Manejo de Datos con R

Sesión 2: Manipulacion de datos

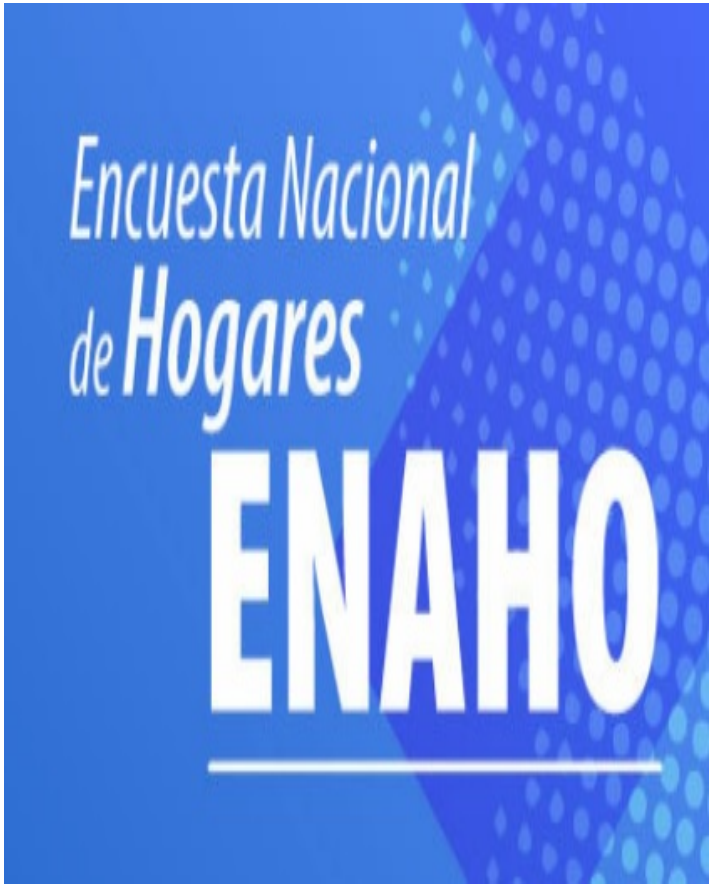
Edinson Tolentino

MSc. Economics
@edutoleraymondi

2023-01-09

Introducción

1. Objetivo :
    - interactuar con datos de encuestas (fuente secundaria)
    - conocimiento de datos a procesar
2. Pasos:
    - Importar datos
    - Manipular datos
3. Base de datos : ENAHO, modulos 500 y 300
4. Descargar datos del Encuesta Nacional de Hogares



Encuesta Nacional de Hogares(ENAHO)

- Se instala la libreria para poder instalar bases de datos de formato distinto a R.

```
  install.packages("readstata13")
library(readstata13)
#?readstata13
library(dplyr)
library(tidyverse)
```

- Se carga la base de datos en formato *stata*

```

# 1 definicion de carpetas -----------------------------

main <- "D:/Dropbox/Docencia/Cientifica/Seminario_2/L1/Data"

# Uniendo temas de texto o formatos de variables en texto
# comando paste0
paste0("Hola", " ","Mundo")
```

ENAHO : Importar datos de STATA a R

- Para la presente seccion usaremos la libreria *readstata13*
- Para la manipulacion de datos se usara: *dplyr* o *tidyverse*

```
library(readstata13)
library(tidyverse)
library(dplyr)
```

ENAHO: importar datos de ENAHO

1. Se usa dos objetos previamente definidos
    - Primero es el objeto main, el cual guarda la direccion o ruta de carpeta
    - Luego se usa el comando read.dta13 , el cual lee formatos de stata en su version 13
    - El nombre de la base de datos es el modulo 500

```
# Cargando la data
data <- read.dta13(paste0(main,"/","enaho01a-2021-500.dta"))
data %>%  names()
```

ENAHO: importar datos de ENAHO

- Otra forma de poder importar es respetando un orden en sus files, es decir, manteniendo una estructura de carpetas , y subcarpetas

```
# Ruta de mi carpeta dropox
ruta   <- "D:/Dropbox"
# Ruta de mi carpeta donde se encuentra la ENAHO
base   <- "/BASES/ENAHO"
# Ruta donde de guardan mis scrips
codigo <- "/Scripts"
# Ruta donde podre guardar la informacion
out    <- "/Docencia/Cientifica/Seminario_2/L1/Data"

data <- read.dta13(paste0(ruta,"/",base,"/","2021","/","enaho01a-202
```

ENAHO: Procesamiento de datos de ENAHO

- Construyendo el ingreso del trabajador, trabajaremos con las variables que se muestran en la imagen

v06Inglab **Ingreso laboral mensual de las ocupaciones principal y secundarias (En nuevos soles). Solo para la PEA ocupada (v03ConAct =1)**

v06IngLab = (I524A1 +D529T +I530A +D536 +I538A1 +D540T +I541A +D543 + D544T) / 12

Donde:

