Análisis de encuestas de hogares con R

Módulo 2: Análisis de variables categóricas

CEPAL - Unidad de Estadísticas Sociales

Tabla de contenidos I

Introducción

Definición del diseño y creación de variables categóricas

Introducción

Motivación

- En el mundo de la estadística y el análisis de datos, nos encontramos con una variedad de variables que pueden ser clasificadas en dos categorías principales: cualitativas y cuantitativas.
- Las variables cualitativas, también conocidas como categóricas, representan características o cualidades que no se pueden medir con números, como el género, el estado civil o el tipo de vivienda.
- Algunas variables cuantitativas se transforman en categóricas al dividir su rango en categorías, y viceversa, algunas variables categóricas se convierten en cuantitativas mediante análisis especializados.
- En esta presentación, exploraremos esta distinción y cómo abordar variables cualitativas en el contexto de encuestas y análisis de datos.

Definición del diseño y creación de variables categóricas

Lectura de la base

Iniciemos con la lectura de la encuesta.

```
encuesta <- readRDS("Imagenes/02_variable_continua/ENIGH_HND_Pers.rds")
```

El paso siguiente es realizar declaración del objeto tipo diseño.

```
options(survey.lonely.psu = "adjust")
library(srvyr)
diseno <- encuesta %>% # Base de datos.
  mutate(estrato = haven::as_factor(F1_A0_ESTRATO),
         Sexo = haven::as_factor(F2_A6_P3_SEXO),
         Area = haven::as_factor(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
  )
```

Creación de nuevas variables

Durante los análisis de encuesta surge la necesidad de crear nuevas variables a partir de las existentes, aquí mostramos la definición de algunas de ellas.

```
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 +")
  )
)
```

Se ha introducido la función `case_when` la cual es una extensión del a función `ifelse` que permite crear múltiples categorías a partir de una o varias condiciones.

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

Estimación del tamaño.

El primer parámetro estimado serán los tamaños de la población y subpoblaciones.

```
(tamano_zona <- diseno %>% group_by(Area) %>%
   summarise(
     n = unweighted(n()), # Observaciones en la muestra.
     Nd = survey_total(vartype = c("se","ci"))))
```

Area	n	Nd	Nd_se	Nd_low	Nd_upp
1. Urbana	26923	5445857	78340	5292056	5599658
2. Rural	5106	4340256	162236	4021748	4658764

En la tabla *n* denota el número de observaciones en la muestra por Zona y *Nd* denota la estimación del total de observaciones en la población.

Estimación de tamaño

Empleando una sintaxis similar es posible estimar el número de personas en condición en el un decil dado el ingreso disponible percápita del hogar

```
(
  tamano_decil <- diseno %>%
      mutate(DECIL = haven::as_factor(DECIL_YDISPO_PER) ) %>%
    group_by(DECIL) %>%
    summarise(Nd = survey_total(vartype = c("se", "ci")))
)
```

DECIL	Nd	Nd_se	Nd_low	Nd_upp
1	1209670	105509	1002530	1416810
2	1146871	76995	995711	1298030
3	1121981	67205	990041	1253921
4	1110289	63055	986496	1234081
5	1014202	60200	896014	1132390
6	966416	50825	866634	1066197
7	914053	51030	813868	1014239
8	851633	52534	748496	954771

Estimación de tamaño

En forma similar es posible estimar el número de personas con tenencia de celular

```
(tamano_etnia<- diseno %>%
     mutate(etnia = haven::as_factor(F2_A6_P5_ETNIA) ) %>%
   group_by(etnia) %>%
   summarise(
       Nd = survey_total(vartype = c("se","ci"))))
```

etnia	Nd	Nd_se	Nd_low	Nd_upp
1. Indigena	663165	123041	421605.7	904724
2. Afrohondureño(a)	29540	5907	17943.6	41136
3. Negro(a)	34682	6142	22623.8	46740
4. Mestizo(a)	8324693	199547	7932934.3	8716451
5. Blanco(a)	731372	50477	632272.6	830472
6. Otro (especifique)	2661	1470	-224.5	5547

