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("Imagenes/02_variable_continua/ENIGH_HND_Pers.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.
  mutate(estrato = haven::as_factor(F1_A0_ESTRATO),
         Sexo = F2_A6_P3_SEXO,
         Area = F1_A0_AREA) %>%
  as_survey_design(
    strata = estrato,  # Id de los estratos.
    ids = F1_A0_UPM,          # Id para las observaciones.
    weights = Factor,       # Factores de expansión.
    nest = TRUE            # Valida el anidado dentro del estrato
  )
```

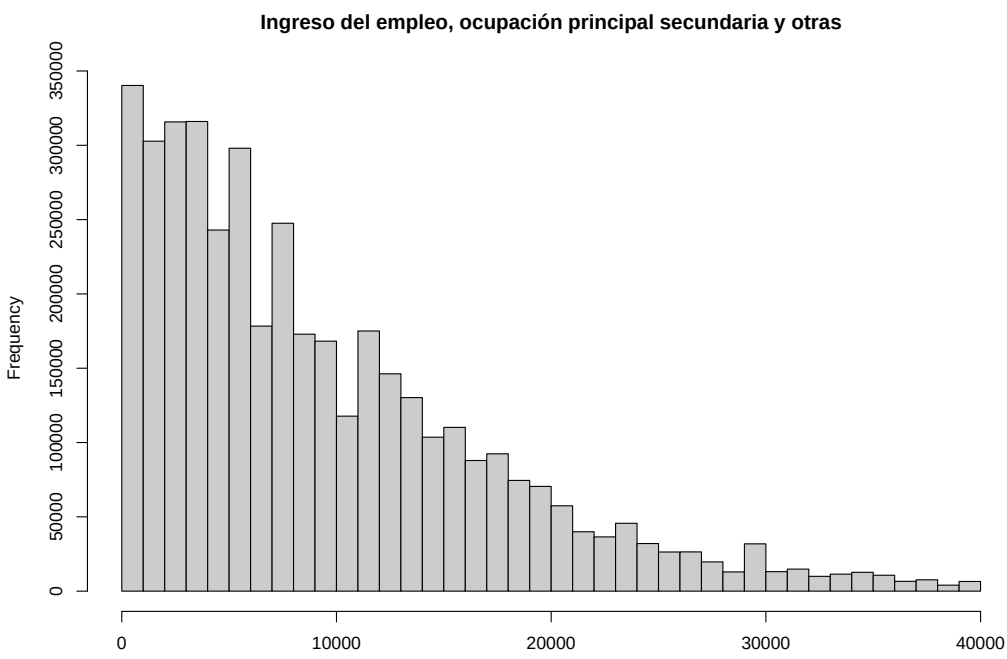
Análisis gráfico

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`

```
# 14 registros con valores menores que cero
svyhist(
  ~ YEMPLEO, #  Ingreso del empleo, ocupación principal secundaria y otras
  diseno %>% filter(YEMPLEO < 40000, YEMPLEO > 0),
  main = "Ingreso del empleo, ocupación principal secundaria y otras",
  col = "grey80", breaks = 50,
  xlab = "Ingreso",
  probability = FALSE
)
```

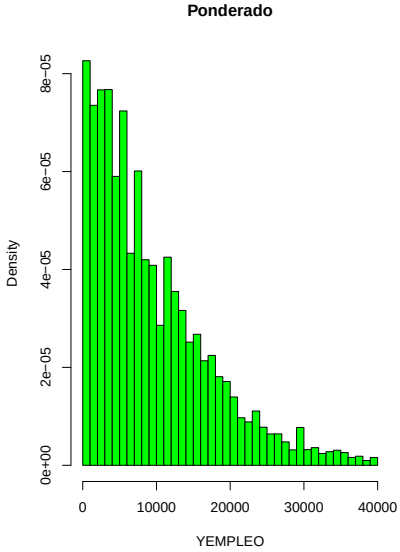
Histograma ponderado para la variable ingreso

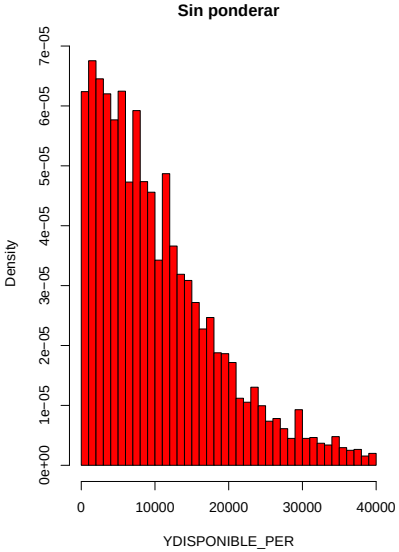


Comparación de histogramas

