R Package "calidad"

Methodologies and Data Science for Statistical Production

NSO Chile

December 2022

Contents

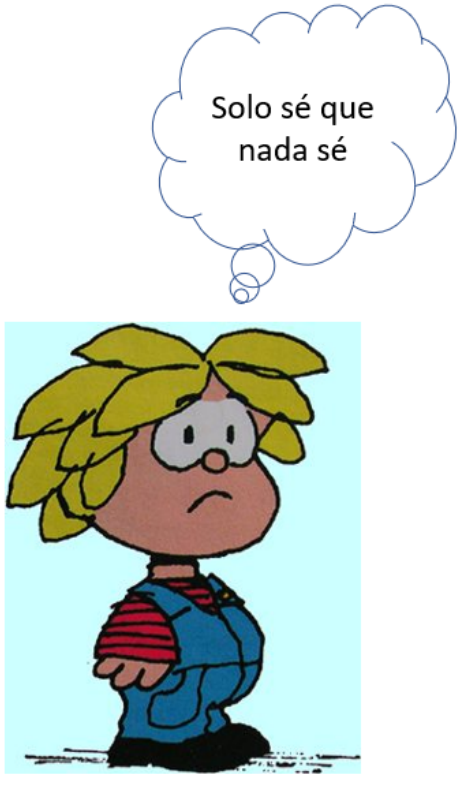
- Estándar de calidad INE
- Estándar de calidad CEPAL
- Paquete de R para implementar estándares
- NSO Chile quality approach
- ECLAC quality approach
- R package implementation

Estándar de calidad en encuestas de hogares INE

Antes de empezar...

Before starting...

What is the meaning of statistical quality?



Estándar INE Chile (NSO Chile approach)

En 2020 el INE publica un estándar de calidad para las estimaciones con **encuestas de hogares**

In 2020 INE Chile published its statistical quality criteria for **household surveys** estimations

Información

Fecha:
Marzo 2020

Fundamentos del Estándar para la evaluación de la calidad de las estimaciones en encuestas de hogares

Documento de trabajo - Metodológico

En este documento se presenta el trabajo metodológico realizado por el INE en relación con la evaluación de las medidas de calidad de las estimaciones provenientes de encuestas a hogares por muestreo, con el propósito de proveer al Sistema Estadístico Nacional (SEN) de un conjunto de directrices que permitan orientar a la población usuaria respecto a la evaluación de la calidad, uso, análisis e interpretación de la información que produce el INE.
Bajo un marco de aseguramiento de la calidad, la importancia de contar con lineamientos para la evaluación de las estimaciones a través de criterios que consideren diferentes dimensiones (como el tamaño muestral, grados de libertad, coeficiente de variación, error estándar, entre otros) radica en que la fiabilidad estadística, entendida como el grado en que las mediciones obtenidas reflejan la realidad, permite definir cuándo una estimación cumple un requisito mínimo de calidad. En este sentido, el documento entrega un panorama respecto al marco conceptual de las medidas de calidad, su uso y criterios utilizados en las encuestas del INE, así como también en diversas Oficinas Nacionales de Estadística (ONE). De acuerdo con los antecedentes expuestos, se desarrolla una propuesta de lineamientos a través de flujogramas que orienta al conjunto de usuarios en la toma de decisiones al momento de analizar y publicar información.

Descargar Documento

Estándar INE Chile (NSO Chile approach)

Cuadro estadístico: arreglo ordenado de datos procesados para facilitar la lectura e interpretación

Statistical chart: data array sorted with reading and interpretation purposes

Ejemplo ilustrativo : **Personas microemprendedoras por sexo según categoría ocupacional**

	N° personas micro emprendedores			Distribución			Concentración		
Categoría ocupacional	**Total**	**Hombres**	**Mujeres**	**Total**	**Hombres**	**Mujeres**	**Total**	**Hombres**	**Mujeres**
Total	**53.317**	30.278	23.040	**100**	56,79	43,21	**100**	**100**	**100**
Empleador	**10.729**	6.736	3.994	**100**	62,78	37,23	20,12	22,25	17,34
Cuenta propia	**42.588**	23.542	19.046	**100**	55,28	44,72	79,88	77,75	82,66