Estimación de tamaño

Otra variable de interés es conocer el estado de ocupación de la personas.

```
(
  tamano_ocupacion <- diseno %>%
    mutate(ocupacion = haven::as_factor(F2_A9_P3_TIPOEMPLEADO)) %>%
    group_by(ocupacion) %>%
    summarise(Nd = survey_total(vartype = c("se", "ci")))
)
```

Estimación de tamaño

```
(
  tamano_ocupacion <- diseno %>%
    mutate(ocupacion = haven::as_factor(F2_A9_P3_TIPOEMPLEADO)) %>%
    group_by(ocupacion) %>%
    summarise(Nd = survey_total(vartype = c("se", "ci")))
)
```

Estimación de tamaño

Utilizando la función `group_by` es posible obtener resultados por más de un nivel de agregación.

```
(tamano_etnia_sexo <- diseno %>%
     mutate(etnia = haven::as_factor(F2_A6_P5_ETNIA)) %>%
   group_by(etnia, Sexo) %>%
   cascade(
       Nd = survey_total(vartype = c("se","ci")),
       .fill = "Total") %>%
   data.frame()
   )
```

Estimación de tamaño

etnia	Sexo	Nd	Nd_se	Nd_low	Nd_upp
1. Indigena	1. Hombre	330443	60571.8	211525.7	449360
1. Indigena	2. Mujer	332722	63593.6	207872.4	457572
1. Indigena	Total	663165	123041.0	421605.7	904724
2. Afrohondureño(a)	1. Hombre	13536	3001.7	7643.3	19429
2. Afrohondureño(a)	2. Mujer	16004	3369.5	9388.6	22619
2. Afrohondureño(a)	Total	29540	5906.8	17943.6	41136
3. Negro(a)	1. Hombre	17841	3454.6	11059.2	24624
3. Negro(a)	2. Mujer	16841	3380.2	10204.4	23477
3. Negro(a)	Total	34682	6142.0	22623.8	46740
4. Mestizo(a)	1. Hombre	3941279	106130.9	3732918.6	4149640
4. Mestizo(a)	2. Mujer	4383414	102802.7	4181587.2	4585240
4. Mestizo(a)	Total	8324693	199546.8	7932934.3	8716451
5. Blanco(a)	1. Hombre	330192	23598.8	283861.6	376522
5. Blanco(a)	2. Mujer	401180	30228.8	341834.0	460527
5. Blanco(a)	Total	731372	50477.5	632272.6	830472
6. Otro (especifique)	1. Hombre	1326	788.0	-221.5	2873
6. Otro (especifique)	2. Mujer	1336	732.1	-101.7	2773
6. Otro (especifique)	Total	2661	1469.9	-224.5	5547
Total	Total	9786113	180160.3	9432414.8	10139811

Estimación de Proporciones Poblacionales

En encuestas de hogares, a menudo es importante estimar la proporción de una característica particular en una población, como la proporción de personas que tienen un cierto nivel de educación, la proporción de hogares con acceso a servicios básicos, entre otros.

La estimación de una proporción poblacional se puede hacer utilizando la siguiente ecuación:

$$\hat{\pi} = p = \frac{\sum_{i=1}^{n} \omega_i y_i}{\sum_{i=1}^{n} \omega_i}$$

Donde:

- $\hat{\pi}$ es la estimación de la proporción poblacional.
- n es el tamaño de la muestra.
- ω_i son los pesos de muestreo para cada unidad de la muestra.
- y_i es la variable binaria que indica si la unidad de muestreo tiene la característica de interés (1 si la tiene, 0 si no la tiene).

Estimación de proporción de urbano y rural

El procedimiento estándar para el calculo de proporciones es crear una *variable dummy* y sobre está realizar las operaciones. Sin embargo, la librería `srvy` nos simplifica el calculo, mediante la sintaxis.

```
(prop_Area <- diseno %>% group_by(Area) %>%
   summarise(
     prop = survey_mean(vartype = c("se","ci"),
                        proportion = TRUE )))
```

Area	prop	prop_se	prop_low	prop_upp
1. Urbana	0.5565	0.0099	0.5370	0.5758
2. Rural	0.4435	0.0099	0.4242	0.4630

Note que, se utilizo la función `survey_mean` para la estimación.