```
par(mfrow = c(1,2))
svyhist(   ~ YEMPLEO,
  diseno  %>% filter(YEMPLEO < 40000, YEMPLEO > 0),
  main = "Ponderado", col = "green", breaks = 50
)
hist( encuesta$YEMPLEO[encuesta$YEMPLEO < 40000 &
                         encuesta$YEMPLEO > 0],
  main = "Sin ponderar", xlab = "YDISPONIBLE_PER",
  col = "red", prob = TRUE, breaks = 50
)
```

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(Area == 1) # 1. Urbana
sub_Rural  <- diseno %>%  filter(Area == 2) # 2. Rural
sub_Mujer  <- diseno %>%  filter(Sexo == 2) # 2. Mujer
sub_Hombre <- diseno %>%  filter(Sexo == 1) # 1. Hombre
```

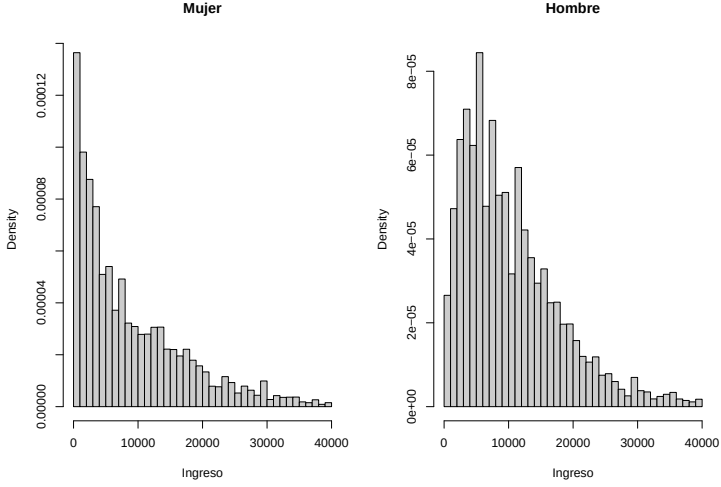
Histograma ponderado en sub-grupos

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

```
par(mfrow = c(1,2))
svyhist(
  ~ YEMPLEO ,
  design = sub_Mujer %>% filter(YEMPLEO < 40000, YEMPLEO > 0,
                                F2_A6_P4_EDAD >= 18),
  main = "Mujer",
  breaks = 30,
  col = "grey80",
  xlab = "Ingreso"
)


svyhist(
  ~ YEMPLEO ,
  design =sub_Hombre %>% filter(YEMPLEO < 40000, YEMPLEO > 0,
                                F2_A6_P4_EDAD >= 18),
  main = "Hombre",
  breaks = 30,
  col = "grey80",
  xlab = "Ingreso"
)
```