Fuente: INE, Quinta Encuesta de Microemprendimiento, 2017

Estándar INE Chile (NSO Chile approach)

- **Tamaño muestral (tm)**: unidades de análisis que nutren las estimaciones (viviendas, hogares y/o personas)
- **Grados de libertad (gl)**
  - Tratamiento diferenciado para las **proporciones y razones definidas entre 0 y 1**

$df_d(1)=$	#UPM con observaciones en la subpoblación menos #Estratos con observaciones en la subpoblación

- **Coeficiente de variación / error estándar**
  - Proporción y razones definidas entre 0 y 1: **SE**
  - Resto: **CV**

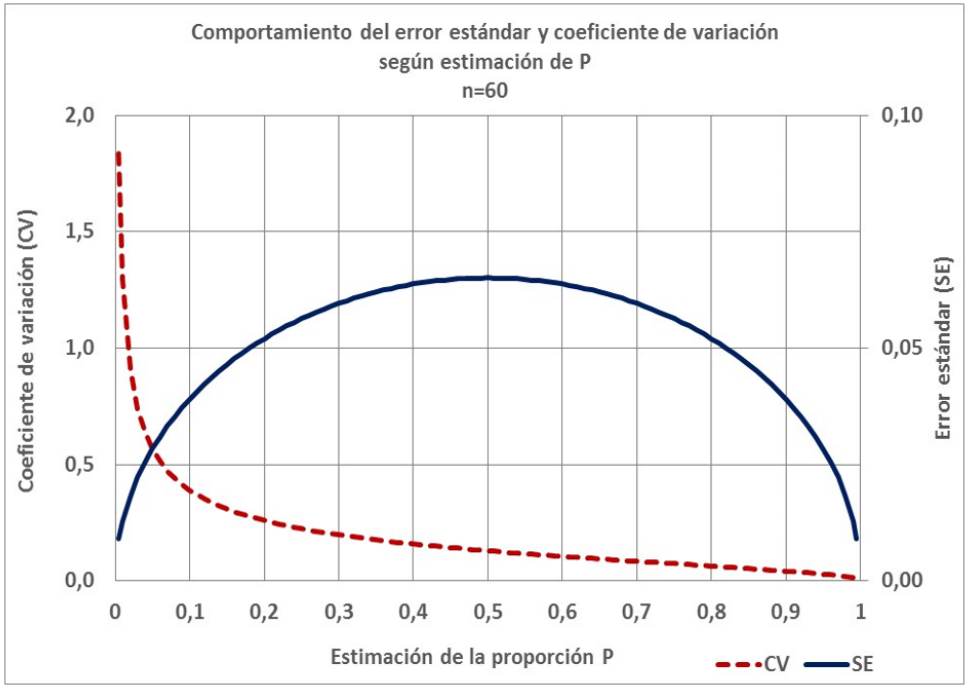
- **Sample size**: analysis units contributing to estimations (houses, households or people)
- **Degrees of freedom**
  - Different procedure for **Proportions and ratios between 0 and 1**

$df_d(1)=$	#UPM con observaciones en la subpoblación menos #Estratos con observaciones en la subpoblación

- **Coefficient of variation / standard error**
  - Proportions and ratios between 0 and 1: **SE**
  - Other: **CV**

Dicotomía de las proporciones

Dicothomy of proportions



P	Error estándar	CV
0,1	0,039	0,391
0,9	0,039	0,043

Es ilógico concluir que la estimación de p no tiene una calidad aceptable pero la de 1-p sí la tiene.

It makes no sense to conclude that p has acceptable quality and 1-p does not.