Estimación de proporción de urbano y rural

La función idónea para realizar la estimación de las proporciones es `survey_prop` y la sintaxis es como sigue:

```
(prop_area2 <- diseno %>% group_by(Area) %>%
   summarise(
     prop = survey_prop(vartype = c("se","ci") )))
```

Area	prop	prop_se	prop_low	prop_upp
1. Urbana	0.5565	0.0099	0.5370	0.5758
2. Rural	0.4435	0.0099	0.4242	0.4630

Proporción de hombres y mujeres en la área urbana

Si el interés es obtener la estimación para una subpoblación, procedemos así:

```
(prop_sexoU <- sub_Urbano %>% group_by(Sexo) %>%
   summarise(
       prop = survey_prop(vartype = c("se","ci"))))
```

Sexo	prop	prop_se	prop_low	prop_upp
1. Hombre	0.4616	0.0031	0.4555	0.4678
2. Mujer	0.5384	0.0031	0.5322	0.5445

¿Cómo estimar el Proporción de hombres dado que están en zona rural?

Proporción de hombres y mujeres en la zona rural

```
(prop_sexoR <- sub_Rural %>% group_by(Sexo) %>%
   summarise(
     n = unweighted(n()),
     prop = survey_prop(vartype = c("se","ci"))))
```

Sexo	n	prop	prop_se	prop_low	prop_upp
1. Hombre	2490	0.4886	0.006	0.4766	0.5006
2. Mujer	2616	0.5114	0.006	0.4994	0.5234

¿Cómo estimar el Proporción de hombres en la área rural dado que es hombre?

Proporción de hombres en la área urbana y rural

```
(prop_ZonaH <- sub_Hombre %>% group_by(Area) %>%
   summarise(
     prop = survey_prop(vartype = c("se","ci"))))
```

Area	prop	prop_se	prop_low	prop_upp
1. Urbana	0.5424	0.011	0.5208	0.5639
2. Rural	0.4576	0.011	0.4361	0.4792

¿Cómo estimar el Proporción de mujeres en la área rural dado que es mujer?

Proporción de mujeres en la área urbana y rural

```
(prop_ZonaM <- sub_Mujer %>% group_by(Area) %>%
   summarise(
    prop = survey_prop(vartype = c("se","ci"))))
```

Area	prop	prop_se	prop_low	prop_upp
1. Urbana	0.5691	0.0099	0.5496	0.5884
2. Rural	0.4309	0.0099	0.4116	0.4504

Proporción de hombres en la área urbana y rural

Con el uso de la función `group_by` es posible estimar un mayor numero de niveles de agregación al combinar dos o más variables.

```
(prop_ZonaH_edad <- sub_Hombre %>%
   group_by(Area, Edad_cat ) %>%
   summarise(
     prop = survey_prop(vartype = c("se","ci")))%>%
   data.frame())
```

Proporción de hombres en la zona urbana y rural

Area	Edad_cat	prop	prop_se	prop_low	prop_upp
1. Urbana	0 - 15	0.3128	0.0052	0.3027	0.3231
1. Urbana	16 - 30	0.2573	0.0049	0.2478	0.2671
1. Urbana	31 - 45	0.1823	0.0042	0.1743	0.1907
1. Urbana	46 - 60	0.1254	0.0037	0.1184	0.1329
1. Urbana	60 +	0.1047	0.0034	0.0982	0.1115
1. Urbana	NA	0.0174	0.0013	0.0151	0.0202
2. Rural	0 - 15	0.3386	0.0114	0.3167	0.3612
2. Rural	16 - 30	0.2484	0.0115	0.2265	0.2716
2. Rural	31 - 45	0.1762	0.0076	0.1617	0.1917
2. Rural	46 - 60	0.1070	0.0068	0.0944	0.1210
2. Rural	60 +	0.1092	0.0086	0.0934	0.1274
2. Rural	NA	0.0206	0.0037	0.0145	0.0293

Proporción de mujeres en la zona urbana y rural

```
(prop_ZonaM_Pobreza <- sub_Mujer %>%
   group_by(Area, Edad_cat) %>%
   summarise(
     prop = survey_prop(vartype = c("se","ci"))) %>%
   data.frame())
```

