure	370.5	13.29	6.016

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

	2.5 %	97.5 %
Expenditure	344.5	396.6

Estimación de la media por sub-grupos

La función cascade regresa el resultado promedio ignorando los niveles.

```
diseno %>% group_by(Sex) %>%
  cascade(
    Media = survey_mean(
      Expenditure, level = 0.95,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio"  ) %>%
  arrange(desc(Sex)) # Ordena la variable.
```

Sex	Media	Media_se	Media_low	Media_upp
Male	374.4	16.06	342.6	406.2
Female	367.0	12.34	342.6	391.5
El gasto medio	370.5	13.29	344.2	396.9

Estimación de la media por sub-grupos

```
diseno %>% group_by(Zone) %>%
  cascade(
    Media = survey_mean(
      Expenditure, level = 0.95,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio")%>%
  arrange(desc(Zone))
```

Zone	Media	Media_se	Media_low	Media_upp
Urban	459.6	22.21	415.6	503.6
Rural	273.9	10.26	253.6	294.3
El gasto medio	370.5	13.29	344.2	396.9

Estimación de medias por sub-grupos

```
diseno %>% group_by(Zone, Sex) %>%
  cascade(
    Media = survey_mean(
      Expenditure, level = 0.95,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio") %>%
  arrange(desc(Zone), desc(Sex)) %>%
  data.frame()
```

Zone	Sex	Media	Media_se	Media_low	Media_upp
Urban	Male	469.8	26.96	416.4	523.2
Urban	Female	450.8	20.12	411.0	490.7
Urban	El gasto medio	459.6	22.21	415.6	503.6
Rural	Male	275.3	10.25	255.0	295.6
Rural	Female	272.7	11.61	249.7	295.7
Rural	El gasto medio	273.9	10.26	253.6	294.3
El gasto medio	El gasto medio	370.5	13.29	344.2	396.9

Estimación de medidas de dispersión y localización

- Es fundamental estimar medidas de dispersión en encuestas de hogares para comprender la variabilidad de las variables estudiadas.
- Una medida común es la desviación estándar, que permite medir qué tan disímiles son los ingresos medios de los hogares en un país.
- El estimador de la desviación estándar se puede expresar como:

$$s\left(y\right)_{\omega} = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}\left(y_{h\alpha i}-\bar{Y}_{\omega}\right)^2}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}-1}$$

.

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

La estimación de la desviación estándar se obtiene con `survey_var`

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

Zone	Sd	Sd_se	Sd_low	Sd_upp
Rural	310.3	117.4	262.6	351.6
Urban	581.9	285.0	421.6	706.7

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

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

Zone	Sex	Sd	Sd_se	Sd_low	Sd_upp
Rural	Female	294.9	111.6	249.6	334.1
Rural	Male	325.8	125.0	274.2	370.2
Urban	Female	568.4	286.5	400.7	696.8
Urban	Male	596.8	288.9	436.8	722.1

Estimación de la mediana

- Las medidas de posición no central, como la mediana, cuartiles y percentiles, son fundamentales en encuestas de hogares para comprender la distribución de las variables estudiadas.
- La mediana es una medida robusta de tendencia central que divide la población en dos partes iguales.
- La estimación de percentiles es esencial para definir políticas públicas, por ejemplo, para impuestos o subsidios.
- Los cuantiles se estiman utilizando la función de distribución acumulativa (CDF). El cuantil q-ésimo es el valor de y tal que la CDF es mayor o igual a q.

$$F(x) = \frac{\sum_{i=1}^{N} I(y_i \leq x)}{N}$$

Donde, $I(y_i \leq x)$ es una variable indicadora la cual es igual a 1 si y_i es menor o igual a un valor específico x, 0 en otro caso.

EStimación de a función de distribución acumulativa (CDF)

Un estimador de la CDF en un diseño complejo (encuesta de hogares) de tamaño n está dado por:

$$\hat{F}(x) = \frac{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}I(y_i \leq x)}{\sum_{h=1}^{H}\sum_{\alpha=1}^{a_h}\sum_{i=1}^{n_{h\alpha}}\omega_{h\alpha i}}$$

El cuantil q-ésimo de una variable y es el valor más pequeño de y tal que la CDF de la población es mayor o igual que q. Como es bien sabido, la mediana es aquel valor donde la CDF es mayor o igual a 0.5