Estándar INE Chile (NSO Chile approach)

Indicadores de calidad

- Proporciones y razones definidas entre 0 y 1: **Error estándar**
- Otras: Coeficiente de variación

Quality criteria

- Proportions and ratios between 0 and 1: **Standard error**
- Else: Coefficient of variation

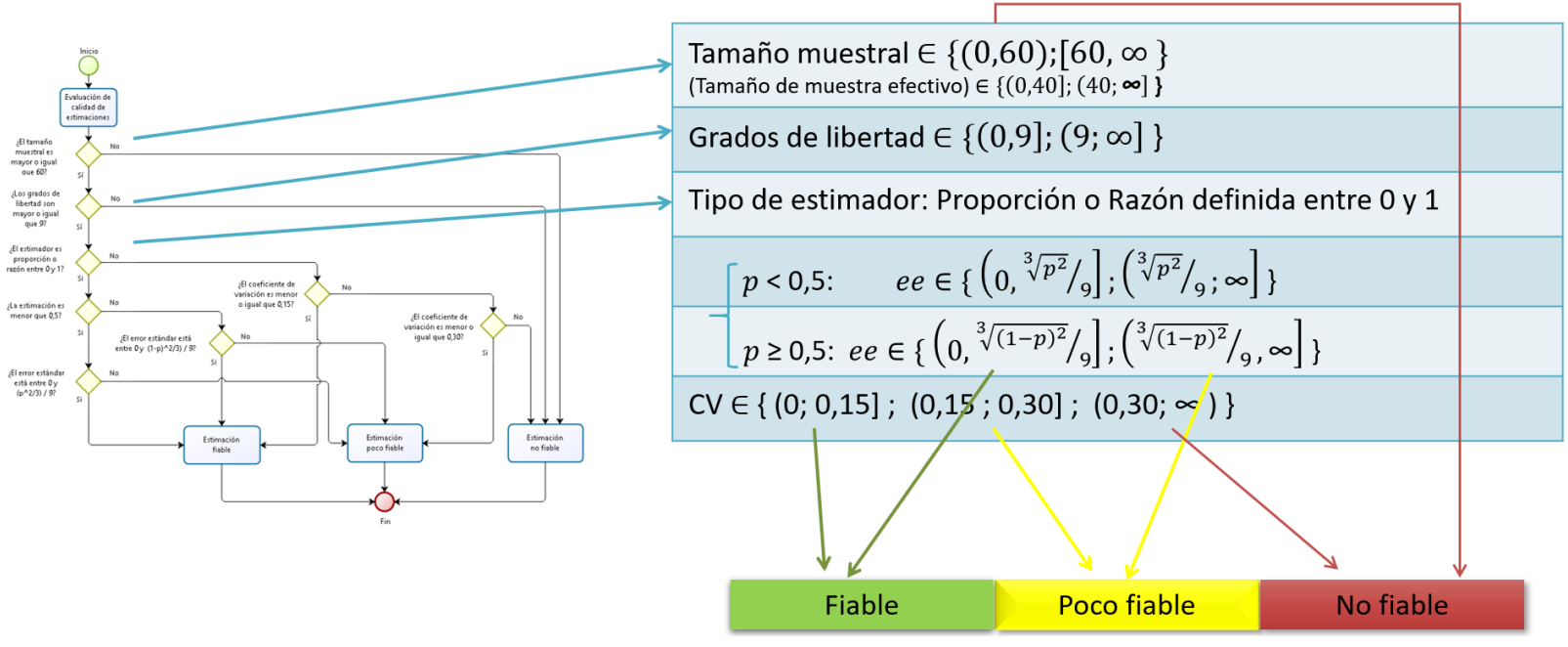
Estándar INE Chile (NSO Chile approach)

Primera etapa de aplicación de estándar

Flujograma para evaluación de calidad de las estimaciones

First step of quality assesing

Quality assesing flow



Categories: Reliable, weakly reliable and non-reliable

Introducción paquete calidad (introduction)

¿Cómo llevar a la práctica los conceptos de calidad?

- Existen múltiples herramientas (Stata, R, SAS, Python) y todas son válidas
- Una posibilidad es el uso de un **paquete (librería)**
- El paquete `calidad` implementa el estándar mediante `R`

¿How to put the quality concepts into practice?