Histograma ponderado en sub-grupos

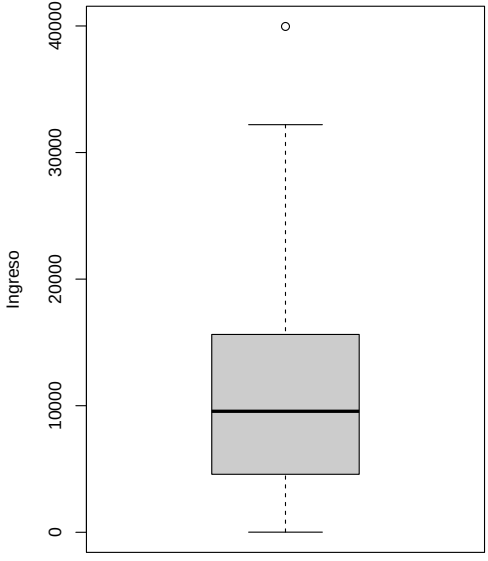


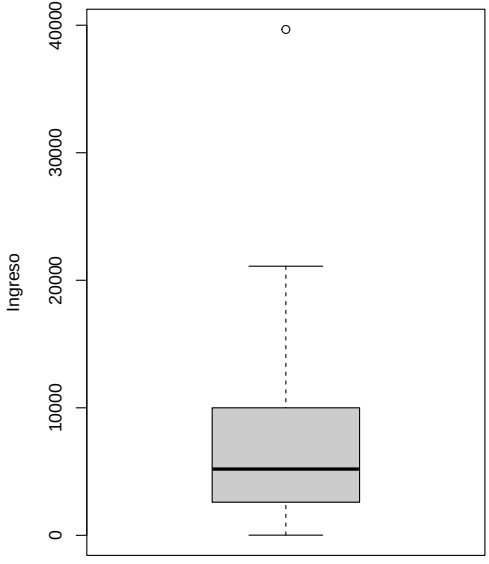
Boxplot ponderado del ingreso por sub-grupos

```r
par(mfrow = c(1,2))
svyboxplot(
  YEMPLEO ~1 ,
  sub_Urbano %>% filter(YEMPLEO < 40000, YEMPLEO > 0),
  col = "grey80",
  ylab = "Ingreso",
  xlab = "Urbano")

svyboxplot(
  YEMPLEO ~ 1 ,
  sub_Rural%>% filter(YEMPLEO < 40000, YEMPLEO > 0),
  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(~YEMPLEO, diseno, deff=T) %>%
  data.frame() %>% tba()
```

	total	YEMPLEO	deff
YEMPLEO	51098431067	1786411294	4.461

```
confint(svytotal (~YEMPLEO, diseno, deff=T)) %>% tba()
```

	2.5 %	97.5 %
YEMPLEO	47597129270	54599732865

Estimación de totales e intervalos de confianza del gasto

```
svytotal ( ~ COSTOALQUI, diseno, deff = T, na.rm = TRUE) %>%
  data.frame() %>% tba()
```

	total	COSTOALQUI	deff
COSTOALQUI	226385712	15952512	3.785

```
confint(svytotal (
  ~ COSTOALQUI,
  diseno,
  deff = T,
  na.rm = TRUE
)) %>% tba()
```

	2.5 %	97.5 %
COSTOALQUI	195119364	257652060

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 %>% mutate(Sexo = haven::as_factor(Sexo)) %>%
  group_by(Sexo) %>%
  cascade(Total = survey_total(YEMPLEO, level = 0.95,
                               vartype =  c("se", "ci")),
          .fill = "Total ingreso") %>% tba()
```

Sexo	Total	Total_se	Total_low	Total_upp
1. Hombre	32819599238	1355038968	30159330557	35479867919
2. Mujer	18278831829	717302754	16870593323	19687070334
Total ingreso	51098431067	1786411294	47591274352	54605587783

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(~YEMPLEO, diseno, deff=T) %>%
  data.frame() %>% tba()
```

	mean	YEMPLEO	deff
YEMPLEO	5222	169.1	3.827

```
confint(svymean (~YEMPLEO, diseno, deff=T)) %>% tba()
```

	2.5 %	97.5 %
YEMPLEO	4890	5553

Estimación de la media e intervalo de confianza del gasto

```
svymean (~COSTOALQUI, diseno, deff=T,na.rm = TRUE) %>%
  data.frame() %>% tba()
```

	mean	COSTOALQUI	deff
COSTOALQUI	86.46	5.382	2.954

```
confint(svymean (~COSTOALQUI, diseno, deff=T,na.rm = TRUE))%>% tba()
```

	2.5 %	97.5 %
COSTOALQUI	75.91	97.01

Estimación de la media por sub-grupos

La función cascade regresa el resultado promedio ignorando los niveles.

```
diseno <- diseno %>%
  mutate(Sexo = haven::as_factor(F2_A6_P3_SEXO))
diseno %>% group_by(Sexo) %>%
  cascade(
    Media = survey_mean(
      COSTOALQUI, level = 0.95,na.rm = TRUE,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio"  ) %>%
  arrange(desc(Sexo)) %>% tba() # Ordena la variable.
```