Estimación de la mediana

Siguiendo las recomendaciones de *Heeringa et al (2017)* para estimar cuantiles, primero se considera las estadísticas de orden que se denotan como $y_1, \ldots, y_n$, y encuentra el valor de j $(j = 1, \ldots, n)$ tal que:

$$\hat{F}\left(y_j\right) \leq q \leq \hat{F}\left(y_{j+1}\right)$$

- La estimación del cuantil q-ésimo en un diseño complejo se calcula utilizando esta fórmula:

$$\hat{Y}_q = y_j + \frac{q - \hat{F}(y_j)}{\hat{F}(y_{j+1}) - \hat{F}(y_j)}(y_{j+1} - y_j)$$

Estimación de la mediana para gastos

La estimación de la median se obtiene con `survey_median`

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

Mediana	Mediana_se	Mediana_low	Mediana_upp
298.3	8.825	282.2	317.2

Estimación de la mediana por sub-grupo

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

Zone	Mediana	Mediana_se	Mediana_low	Mediana_upp
Rural	240.7	11.00	214.2	258.3
Urban	380.7	19.84	337.1	416.3

Estimación de la mediana por sub-grupo

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

Sex	Mediana	Mediana_se	Mediana_low	Mediana_upp
Female	299.9	10.499	282.2	323.8
Male	297.3	9.287	277.3	314.1

Estimación del quantil 0.5 para el gasto

La estimación de la median se obtiene con `survey_quantile`

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

Q_q50	Q_q50_se	Q_q50_low	Q_q50_upp
298.3	11.96	264.8	312.1

Estimación del quantil 0.25 para el gasto por sub-grupo

```r
diseno %>% group_by(Sex) %>%
  summarise(
    Q =  survey_quantile(
    Expenditure,
    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
Female	209.7	14.91	169.0	228.1
Male	192.5	10.41	163.5	204.7

Estimación del quantile 0.25 para el gasto por sub-grupo

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

Zone	Q_q25	Q_q25_se	Q_q25_low	Q_q25_upp
Rural	159.9	4.641	144.9	163.5
Urban	258.2	9.048	256.0	292.2

Estimando razones en encuestas de hogares

- La razón poblacional es el cociente de dos totales poblacionales de características de interés, como la cantidad de hombres por cada mujer en un país.
- Para estimar esta razón en encuestas de hogares, se calculan por separado los totales de las variables de interés.
- El estimador puntual de la razón se define como el cociente de los totales estimados:

$$\hat{R} = \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{\alpha_h} \sum_{i=1}^{nh\alpha} \omega_{h\alpha i} y_{h\alpha i}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{\alpha_h} \sum_{i=1}^{nh\alpha} \omega_{h\alpha i} x_{h\alpha i}}$$

- Sin embargo, el cálculo de la varianza de este estimador no es trivial, por lo que se requiere aplicar la técnica de linealización de Taylor *(Gutiérrez, 2016)*.

Estimación de la razón entre el gasto y el ingreso

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

```
diseno %>% summarise(
    Razon =  survey_ratio(
      numerator = Expenditure,
      denominator = Income,
      level = 0.95,
    vartype =  c("se", "ci")
    ))
```

Razon	Razon_se	Razon_low	Razon_upp
0.649	0.0232	0.6031	0.6949

Estimación de la razón entre hombres y mujeres

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

Razon	Razon_se	Razon_low	Razon_upp
1.114	0.0351	1.045	1.184

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

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

Razon	Razon_se	Razon_low	Razon_upp
1.068	0.0352	0.9975	1.139

Estimación de la razón del gastos y los ingreso entre las mujeres

```
sub_Mujer %>% summarise(
    Razon =  survey_ratio(
      numerator = Expenditure,
      denominator = Income,
      level = 0.95,
    vartype =  c("se", "ci")
    ))
```

Razon	Razon_se	Razon_low	Razon_upp
0.6583	0.0199	0.619	0.6976

Estimación de la razón del gasto y los ingresos entre los hombres