- There are many tools (Stata, R, SAS, Python) and all of them are valid
- One possibility is the use of a **package (library)**
- The `calidad` package implements the criteria using `R`

Introducción paquete calidad (introduction)

Objetivos del paquete

- Facilitar la aplicación del estándar a usuarios externos
- Aumentar la eficiencia de los analistas de datos
- Reducir la probabilidad de errores en la implementación

Objectives of the package

- Make it easier for external users to apply the standard
- Increase the data analysts efficiency
- Reduce the probability of implementations errors

Introducción paquete calidad (introduction)

El paquete `calidad` combina conceptos de calidad con el paquete `survey`, desarrollado por Thomas Lumley

`calidad` combines quality concepts with the `survey` package, developed by Thomas Lumley

Desde hace un año el paquete está en github y hace algunas semanas lo subimos a CRAN

- media
- proporciones
- ~~mediana (con réplicas)~~
- tamaños
- totales

The package has been available on github for a year and 2 months ago we uploaded it to CRAN

- mean
- proportions
- ~~median (replicates)~~
- sizes
- totals

Implementación en R

Implementation on R

Demostración (Demonstration)

Lo primero, es descargar el paquete desde CRAN

The first step is to download the package from CRAN

```
install.packages("calidad")
```

... o versión en desarrollo desde github

... or developing version from github

```
devtools::install_github("inesscc/calidad")
```

Cargamos el paquete en la sesión y otras dependencias que usaremos

We load the package with other dependencies we will be using

```
library(calidad)
library(survey)
library(dplyr)
```

Demostración paquete calidad

Trabajaremos con los datos de la Encuesta de Caracterización Socioeconómica (CASEN) 2020 (cargados en el paquete)

Construyamos algunas variables necesarias para calcular indicadores relevantes del mercado laboral

```
casen_edit <- casen %>%
  mutate(fdt = if_else(activ %in% c(1, 2), 1, 0, missing = 0), # fuerza de trabajo
         ocupado = if_else(activ == 1, 1, 0, missing = 0), # persona ocupada
         desocupado = if_else(activ == 2, 1, 0, missing = 0), # persona desocupada
         metro = if_else(region == 13, 1, 0))
```

Variables *dummy*:

- fuerza de trabajo
- ocupado
- desocupado
- metro (pertenece a la región metropolitana)

Panorama general

Declaramos el diseño complejo con la función `svydesign` de `survey`

```
dc <- svydesign(weights = ~expr, ids = ~cod_upm, strata = ~estrato, data = casen_edit )
options(survey.lonely.psu = "certainty")
```

Debemos decirle a `R` qué hacer con la varianza cuando encuentra estratos con una sola UPM

Panorama general

El paquete `calidad` tiene 2 tipos de funciones:

- *create_*: **crean** los insumos para el estándar
- *evaluate*: **evaluación** del estándar

Podemos hacer los siguientes cálculos

- media (create_mean)
- proporción o razón (create_prop)
- suma de variables continuas (create_total)
- conteo de unidades (create_size)

Creando los insumos: create_mean

Queremos calcular la edad media para mujeres y hombres

```
create_mean(var = "edad", domains = "sexo", design = dc)
```

```
##   sexo     stat        se    df     n          cv
## 1    1 35.81776 0.1320879 10701 86096 0.003687776
## 2    2 38.88116 0.2030783 10818 99341 0.005223053
```

- `var`: variable a estimar
- `domains`: desagregaciones
- `design`: diseño muestral creado con `svydesign`

La función genera:

- estimación (stat)
- error estándar (se)
- coeficiente de variación (CV)
- grados de libertad (df)
- tamaño muestral (n)

Creando los insumos: create_prop

¿Y si queremos calcular la tasa de desempleo?