I524A1	=	Ingreso en la actividad principal por trabajo dependiente.
D529T	=	Pago en especie en la actividad principal por trabajo dependiente.
I530A	=	Ingreso en la actividad principal por trabajo independiente.
D536	=	Autoconsumo de los trabajadores independientes.
I538A1	=	Ingreso en la actividad secundaria por trabajo dependiente.
D540T	=	Pago en especie en la actividad secundaria por trabajo dependiente.
I541A	=	Ingreso en la actividad secundaria por trabajo independiente.
D543	=	Autoconsumo en la actividad secundaria por trabajo independiente.
D544T	=	Gratificación de navidad, Gratificación de fiestas patrias, Bonificación por últimas vacaciones, Bonificación por escolaridad, Participación de utilidades de la empresa, Bonificación por otro concepto relacionado con su trabajo, Compensación por tiempo de servicio (CTS), Otro ingreso por trabajo (reintegros, etc.).

Las variables de ingresos son indexadas y anualizadas por el INEI.

ENAHO: Procesamiento de datos de ENAHO

- Generando el codigo llave por cada persona
- La variable la llamaremos rid , que semeja una proxy de DNI en la encuesta
- Se mostrara las 5 primeras obsrvaciones

```
# Variable llave
data <- data %>%
  mutate(rid=paste0(conglome,vivienda, hogar,codperso))

# Mostrando las 5 observaciones
data$rid %>%  head(5)
```

```
## [1] "00500700311O1" "00500701211O1" "00500702211O1" "00500702211O2"
## [5] "00500702211O3"
```

ENAHO: Procesamiento de datos de ENAHO

- Se genera los departamentos con la variable *ubigeo*
- Para dicha variable se tiene en cuenta el comando *substr*

```
# Generacion de la variable departamento
data$rDpto <- substr(data$ubigeo,1,2)

# Realizar un tabulado de los 25 regiones
table(data$rDpto)
```

```
##
##    01    02    03    04    05    06    07    08    09    10    11    12
##  3091  3662  2302  3845  2604  3353  2701  3107  2431  3064  3807  3690
##    14    15    16    17    18    19    20    21    22    23    24    25
##  3997 12194  4113  1458  2244  2203  4459  2621  3467  3206  2185  2983
```

ENAHO: Procesamiento de datos de ENAHO

```
# Seleccionamos la informacion de variables
ingresos <- data %>%
  dplyr::select(i524a1, d529t, i530a, d536, i538a1, d540t, i541a, d54

# Sumamos las filas de cada columna de variable, solo para los ocupad
data$r6 <- ifelse(data$ocu500==1, rowSums(ingresos, na.rm=T)/12,0)

# Mostramos la informacion
data$r6 %>% head(5)
```

```
## [1] 6433.594 1062.166 2224.665 1619.210    0.000
```

ENAHO: Procesamiento de datos de ENAHO

- Se filtra la informacion que se va a trabajar
- se filtra las variables: rid , r6
- comando a usar *filter*

```
data_empleo <- data %>%
  dplyr::select(rid, r6,rDpto,rneduca)

# Mostrando las primeras 5 observacioes
data_empleo %>%  head(5)
```

```
##             rid       r6 rDpto
## 1 0050070031101 6433.594    01
## 2 0050070121101 1062.166    01
## 3 0050070221101 2224.665    01
## 4 0050070221102 1619.210    01
## 5 0050070221103    0.000    01
```

ENAHO: Procesamiento de datos de ENAHO

- Informacion por filtrar solo el caso de ingresos (r6) positivos, mayores de cero
-

```
Empleo_2021 <- data %>%
  filter(r6>0) %>%
  dplyr::select(rid, r6, rDpto) %>%
  drop_na()

Empleo_2021 %>%  head(5)
```

```
##              rid        r6 rDpto
## 1 0050070031101 6433.5939    01
## 2 0050070121101 1062.1659    01
## 3 0050070221101 2224.6652    01
## 4 0050070221102 1619.2104    01
## 5 0050090411101  397.5833    01
```

ENAHO: Analisis descriptivo

- Se realizara una tabla entre departamento e ingresos

```
#Ingresos por departamento
Tabla <- Empleo_2021 %>%
  group_by(rDpto) %>%
  summarise(Promedio=mean(r6),
            Desv = sd(r6))
```

```
## # A tibble: 25 × 3
##    rDpto Promedio  Desv
##    <chr>    <dbl> <dbl>
##  1 01       1037. 1182.
##  2 02       1123. 1457.
##  3 03       1082. 1125.
##  4 04       1551. 1931.
##  5 05       1001. 1210.
##  6 06        944. 1402.
##  7 07       1476. 1364.
##  8 08       1083. 1211.
##  9 09        759.  861.
## 10 10       1020. 1278.
## # … with 15 more rows
```

CONCLUSIONES

- La informacion final es una base de datos dado las personas encuestadas, la cual contiene las siguientes variables:
  - Ingresos : r6
  - Los departamenos: rDpto
- Realizar analisis descriptivos sobre la base de datos
- Realizar cruces de otras bases de datos como por ejemplo el modulo 300