```
sub_Hombre %>% summarise(
    Razon =  survey_ratio(
      numerator = Expenditure,
      denominator = Income,
      level = 0.95,
    vartype =  c("se", "ci")
    ))
```

Razon	Razon_se	Razon_low	Razon_upp
0.6391	0.0288	0.5821	0.696

Estimación del coeficiente de Gini en encuestas de hogares

Reflexión

Definir lo justo siempre será difícil y es algo a lo que quizá sea poco realista aspirar a conseguir. Sin embargo si estamos un poco más conscientes de cómo la desigualdad afecta nuestra libertad y cómo se refleja en el bienestar y calidad de vida de las personas, podremos poner en contexto una discusión que tendremos cada vez más presente en el mundo y en el país.

Índice de Gini

- El índice de Gini es un indicador ampliamente utilizado para medir la desigualdad económica en los hogares de un país. Su valor oscila entre 0 y 1, donde 0 representa una distribución perfectamente igualitaria y valores más altos indican una creciente desigualdad en la distribución de la riqueza.
- El estimador del coeficiente de Gini, según *Binder y Kovacevic (1995)*, se define como:

$$\widehat{G}(y) = \frac{2 \times \sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n_{h\alpha}} \omega^*_{h\alpha i} \widehat{F}_{h\alpha i} y_{h\alpha i} - 1}{\bar{y}_\omega}$$

Donde: - $\omega^*_{h\alpha i}$ es el peso de diseño normalizado.

- $\widehat{F}_{h\alpha i}$ es la estimación de la función de distribución acumulativa en el conglomerado α del estrato h.
- $\bar{y}_\omega$ es la estimación del promedio de la variable de interés.

Estimación del índice de GINI

La estimación del índice de GINI se realiza haciendo uso de la librería `convey`, para ello se procede así:

```
library(convey)
## Definir el diseño
diseno_gini <- convey_prep(diseno)
## Calculo del indice para el ingreso
svygini( ~Income, design = diseno_gini) %>%
  data.frame()
```

	gini	Income
Income	0.4133	0.0187

Estimación del índice de GINI

En forma análoga es posible obtener el indice de GINI para el gasto.

```
svygini( ~Expenditure, design = diseno_gini) %>%
  data.frame()
```

	gini	Expenditure
Expenditure	0.3509	0.0141

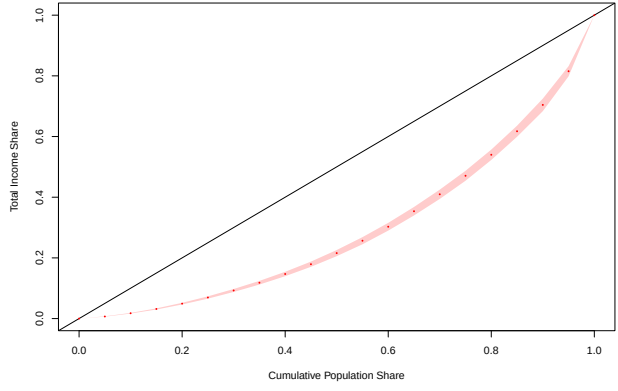
Curva de Lorenz

- La **curva de Lorenz** es una herramienta fundamental para analizar la desigualdad en la distribución de ingresos en una población. Esta curva representa el porcentaje acumulado de la población, ordenada de menor a mayor ingreso, frente a su participación en el ingreso total. Cuanto más cerca esté la curva de Lorenz de la línea de 45 grados, más equitativa es la distribución de ingresos.
- El área entre la curva de Lorenz y la línea de 45 grados se conoce como el **área de Lorenz**. El índice de Gini es igual al doble del área de Lorenz. Si todos los ingresos son iguales, la curva de Lorenz se convierte en una línea de 45 grados.

Estimación del curva de Lorenz.

La **curva de Lorenz** es una representación gráfica de la desigualdad en la distribución de la renta, para obtener la representación gráfica de está usamos la función `svylorenz`.