Para ello, contamos con la función `create_prop`

```
create_prop(var = "desocupado", domains = "sexo", design = dc)
```

El problema es que el desempleo debe calcularse sobre una subpoblación específica (fuerza de trabajo)

Para ello, utilizamos el argumento `subpop`

```
create_prop(var = "desocupado", domains = "sexo", subpop = "fdt", design = dc)
```

Es muy importante considerar que la variable **subpop debe ser dummy**

Con subpop evitamos error en el cálculo de la varianza

Creando los insumos: create_prop

¿Qué pasa si queremos desagregar por más variables?

Se debe agregar otra variable utilizando un signo +

```
create_prop(var = "desocupado", domains = "sexo+metro", subpop = "fdt", design = dc)
```

```
##   sexo metro      stat          se   df     n         cv
## 1    1     0 0.1171641 0.002540917 7977 34097 0.02168682
## 2    2     0 0.1417696 0.003012865 7750 27786 0.02125184
## 3    1     1 0.1097366 0.004349727 2019  9901 0.03963790
## 4    2     1 0.1364608 0.008565356 1972  9055 0.06276788
```

Creando los insumos: create_prop

Queremos calcular el número de ocupad**os** respecto al número de ocupad**as**

$$\frac{SumaOcupadosHombre}{SumaOcupadasMujer}$$

Lo primero que debemos hacer es crear variables auxiliares

```
casen_edit <- casen_edit %>%
  mutate(ocupado_hombre = if_else(sexo == 1, ocupado, 0),
         ocupada_mujer  = if_else(sexo == 2, ocupado, 0))
```

Volvemos a declarar el diseño para incluir las variables recién creadas

```
dc <- svydesign(weights = ~expr, ids = ~cod_upm, strata = ~estrato, data = casen_edit )
```

Creando los insumos: create_prop

La función `create_prop` permite incluir el argumento `denominator`

```
create_prop(var = "ocupado_hombre", denominator = "ocupada_mujer",
            subpop = "fdt", design = dc)
```

```
##       stat           se    df     n         cv
## 1 1.186844 0.0001418034 10590 80839 0.01003344
```

Podemos incluir el parámetro `domains`, si queremos desagregar

```
create_prop(var = "ocupado_hombre", denominator = "ocupada_mujer",
            domains = "metro" , subpop = "fdt", design = dc)
```

```
##   metro     stat          se   df     n         cv
## 1     0 1.238095 0.01312396 8510 61883 0.01060013
## 2     1 1.127986 0.02073977 2080 18956 0.01838654
```

Argumentos adicionales

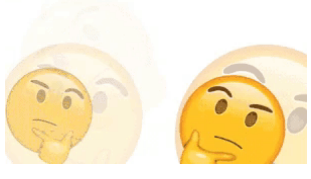
Solo hemos revisado `create_prop` y `create_mean`

Todas las funciones del paquete operan de manera similar

Existen más argumentos

- ci
- deff
- rel_error
- ...

Hasta el momento solo hemos visto la generacion de insumos



Evaluación del estándar

Evaluemos si la media de edad por sexo cumple con el estándar

```
est <- create_mean(var = "edad", domains = "sexo", design =  dc)
evaluate(est)
```

```
##   sexo     stat        se    df     n          cv                 eval_n
## 1    1 35.81776 0.1320879 10701 86096 0.003687776 sufficient sample size
## 2    2 38.88116 0.2030783 10818 99341 0.005223053 sufficient sample size
##         eval_df    eval_cv    label
## 1 sufficient df cv <= 0.15 reliable
## 2 sufficient df cv <= 0.15 reliable
```

Tenemos 4 columnas nuevas

- `eval_n`: indica si el tamaño muestral es sufciente
- `eval_df`: indica si los gl son suficientes
- `eval_cv`: indica el tramo en el que está el cv
- `label`: evaluación final de la estimación

Por defecto, las funciones de evaluación consideran el estándar INE

- **Grados de libertad**: 9
- **Tamaño de muestra**: 60
- **Tramos de CV**: 0.15, 0.3