Sexo	Media	Media_se	Media_low	Media_upp
El gasto medio	86.46	5.382	75.90	97.03
2. Mujer	99.92	11.525	77.29	122.54
1. Hombre	78.46	4.633	69.36	87.55

Estimación de la media por sub-grupos

```
diseno <- diseno %>%
  mutate(Area = haven::as_factor(F1_A0_AREA))
diseno %>% group_by(Area) %>%
  cascade(
    Media = survey_mean(
      COSTOALQUI, level = 0.95, na.rm = TRUE,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio")%>%
  arrange(desc(Area))  %>% tba()
```

Area	Media	Media_se	Media_low	Media_upp
El gasto medio	86.46	5.382	75.90	97.03
2. Rural	84.44	10.349	64.12	104.75
1. Urbana	87.96	5.443	77.27	98.64

Estimación de medias por sub-grupos

```
diseno %>% group_by(Area, Sexo) %>%
  cascade(
    Media = survey_mean(
      COSTOALQUI, level = 0.95,na.rm = TRUE,
       vartype =  c("se", "ci")),
        .fill = "El gasto medio") %>%
  arrange(desc(Area), desc(Sexo)) %>%
  data.frame()  %>% tba()
```

Area	Sexo	Media	Media_se	Media_low	Media_upp
El gasto medio	El gasto medio	86.46	5.382	75.90	97.03
2. Rural	El gasto medio	84.44	10.349	64.12	104.75
2. Rural	2. Mujer	111.14	27.019	58.10	164.19
2. Rural	1. Hombre	72.32	7.192	58.20	86.44
1. Urbana	El gasto medio	87.96	5.443	77.27	98.64
1. Urbana	2. Mujer	93.75	9.941	74.23	113.26
1. Urbana	1. Hombre	83.79	6.007	72.00	95.59

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(Area) %>%
   summarise(Sd = sqrt(
  survey_var(
    YEMPLEO,
    level = 0.95,
    vartype =  c("se", "ci"),
  ) )))  %>% tba()
```

Area	Sd	Sd_se	Sd_low	Sd_upp
1. Urbana	16104	6996	12776	18852
2. Rural	14594	8426	8579	18772

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

```
(tab_sd <- diseno %>% group_by(Area, Sexo) %>%
   summarise(Sd = sqrt(
  survey_var(
    YEMPLEO,
    level = 0.95,
    vartype =  c("se", "ci"),
   )
))) %>% data.frame() %>% tba()
```

Area	Sexo	Sd	Sd_se	Sd_low	Sd_upp
1. Urbana	1. Hombre	19464	9861	13709	23870
1. Urbana	2. Mujer	12249	5342	9696	14356
2. Rural	1. Hombre	16666	10647	7432	22368
2. Rural	2. Mujer	12078	9202	NaN	17667

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 %>% filter(COSTOALQUI > 0) %>%
  summarise(Mediana =
  survey_median(
    COSTOALQUI, na.rm = TRUE,
    level = 0.95,
    vartype =  c("se", "ci"),
   )) %>% tba()
```

Mediana	Mediana_se	Mediana_low	Mediana_upp
173.3	15.17	148.8	208.3

Estimación de la mediana por sub-grupo

```
diseno %>% filter(COSTOALQUI > 0) %>%
  group_by(Area) %>%
  summarise(Mediana =
  survey_median(
    COSTOALQUI,
    level = 0.95,  na.rm = TRUE,
    vartype =  c("se", "ci"),
   )) %>% tba()
```

Area	Mediana	Mediana_se	Mediana_low	Mediana_upp
1. Urbana	191.2	11.03	166.7	210
2. Rural	166.7	31.49	125.0	250

Estimación de la mediana del salario por sub-grupo

```
diseno %>% filter(SALARIO_IMPUT_2 > 0) %>%
  group_by(Sexo) %>%
  summarise(Mediana =
  survey_median(
    SALARIO_IMPUT_2,
    level = 0.95, na.rm = TRUE,
    vartype =  c("se", "ci"),
   )) %>% tba()
```