```
svylorenz( ~Income, diseno_gini,
           seq(0,1,.05), alpha = .01 )
```



Prueba de hipótesis para la diferencia de medias en encuestas de hogares

Prueba de Hipótesis:

Se plantean dos hipótesis: nula (H0) y alternativa (H1).

$$\begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta \neq \theta_0 \end{cases} \quad \begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta > \theta_0 \end{cases} \quad \begin{cases} H_0: & \theta = \theta_0 \\ H_1: & \theta < \theta_0 \end{cases}$$

El proceso de selección entre las dos hipótesis se llama prueba de hipótesis.

Combinaciones Lineales de Estadísticas Descriptivas

- Parámetros importantes se expresan como combinaciones lineales de medidas descriptivas.
- Ejemplo: suma ponderada de medias para construir índices económicos, es decir, la función de combinación lineal: $f(\theta_1, \theta_2, ..., \theta_j) = \sum_{j=1}^{J} a_j \theta_j$
- Estimación de la función:

$$f(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_j) = \sum_{j=1}^{J} a_j \hat{\theta}_j$$

Varianza del Estimador:

- La varianza del estimador se calcula como:

$$var\left(\sum_{j=1}^{J} a_j \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 var\left(\hat{\theta}_j\right) + 2\sum_{j=1}^{J-1}\sum_{k>j}^{J} a_j a_k \, cov\left(\hat{\theta}_j, \hat{\theta}_k\right)$$

Diferencia de Medias y Prueba de Hipótesis

La diferencia de medias se expresa como $\bar{Y}_1 - \bar{Y}_2$.

- Ejemplo: Diferencia entre los ingresos medios de padres e ingresos medios de madres en un hogar.

Hipótesis

- H_0: No hay diferencia entre las medias ($H_0 : \bar{Y}_1 - \bar{Y}_2 = 0$).
- H_1: Existe diferencia entre las medias:

$$\begin{cases} H_0 : \bar{Y}_1 - \bar{Y}_2 = 0 \\ H_1 : \bar{Y}_1 - \bar{Y}_2 \neq 0 \end{cases} \qquad \begin{cases} H_0 : \bar{Y}_1 - \bar{Y}_2 = 0 \\ H_1 : \bar{Y}_1 - \bar{Y}_2 > 0 \end{cases} \qquad \begin{cases} H_0 : \bar{Y}_1 - \bar{Y}_2 = 0 \\ H_1 : \bar{Y}_1 - \bar{Y}_2 < 0 \end{cases}$$

Estadístico de Prueba t

- El estadístico de prueba t se utiliza para probar las hipótesis y se distribuye como una t-Student.
- Fórmula del estadístico de prueba t:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{se(\bar{Y}_1 - \bar{Y}_2)}$$

- Donde $se(\bar{Y}_1 - \bar{Y}_2)$ es la desviación estándar de la diferencia de medias:

$$se(\bar{Y}_1 - \bar{Y}_2) = \sqrt{var(\bar{y}_1) + var(\bar{y}_2) - 2cov(\bar{y}_1, \bar{y}_2)}$$

Intervalo de Confianza para la Diferencia de Medias

- Para construir un intervalo de confianza para la diferencia de medias:

$$\left(\bar{Y}_1 - \bar{Y}_2\right) \pm t_{gl,\alpha/2}\, se(\bar{Y}_1 - \bar{Y}_2)$$

- Las pruebas de hipótesis y los intervalos de confianza son herramientas clave para la toma de decisiones y evaluación en estadísticas.

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 = 1.4, df = 118, p-value = 0.2
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -12.82  69.39
sample estimates:
difference in mean
             28.28
```

El resultando indica que no hay diferencia entre los ingreso medios.

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.

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

```
    Design-based t-test

data:  Income ~ Sex
t = 1.6, df = 63, p-value = 0.1
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -12.32 101.74
sample estimates:
difference in mean
             44.71
```

El resultando indica que no hay diferencia entre los ingreso medios.

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

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

```
    Design-based t-test

data:  Income ~ Sex
t = 1.5, df = 118, p-value = 0.1
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -10.73  82.85
sample estimates:
difference in mean
             36.06
```

Contrastes en Encuestas de Hogares

- En encuestas de hogares, a menudo se necesita comparar más de dos poblaciones simultáneamente.
- Ejemplo: Comparar los ingresos medios en 3 regiones o municipalidades en la postpandemia para evaluar el impacto de COVID-19 en los hogares.
- La diferencia de medias, utilizada previamente, es limitada para comparar solo dos poblaciones.
- Los contrastes ofrecen una solución efectiva para abordar problemas de comparación múltiple en encuestas de hogares.

Contrastes y Combinaciones Lineales de Parámetros

- Un contraste es una combinación lineal de parámetros:

$$f(\theta_1, \theta_2, ..., \theta_j) = \sum_{j=1}^{J} a_j \theta_j$$

- La estimación de esta función se expresa como:

$$f(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_j) = \sum_{j=1}^{J} a_j \hat{\theta}_j$$

- La varianza del estimador se calcula de la siguiente manera:

$$var\left(\sum_{j=1}^{J} a_j \hat{\theta}_j\right) = \sum_{j=1}^{J} a_j^2 var(\hat{\theta}_j) + 2 \sum_{j=1}^{J-1} \sum_{k>j}^{J} a_j a_k \, cov(\hat{\theta}_j, \hat{\theta}_k)$$

Contrastes

Ahora, el interés es realizar contrastes entre más de dos subpobaciones, por ejemplo por regiones geográficas.