Evaluación del estándar

Veamos el caso de la tasa de desempleo

```
est <-  create_prop(var = "desocupado", subpop = "fdt", domains = "sexo", design = dc)
evaluate(est)
```

```
##   sexo      stat          se   df     n         cv                 eval_n
## 1    1 0.1138937 0.002402195 9996 43998 0.02109156 sufficient sample size
## 2    2 0.1393067 0.004271094 9722 36841 0.03065964 sufficient sample size
##         eval_df prop_est eval_type  quadratic       eval_se eval_cv    label
## 1 sufficient df   <= 0.5   Eval SE 0.02610701 admissible SE    <NA> reliable
## 2 sufficient df   <= 0.5   Eval SE 0.02985879 admissible SE    <NA> reliable
```

Además de las columnas ya vistas, tenemos

- `prop_est`
- `eval_type`
- `quadratic`
- `eval_se`
- `eval_cv`

Evaluación del estándar

El estándar establece que un tabulado puede ser publicado si el 50% de sus celdas es fiable

Para saber si el tabulado debe ser publicado, usamos el argumento `publish`

```
est <-  create_size(var = "desocupado", subpop = "fdt", domains = "region+sexo", design = dc)
evaluate(est, publish = T) %>%
  select(region, sexo, stat, label, publication, pass) %>%
  slice(1:6)
```

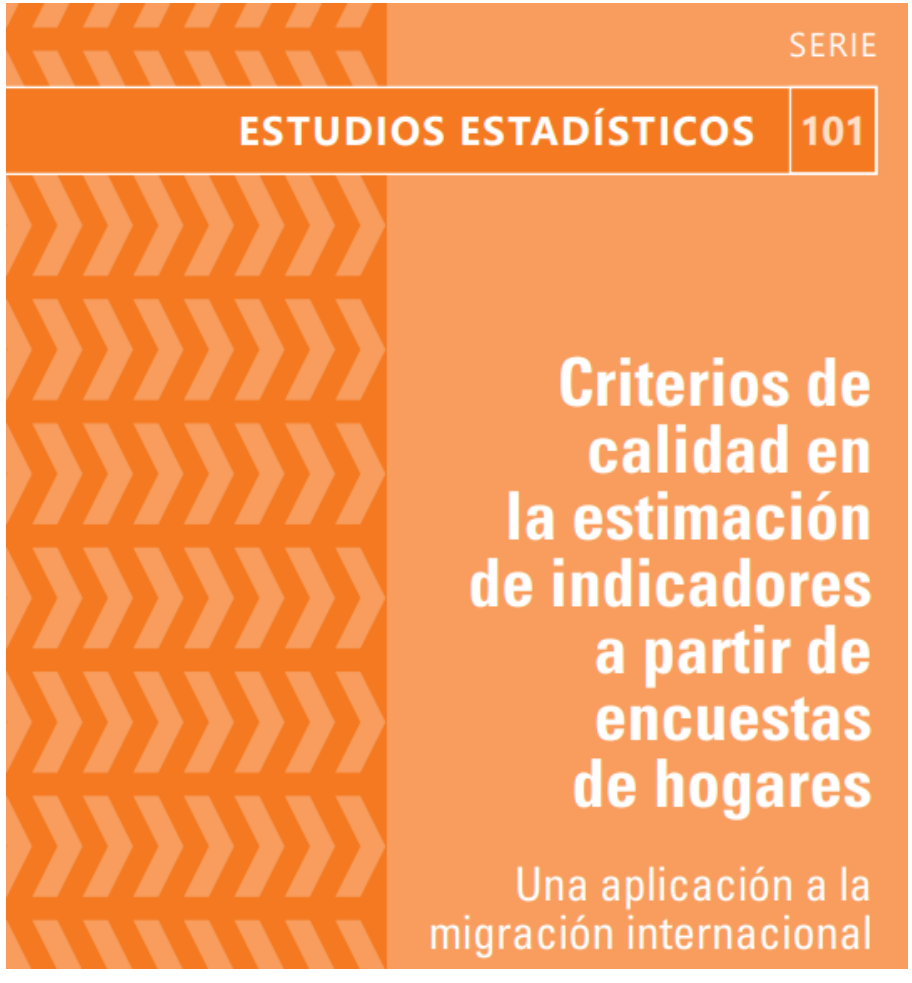
```
##   region sexo  stat    label publication                    pass
## 1      1    1  9436 reliable     publish 100% reliable estimates
## 2      2    1 21139 reliable     publish 100% reliable estimates
## 3      3    1  8586 reliable     publish 100% reliable estimates
## 4      4    1 22801 reliable     publish 100% reliable estimates
## 5      5    1 56607 reliable     publish 100% reliable estimates
## 6      6    1 24507 reliable     publish 100% reliable estimates
```