Sexo	Mediana	Mediana_se	Mediana_low	Mediana_upp
1. Hombre	12860	2901	10195	23322
2. Mujer	14800	4778	11318	32935

Estimación del quantil 0.5 para el salario

La estimación de la median se obtiene con `survey_quantile`

```
diseno %>% filter(!is.na(SALARIO_IMPUT_2)) %>%
  summarise(
    Q =  survey_quantile(
    SALARIO_IMPUT_2,
    quantiles = 0.5,
    level = 0.95,
    vartype =  c("se", "ci"),
    interval_type = "score"
   )) %>% tba()
```

Q_q50	Q_q50_se	Q_q50_low	Q_q50_upp
13900	1589	12000	18629

Estimación del quantil 0.25 para el salario por sub-grupo

```
diseno %>% filter(!is.na(SALARIO_IMPUT_2)) %>%
  group_by(Sexo) %>%
  summarise(
    Q =  survey_quantile(
    SALARIO_IMPUT_2,
    quantiles = 0.25,
    level = 0.95,
    vartype =  c("se", "ci"),
    interval_type = "score"
   ))  %>% tba()
```

Sexo	Q_q25	Q_q25_se	Q_q25_low	Q_q25_upp
1. Hombre	9360	977.1	8439	12860
2. Mujer	11318	1158.2	9560	14800

Estimación del quantile 0.25 por sub-grupo

```
diseno %>% filter(YEMPLEO > 0) %>%
  group_by(Area) %>%
  summarise(
    Q =  survey_quantile(
    YEMPLEO,
    quantiles = 0.25,
    level = 0.95,
    vartype =  c("se", "ci"),
    interval_type = "score"
   ))  %>% tba()
```

Area	Q_q25	Q_q25_se	Q_q25_low	Q_q25_upp
1. Urbana	4800	134.4	4668	5196
2. Rural	2598	227.1	2099	3000

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 %>%  filter(YEMPLEO > 0, COSTOALQUI > 0) %>%
   summarise(
    Razon =  survey_ratio(
      numerator = COSTOALQUI,
      denominator = YEMPLEO,
      level = 0.95,
    vartype =  c("se", "ci")
    ))  %>% tba()
```

Razon	Razon_se	Razon_low	Razon_upp
0.01818	0.001134	0.01596	0.02041

Estimación de la razón entre hombres y mujeres

```
diseno %>% summarise(
    Razon =  survey_ratio(
      numerator = (Sexo == "2. Mujer"),# creando dummy.
      denominator = (Sexo == "1. Hombre"),# creando dummy.
      level = 0.95,
    vartype =  c("se", "ci")
    ))  %>% tba()
```

Razon	Razon_se	Razon_low	Razon_upp
1.112	0.01445	1.083	1.14

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

```
sub_Rural <- sub_Rural %>%
  mutate(Sexo = haven::as_factor(F2_A6_P3_SEXO))
sub_Rural %>% summarise(
    Razon =  survey_ratio(
    numerator = (Sexo == "2. Mujer"),# creando dummy.
    denominator = (Sexo == "1. Hombre"),# creando dummy.
      level = 0.95,
    vartype =  c("se", "ci")
    ))  %>% tba()
```

Razon	Razon_se	Razon_low	Razon_upp
1.047	0.02534	0.9964	1.097

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

```
sub_Mujer %>% filter(YEMPLEO > 0, COSTOALQUI > 0) %>%
  summarise(
    Razon =  survey_ratio(
      numerator = COSTOALQUI,
      denominator = YEMPLEO,
      level = 0.95,
    vartype =  c("se", "ci")
    ))  %>% tba()
```

Razon	Razon_se	Razon_low	Razon_upp
0.02478	0.003537	0.01782	0.03173

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