	Region	Income	se	ci_l	ci_u
Norte	Norte	552.4	55.36	443.9	660.9
Sur	Sur	625.8	62.41	503.5	748.1
Centro	Centro	650.8	61.47	530.3	771.3
Occidente	Occidente	517.0	46.22	426.4	607.6
Oriente	Oriente	541.8	71.66	401.3	682.2

Por ejemplo, la diferencia media entre las regiones Norte y Sur $\hat{\bar{y}}_{Norte} - \hat{\bar{y}}_{Sur}$

Procedimiento para realizar los contrastes

```
# Paso 1: diferencia de estimaciones (Norte - Sur)
552.4 - 625.8
```

```
[1] -73.4
```

```
# Paso 2: error estándar de la diferencia
vcov(prom_region)
```

	Norte	Sur	Centro	Occidente	Oriente
Norte	3065	0	0	0	0
Sur	0	3894	0	0	0
Centro	0	0	3778	0	0
Occidente	0	0	0	2136	0
Oriente	0	0	0	0	5136

```
sqrt(3065 + 3894 - 2*0)
```

```
[1] 83.42
```

Procedimiento para realizar los contrastes

El procedimiento anterior se reduce a la sintaxis:

```
svycontrast(prom_region,
            list(diff_NS = c(1, -1, 0, 0, 0))) %>%
  data.frame()
```

	contrast	diff_NS
diff_NS	-73.41	83.42

Creado una matriz de contrastes

Ahora el interés es realizar los contrastes siguientes:

- $\hat{\bar{y}}_{Norte} - \times \hat{\bar{y}}_{Centro'}$
- $\hat{\bar{y}}_{Sur} - \hat{\bar{y}}_{Centro}$
- $\hat{\bar{y}}_{Occidente} - \hat{\bar{y}}_{Oriente}$

Escrita de forma matricial es:

$$\begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

Creado una matriz de contrastes en R

```
svycontrast(prom_region, list(
           Norte_sur = c(1, 0, -1, 0, 0),
         Sur_centro = c(0, 1, -1, 0, 0),
   Occidente_Oriente = c(0, 0, 0, 1, -1)
           )) %>% data.frame()
```

	contrast	SE
Norte_sur	-98.42	82.72
Sur_centro	-25.01	87.60
Occidente_Oriente	-24.75	85.28

```
c(sqrt(3065 + 3778 - 2*0), sqrt(3894 + 3778 - 2*0),
  sqrt(2136 + 5136 - 2*0))
```

```
[1] 82.72 87.59 85.28
```

Contrastes no independiente

Es posible que las variables estén correlacionadas.Por ejemplo, Ingreso y Sexo.

```
(prom_sexo <-
   svyby(~Income, ~Sex, diseno,
         svymean, na.rm=T,covmat = TRUE,
         vartype = c("se", "ci")))
```

	Sex	Income	se	ci_l	ci_u
Female	Female	557.6	25.83	506.9	608.2
Male	Male	585.8	34.59	518.1	653.6

Contrastes no independiente

El contraste

$$\hat{\bar{y}}_F - \hat{\bar{y}}_M$$

Es calculado como sigue:

```
svycontrast(prom_sexo,
            list(diff_Sexo = c(1, -1))) %>%
  data.frame()
```

	contrast	diff_Sexo
diff_Sexo	-28.28	20.76

Contrastes no independiente

```
vcov(prom_sexo)
```

	Female	Male
Female	667.2	716.3
Male	716.3	1196.3

```
# Note que el error estándar de la diff  es igual a
sqrt(667.2 + 1196.3 - 2*716.3)
```

```
[1] 20.76
```

Contrastes no independiente

Otra posibilidad es poder obtener resultados agregados, por ejemplo:

$$\hat{\bar{y}}_{Norte} + \hat{\bar{y}}_{Sur} + \hat{\bar{y}}_{Centro}$$

```
(sum_region <- svyby( ~ Income,  ~ Region,
                      diseno, svytotal, na.rm = T,
                      covmat = TRUE,
                      vartype = c("se", "ci")))
```