Tenemos 2 nuevas columnas

- `publication`: evaluación general del tabulado
- `pass`: porcentaje de celdas con categoría fiable

Estándar de calidad CEPAL

Estándar de calidad CEPAL



Estándar de calidad CEPAL

- **El estándar CEPAL considera:**
  - coeficiente de variación
  - coeficiente de variación logarítmico
  - tamaño de muestra
  - tamaño de muestra efectivo
  - conteo de casos no ponderado
  - grados de libertad

Nuevas funcionalidades

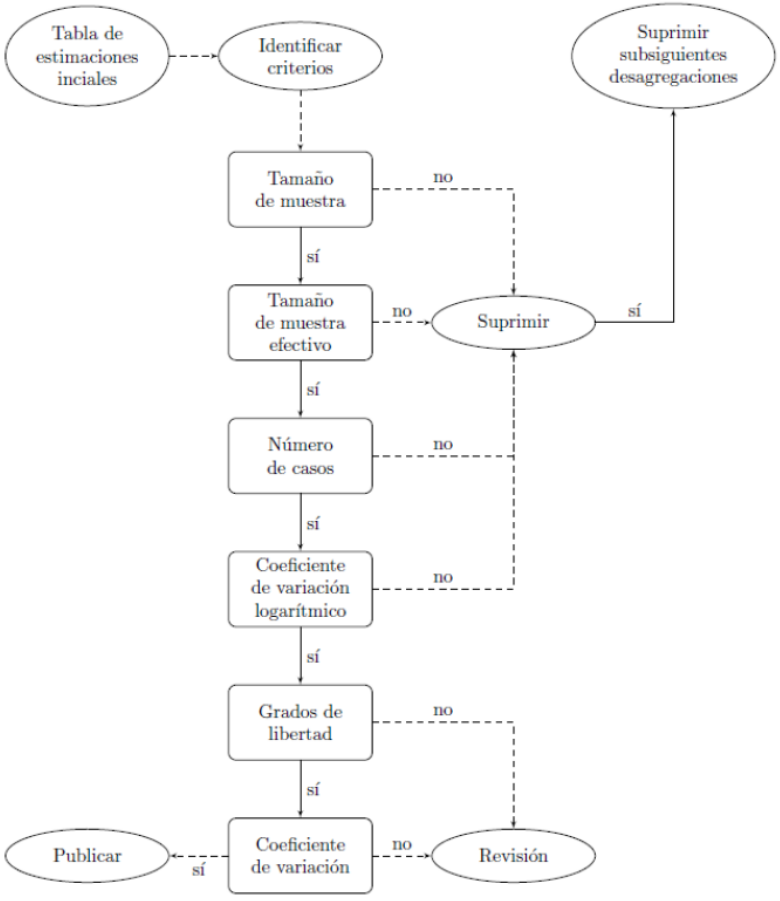
Nuevos desarrollos con CEPAL

Nuevos indicadores de calidad

Flexibilización de umbrales

Alinear nombres con la teoría

Estándar de calidad CEPAL



Suprimir

Revisar

Publicar

¡Veamos un poco de código!