```
sub_Hombre %>%  filter(YEMPLEO > 0, COSTOALQUI > 0) %>%
  summarise(
    Razon =  survey_ratio(
      numerator = COSTOALQUI,
      denominator = YEMPLEO,
      level = 0.95,
    vartype =  c("se", "ci")
    ))  %>% tba()
```

Razon	Razon_se	Razon_low	Razon_upp
0.01613	0.00163	0.01293	0.01933

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:

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

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

- $\hat{F}_{h\alpha i}$ 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( ~YEMPLEO,
         design = diseno_gini %>% filter(YEMPLEO > 0)) %>%
  data.frame() %>% tba()
```

	gini	YEMPLEO
YEMPLEO	0.5329	0.009716

Estimación del índice de GINI

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

```
svygini( ~COSTOALQUI,
         design = diseno_gini %>% filter(COSTOALQUI > 0)) %>%
  data.frame() %>% tba()
```

	gini	COSTOALQUI
COSTOALQUI	0.6017	0.01351

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( ~YEMPLEO, diseno_gini %>% filter(YEMPLEO > 0),
           seq(0,1,.05), alpha = .01 )
```

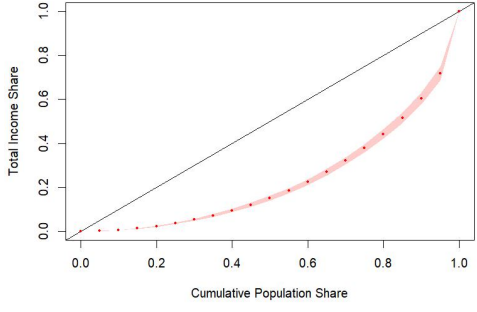


Figura 1: *curva de Lorenz*

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:

```
diseno <- diseno %>% filter(YEMPLEO > 0)
svyttest(YEMPLEO ~ Sexo, diseno)
```

```
    Design-based t-test

data:  YEMPLEO ~ Sexo
t = -5.1, df = 724, p-value = 4e-07
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -3144 -1406
sample estimates:
difference in mean
             -2275
```

El resultando indica que 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.

```
sub_Urbano <- sub_Urbano %>% filter(YEMPLEO > 0)
svyttest(YEMPLEO ~ Sexo, sub_Urbano)
```

```
    Design-based t-test

data:  YEMPLEO ~ Sexo
t = -9.5, df = 624, p-value <2e-16
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -4242 -2790
sample estimates:
difference in mean
             -3516
```

El resultando indica que hay diferencia entre los ingreso medios.

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

```
svyttest(YEMPLEO ~ Sexo, diseno %>% filter(F2_A6_P4_EDAD  > 18))
```

```
    Design-based t-test

data:  YEMPLEO ~ Sexo
t = -5.8, df = 724, p-value = 8e-09
alternative hypothesis: true difference in mean is not equal to 0
95 percent confidence interval:
 -3736 -1856
sample estimates:
difference in mean
             -2796
```

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 departamento.

	dam	YEMPLEO	se	ci_l	ci_u
1. Atlántida	1. Atlántida	12218	1564.7	9151	15284
2. Colon	2. Colon	10428	982.5	8503	12354
3. Comayagua	3. Comayagua	11674	937.5	9836	13511
4. Copan	4. Copan	11865	2087.1	7774	15955
5. Cortes	5. Cortes	13666	691.6	12310	15021
6. Choluteca	6. Choluteca	13662	3244.5	7303	20021
7. El Paraíso	7. El Paraíso	10677	2239.8	6287	15066
8. Francisco Morazán	8. Francisco Morazán	14302	725.4	12880	15723
9. Gracias A Dios	9. Gracias A Dios	7099	710.3	5707	8491
10. Intibuca	10. Intibuca	5003	1130.6	2787	7219
11. Islas De La Bahía	11. Islas De La Bahía	19366	3018.0	13451	25281
12. La Paz	12. La Paz	8005	1066.2	5915	10095
13. Lempira	13. Lempira	10520	1613.2	7358	13681
14. Ocotepeque	14. Ocotepeque	6178	1485.0	3267	9088
15. Olancho	15. Olancho	9434	672.9	8115	10753
16. Santa Bárbara	16. Santa Bárbara	9000	934.0	7169	10831
17. Valle	17. Valle	18361	3588.6	11328	25395
18. Yoro	18. Yoro	10463	1055.9	8394	12533

Por ejemplo, la diferencia media entre los departamentos Colon y Comayagua $\hat{\bar{y}}_{Colon} - \hat{\bar{y}}_{Comayagua}$

Procedimiento para realizar los contrastes