Area	Edad_cat	prop	prop_se	prop_low	prop_upp
1. Urbana	0 - 15	0.2683	0.0044	0.2599	0.2770
1. Urbana	16 - 30	0.2489	0.0042	0.2407	0.2573
1. Urbana	31 - 45	0.2057	0.0034	0.1992	0.2124
1. Urbana	46 - 60	0.1404	0.0034	0.1338	0.1473
1. Urbana	60 +	0.1212	0.0037	0.1141	0.1287
1. Urbana	NA	0.0155	0.0013	0.0131	0.0182
2. Rural	0 - 15	0.3132	0.0108	0.2924	0.3348
2. Rural	16 - 30	0.2539	0.0095	0.2358	0.2729
2. Rural	31 - 45	0.1888	0.0090	0.1718	0.2070
2. Rural	46 - 60	0.1214	0.0079	0.1068	0.1378
2. Rural	60 +	0.1081	0.0075	0.0942	0.1238
2. Rural	NA	0.0146	0.0028	0.0100	0.0214

Proporción de hombres en la area disponible para trabajar

```
#F2_A8_P13_DISPONIBLETRABAJAR:  Estaba disponible para trabajar

(prop_ZonaH_disponible <- sub_Hombre %>%
    mutate(disponible = haven::as_factor(F2_A8_P13_DISPONIBLETRABAJAR)) %>%
   group_by(Area, disponible) %>%
   summarise(
     prop = survey_prop(vartype = c("se","ci"))) %>%
   data.frame())
```

Proporción de hombres en la área disponible para trabajar

Area	disponible	prop	prop_se	prop_low	prop_upp
1. Urbana	1. Sí	0.0502	0.0027	0.0452	0.0558
1. Urbana	2. No, pero lo estará en 15 días o menos	0.0001	0.0001	0.0000	0.0007
1. Urbana	3. No, pero lo estará en más de 15 días pero menos de 12 meses	0.0004	0.0002	0.0002	0.0013
1. Urbana	4. No	0.2024	0.0045	0.1937	0.2113
1. Urbana	5. No sabe	0.0046	0.0008	0.0032	0.0064
1. Urbana	NA	0.7423	0.0050	0.7324	0.7520
2. Rural	1. Sí	0.0361	0.0050	0.0274	0.0473
2. Rural	3. No, pero lo estará en más de 15 días pero menos de 12 meses	0.0014	0.0014	0.0002	0.0096
2. Rural	4. No	0.1647	0.0099	0.1461	0.1851
2. Rural	5. No sabe	0.0038	0.0016	0.0017	0.0085
2. Rural	NA	0.7941	0.0107	0.7723	0.8143

Proporción de mujeres en la área urbana y rural

```
(prop_ZonaM_disponible <- sub_Mujer %>%
    mutate(disponible = haven::as_factor(F2_A8_P13_DISPONIBLETRABAJAR)) %>%
   group_by(Area, disponible) %>%
   summarise( prop = survey_prop(vartype = c("se","ci"))) %>%
   data.frame())
```

Proporción de mujeres en la área urbana y rural

Area	disponible	prop	prop_se	prop_low	prop_upp
1. Urbana	1. Sí	0.0778	0.0032	0.0718	0.0844
1. Urbana	2. No, pero lo estará en 15 días o menos	0.0004	0.0002	0.0002	0.0010
1. Urbana	3. No, pero lo estará en más de 15 días pero menos de 12 meses	0.0010	0.0003	0.0005	0.0017
1. Urbana	4. No	0.3574	0.0051	0.3475	0.3675
1. Urbana	5. No sabe	0.0084	0.0009	0.0067	0.0105
1. Urbana	NA	0.5549	0.0054	0.5444	0.5654
2. Rural	1. Sí	0.0763	0.0089	0.0605	0.0958
2. Rural	3. No, pero lo estará en más de 15 días pero menos de 12 meses	0.0005	0.0005	0.0001	0.0036
2. Rural	4. No	0.4062	0.0191	0.3694	0.4441

¡Gracias!

Email: andres.gutierrez@cepal.org