Implementación estándar CEPAL

Se deben incluir nuevos parámetros en las funciones `create_`

```
est <- create_size(var = "desocupado", domains = "region+sexo", design = dc,
                   unweighted = T, deff = T, ess = T, df_type = "eclac")
```

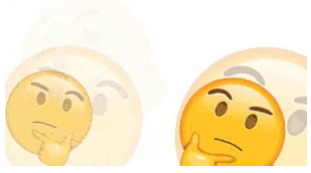
Y agregar cepal en evaluate

```
evaluate(est, scheme = "eclac") %>%
  select(region, sexo, stat, n, df, cv, unweighted, ess, label) %>%
  slice(1:6)
```

```
##   region sexo  stat    n   df         cv unweighted      ess   label
## 1      1    1  9436 3981  419 0.08412127        220 2545.121 publish
## 2      2    1 21139 3572  493 0.08197950        243 2130.673 publish
## 3      3    1  8586 3468  557 0.08745271        205 2076.391 publish
## 4      4    1 22801 3783  495 0.08145155        238 2402.469 publish
## 5      5    1 56607 8397 1119 0.06976443        511 2987.021 publish
## 6      6    1 24507 5830  611 0.07226518        327 3470.302 publish
```

Estándar CEPAL

¿Y si necesito manejar los parámetros del estándar?



```
est <- create_size(var = "desocupado", domains = "region+sexo", design = dc,
                   unweighted = T, deff = T, ess = T, df_type = "ine")

evaluate(est, scheme = "eclac", unweighted = 220, ess = 200) %>%
  select(region, sexo, stat, n, df, cv, unweighted, ess, label) %>%
  slice(1:6)
```

```
##   region sexo  stat   n  df         cv unweighted      ess   label
## 1      1    1  9436 220 144 0.08412127        220 140.6497 supress
## 2      2    1 21139 243 167 0.08197950        243 144.9478 supress
## 3      3    1  8586 205 138 0.08745271        205 122.7394 supress
## 4      4    1 22801 238 146 0.08145155        238 151.1466 supress
## 5      5    1 56607 511 322 0.06976443        511 181.7754 supress
## 6      6    1 24507 327 175 0.07226518        327 194.6464 supress
```

Utilización de loops

Queremos calcular la media para varias variables

En este caso, queremos la media de `edad` y `ing_aut_hog`, según sexo

```
insumos <- data.frame()
for (v in c("edad", "ing_aut_hog")) {
  insumo <- create_mean(var = v, domains = "sexo", design = dc, rm.na = T )
  insumos <- bind_rows(insumos, insumo)
}
```

Podemos hacer lo mismo, utilizando el paquete `purrr` (mucho más recomendado que un for)

```
insumos <- map_df(c("edad", "ing_aut_hog"), ~create_mean(var = .x, domains = "sexo",
                                                         design = dc, rm.na = T ))
```

Combinación de estándares

La idea es generar una herramienta que:

- Implemente los dos estándares
- Ofrezca flexibilidad a los usuarios

¿En qué estamos?

- Mantenimiento constante
- Junto a CEPAL estamos preparando instancias de difusión:
  - RTC (aquí estamos)
  - Intersecretariat Working Group on Household Surveys
- Preparación de material de difusión
- Integración de `calidad` con `dataine`
- Comienzo de nuevos desarrollos

Desarrollo open source

El paquete `calidad` es un desarrollo completamente *open source*

En este repositorio de github pueden proponer nuevos desarrollos

Klaus Lehmann y Ricardo Pizarro son los mantenedores

Pueden generar *issues* o nuevas ramas de desarrollo

Si tienen **propuestas de mejora, reportes de errores o nuevos desarrollos**, estaremos felices de revisarlo e incorporarlo al paquete

😄😄😄😄😄

- Klaus Lehmann: kilehmannm@ine.gob.cl
- Ignacio Agloni: ifaglonij@ine.gob.cl
- Ricardo Pizarro: rapizarros@ine.gob.cl

https://github.com/inesscc/calidad

Presentación librería calidad

Proyecto Estratégico Metodología y Ciencia de Datos para la Producción Estadística

Subdirección Técnica

Octubre 2022