```
# Paso 1: diferencia de estimaciones (Colon - Comayagua)
10428 - 11674
```

```
[1] -1246
```

```
# Paso 2: error estándar de la diferencia
vcov(prom_dam)[2:3,2:3] %>% tba()
```

	2. Colon	3. Comayagua
2. Colon	965391	0
3. Comayagua	0	878904

```
sqrt(965391 + 878904 - 2*0)
```

```
[1] 1358
```

Procedimiento para realizar los contrastes

El procedimiento anterior se reduce a la sintaxis:

```
svycontrast(prom_dam,
            list(diff = c(0, 1, -1, rep(0,15)))) %>%
  data.frame()
```

	contrast	diff
diff	-1246	1358

Creado una matriz de contrastes

Ahora el interés es realizar los contrastes siguientes:

- $\hat{\bar{y}}_{Atlantida} - \hat{\bar{y}}_{Colon'}$
- $\hat{\bar{y}}_{Cortes} - \hat{\bar{y}}_{Choluteca}$
- $\hat{\bar{y}}_{Olancho} - \hat{\bar{y}}_{Yoro}$

Escrita de forma matricial es:

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

Creado una matriz de contrastes en R

```r
contrastes <- matrix(0, nrow = 3, ncol = 18)
# Definir los contrastes
contrastes[1, c(1, 2)]  <- c(1, -1)   # Atlántida - Colón
contrastes[2, c(5, 6)]  <- c(1, -1)   # Cortés - Choluteca
contrastes[3, c(15, 18)] <- c(1, -1)  # Olancho - Yoro

contrastes
```

1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	-1

Contrastes en R

```
svycontrast(prom_dam, list(
           Atlantida_Colón = contrastes[1,],
          Cortés_Choluteca = contrastes[2,],
         Olancho_Yoro = contrastes[3,]
            )) %>% data.frame()
```

	contrast	SE
Atlantida_Colón	1789.401	1921
Cortés_Choluteca	3.329	3317
Olancho_Yoro	-1029.019	1252

Contrastes no independiente

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

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

	Sexo	YEMPLEO	se	ci_l	ci_u
1. Hombre	1. Hombre	12904	455.1	12011	13796
2. Mujer	2. Mujer	10628	402.8	9839	11418

Contrastes no independiente

El contraste

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

Es calculado como sigue:

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

	contrast	diff_Sexo
diff_Sexo	2275	442.8

Contrastes no independiente

```
vcov(prom_sexo)
```

	1. Hombre	2. Mujer
1. Hombre	207159	86645
2. Mujer	86645	162242

```
# Note que el error estándar de la diff  es igual a
sqrt(207159     + 162242 - 2*86645)
```

```
[1] 442.8
```

Contrastes no independiente

Otra posibilidad es poder obtener resultados agregados, por ejemplo:

$$\hat{\bar{y}}_{Yoro} + \hat{\bar{y}}_{Colon} + \hat{\bar{y}}_{Comayagua}$$

```
sum_dam <- svyby( ~ YEMPLEO,  ~ dam,
                      diseno, svytotal, na.rm = T,
                      covmat = TRUE,
                      vartype = c("se", "ci"))
sum_dam %>% tba()
```

Total del ingreso

	dam	YEMPLEO	se	ci_l	ci_u
1. Atlántida	1. Atlántida	2701263523	420318100	1877455185	3525071860
2. Colon	2. Colon	1554542036	297578897	971298114	2137785957
3. Comayagua	3. Comayagua	3042925359	566577280	1932454295	4153396423
4. Copan	4. Copan	2692039043	1162592848	413398931	4970679155
5. Cortes	5. Cortes	11787256071	933305704	9958010504	13616501638
6. Choluteca	6. Choluteca	3098069130	797281658	1535425794	4660712466
7. El Paraíso	7. El Paraíso	2402659633	568508895	1288402673	3516916592
8. Francisco Morazán	8. Francisco Morazán	11133254507	667147172	9825670078	12440838936
9. Gracias A Dios	9. Gracias A Dios	231975221	26727273	179590729	284359713
10. Intibuca	10. Intibuca	666820269	204777487	265463769	1068176769
11. Islas De La Bahía	11. Islas De La Bahía	693330915	141624220	415752544	970909286
12. La Paz	12. La Paz	793885074	219533927	363606484	1224163665
13. Lempira	13. Lempira	1280637715	515860239	269570227	2291705204
14. Ocotepeque	14. Ocotepeque	581542630	295285596	2793498	1160291762
15. Olancho	15. Olancho	2352293352	332978507	1699667469	3004919234
16. Santa Bárbara	16. Santa Bárbara	1795272465	377859325	1054681797	2535863133
17. Valle	17. Valle	1534073032	677887175	205438582	2862707481
18. Yoro	18. Yoro	2778609442	561651007	1677793696	3879425187

Contrastes no independiente

La matriz de contraste queda como:

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

el procedimiento en R es:

```
svycontrast(sum_dam,
            list(
              Agregado = rep(c(1,0), c(3 , 15))
              )) %>% data.frame() %>% tba()
```

	contrast	Agregado
Agregado	7298730917	704623293

Contrastes

```
vcov(sum_dam)[1:3, 1:3] %>% as.data.frame() %>% tba()
```

	1. Atlántida	2. Colon	3. Comayagua
1. Atlántida	176667305015941248	-44868167557122912	0
2. Colon	-44868167557122912	88553200088489904	0
3. Comayagua	0	0	321009814737339968

```
sqrt(176679802169900928 + 88553200088489904 + 321009814737339968 -
       2*44864888457644600)
```

```
[1] 704636814
```

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

```
diseno <- diseno %>%
  mutate(Edad_cat = cut(
    F2_A6_P4_EDAD ,
    c(0, 15, 30 , 45, 60, Inf),
    labels = c("0 - 15", "16 - 30",
               "31 - 45", "46 - 60", "60 +")
  ))
  prom_edad <-
    svyby(
      ~ YEMPLEO,
      ~ Edad_cat,
      diseno %>% filter(F2_A6_P4_EDAD > 15),
      svymean,
      na.rm = T,
      covmat = TRUE
    )
```

Contrastes no independiente

	Edad_cat	YEMPLEO	se
16 - 30	16 - 30	9451	221.5
31 - 45	31 - 45	14432	750.3
46 - 60	46 - 60	13603	531.6
60 +	60 +	11829	1063.6

Contrastes no independiente

La matriz de contraste estaría dada por:

$$\begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

El procedimiento en R es:

```
svycontrast(prom_edad,
  list(
  agregado_edad = c(1/4, 1/4, 1/4, 1/4)
      )) %>% data.frame()
```

	contrast	agregado_edad
agregado_edad	12329	434.6

Contrastes no independiente

```
vcov(prom_edad)
```

	16 - 30	31 - 45	46 - 60	60 +
16 - 30	49065	39352	30798	95779
31 - 45	39352	562921	71905	129515
46 - 60	30798	71905	282640	130462
60 +	95779	129515	130462	1131332

```
(1 / 4)*sqrt(
  49065 + 2*39352 + 2*30798  + 2*95779  +
          562921 + 2*71905  + 2*129515 +
                   282640  + 2*130462 +
                             1131332
)
```

```
[1] 434.6
```

Contrastes no independiente

```
diseno_temp <- diseno %>% filter(COSTOALQUI > 0)
(razon_sexo <- svyby(~YEMPLEO, ~Sexo,
                     denominator = ~COSTOALQUI,
                     diseno_temp, svyratio,
                     na.rm=T, covmat = TRUE,
                     vartype = c("se", "ci")))
```

	Sexo	YEMPLEO/COSTOALQUI	se.YEMPLEO/COSTOALQUI	ci_l	ci_u
1. Hombre	1. Hombre	62.00	6.265	49.73	74.28
2. Mujer	2. Mujer	40.36	5.762	29.07	51.66

Contrastes no independiente

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

	contrast	diff_sexo
diff_sexo	21.64	10.38

Contrastes no independiente

```
vcov(razon_sexo)
```

	1. Hombre	2. Mujer
1. Hombre	39.25	-17.66
2. Mujer	-17.66	33.20

```
sqrt(39.2 + 33.20 - 2*(-17.66))
```

```
[1] 10.38
```

¡Gracias!

Email: andres.gutierrez@cepal.org