	Region	Income	se	ci_l	ci_u
Norte	Norte	14277323	1507575	11322530	17232115
Sur	Sur	16068151	1877989	12387359	19748942
Centro	Centro	16483319	2383556	11811634	21155003
Occidente	Occidente	16853540	1823807	13278944	20428135
Oriente	Oriente	22111335	2833460	16557856	27664814

Contrastes no independiente

La matriz de contraste queda como:

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

el procedimiento en R es:

```
svycontrast(sum_region,
            list(
              Agregado_NCS = c(1, 1, 1, 0, 0)
              )) %>% data.frame()
```

	contrast	Agregado_NCS
Agregado_NCS	46828792	3388357

Contrastes

```
require(kableExtra)
# Note que el error estándar  de la dif. es igual a
vcov(sum_region) %>% data.frame() %>% kable(digits = 10,
        format.args = list(scientific = FALSE)) %>%
  kable_styling(full_width = FALSE)
```

	Norte	Sur	Centro	Occidente	Oriente
Norte	2272782099289	0	0	0	0
Sur	0	3526843231468	0	0	0
Centro	0	0	5681340267222	0	0
Occidente	0	0	0	3326270307526	0
Oriente	0	0	0	0	8028493876790

```
sqrt(2272782099289 + 3526843231468 + 5681340267222 )
```

```
[1] 3388357
```

Contrastes no independiente

La función puede usarse para obtener los promedios por categorías. Por ejemplo:

$$\hat{\bar{y}}_{Edad} = \frac{1}{k} \sum_{k=1}^{K} \hat{\bar{y}}_k$$

donde K es el número de categorías de la variable.

Contrastes no independiente

```
(prom_edad <- svyby(~Income, ~CatAge, diseno,
                    svymean, na.rm=T,covmat = TRUE))
```

	CatAge	Income	se
0-5	0-5	463.8	28.87
6-15	6-15	511.6	34.88
16-30	16-30	607.3	37.42
31-45	31-45	573.4	26.95
46-60	46-60	763.1	58.97
Más de 60	Más de 60	466.6	31.21

Contrastes no independiente

La matriz de contraste estaría dada por:

$$\begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}$$

El procedimiento en R es:

```
svycontrast(prom_edad,
  list(
  agregado_edad = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)
      )) %>% data.frame()
```

	contrast	agregado_edad
agregado_edad	564.3	25.4

Contrastes no independiente

```
vcov(prom_edad)
```

	0-5	6-15	16-30	31-45	46-60	Más de 60
0-5	833.4	548.4	361.1	262.3	132.7	312.6
6-15	548.4	1216.6	739.7	528.1	565.5	120.1
16-30	361.1	739.7	1399.9	534.9	1564.6	412.5
31-45	262.3	528.1	534.9	726.2	642.3	161.5
46-60	132.7	565.5	1564.6	642.3	3477.7	416.6
Más de 60	312.6	120.1	412.5	161.5	416.6	973.9

```
(1 / 6)*sqrt(
  833.4 + 1216.6 + 1399.9 + 726.2 + 3477.7 + 973.9 +
    2*548.4 + 2*361.1  + 2*262.3 + 2*132.7 + 2*312.6 +
    2*739.7 + 2*528.1  + 2*565.5 + 2*120.1 +
    2*534.9 + 2*1564.6 + 2*412.5 +
    2*642.3 + 2*161.5  +
    2*416.6)
```

```
[1] 25.4
```

Contrastes no independiente

```
(razon_sexo <- svyby(~Income, ~Sex,
                     denominator = ~Expenditure,
                     diseno, svyratio,
                     na.rm=T, covmat = TRUE,
                     vartype = c("se", "ci")))
```

	Sex	Income/Expenditure	se.Income/Expenditure	ci_l	ci_u
Female	Female	1.519	0.0458	1.429	1.609
Male	Male	1.565	0.0704	1.427	1.703

Contrastes no independiente

```
svycontrast(razon_sexo,
            list(
              diff_sexo = c(1, -1)
              )) %>% data.frame()
```

	contrast	diff_sexo
diff_sexo	-0.0457	0.0416

Contrastes no independiente

```
vcov(razon_sexo)
```

	Female	Male
Female	0.0021	0.0027
Male	0.0027	0.0050

```
sqrt(0.0021 + 0.0050 - 2*0.0027)
```

```
[1] 0.04123
```

¡Gracias!

Email: andres.gutierrez@cepal.org