Ejemplo de práctica EPH

Núcleo de Innovación Social & R en Buenos Aires

Betsy Cohen & Andrea Gómez Vargas

1 ¿Qué vamos a hacer?

Vamos a elaborar una tabla de datos con indicadores de Mercado de trabajo para el año 2023 similar a la presentada por el INDEC en el Informe de Mercado de trabajo del primer trimestre 2024

Cuadro 1.1 **Principales tasas del mercado de trabajo. Total 31 aglomerados urbanos. Primer trimestre 2023-primer trimestre 2024**

Tasas	Total 31 aglomerados urbanos				
	Año 2023				Año 2024
	1° trimestre	2° trimestre	3° trimestre	4° trimestre	1° trimestre
Actividad	48,3	47,6	48,2	48,6	48,0
Empleo	45,0	44,6	45,5	45,8	44,3
Desocupación abierta	6,9	6,2	5,7	5,7	7,7
Ocupados demandantes de empleo	14,8	16,2	15,2	15,5	16,0
Subocupación	9,4	10,6	10,3	10,5	10,2
Subocupación demandante	6,3	7,4	6,8	6,8	7,0
Subocupación no demandante	3,1	3,2	3,5	3,7	3,2

Fuente: INDEC, Encuesta Permanente de Hogares.

2 Definiciones Básicas

Antes de cálcular los indicadores tengamos en cuenta las siguientes definiciones de lo que vamos a calcular:

- **Población:** es el total de unidades al que expande la muestra.
- **Población Ocupada:** conjunto de personas que tienen por lo menos una ocupación

- **Población Desocupada:** personas que, no teniendo ocupación, están buscando activamente trabajo y están disponibles para trabajar.
- **Población Economicamente Activa**(PEA): la integran las personas que tienen una ocupación o que, sin tenerla, la están buscando activamente. Está compuesta por la población ocupada más la población desocupada.
- **Población Ocupada demandante de empleo:** población ocupada que busca activamente otra ocupación.
- **Población Subocupada:** se refiere a la subocupación por insuficiencia de horas, visible u horaria, y comprende a las personas ocupadas que trabajan menos de 35 horas semanales por causas involuntarias y que están dispuestas a trabajar más horas.
  - **Población Subocupada demandante**: población subocupada (por causas involuntarias y dispuesta a trabajar más horas) que además busca activamente otra ocupación.
  - **Población Subocupada no demandante**: población subocupada (por causas involuntarias y dispuesta a trabajar más horas) que no está en la búsqueda activa de otra ocupación.

Tanto la ocupación como la desocupación constituyen lo que se denomina **condición de actividad** que define la situación en que se encuentran las personas con respecto a su participación o no en la actividad económica.

3 ¿Qué vamos a aprender en R?

1. Cargar los datos directamente desde nuestra consola de R usando el paquete `{eph}`
2. Transformación de datos con `{dplyr}` para calcular indicadores básicos de mercado de trabajo
3. Elaborar un gráfico y una tabla usando `{gt}` y `{ggplot2}`

4 Manos al código

4.1 Antes de arrancar: instalación de programas y paquetes

Para poder trabajar ordenadamente te recomendamos seguir los pasos de instalación de R & Rstudio indicados en la sección de primeros pasos si todavía no los tienes para continuar con la instalación de los paquetes de trabajo:

Vamos a usar con la función `install.packages()`: `{tidyverse}`,`{gt}`,`{gtExtras}` & `{eph}`

```
install.packages("tidyverse")
install.packages("gt")
install.packages("gtExtras")
install.packages("eph")
```

Una buena práctica siempre que trabajes con **encuestas** es tener a mano:

- Los cuestionarios (te van a permitir entender bien la estructura y pases)
- El registro de diseño: donde tenes la referencia del código de pregunta, estructura y en caso que corresponda las categorías posibles que tiene que adoptar.
- Definiciones operacionales que realiza INDEC sobre condición de actividad, subocupación horaria y categoría ocupacional disponible aquí

4.2 Setup paquetes

El primer paso es llamar a los paquetes que vamos a usar con la función `library()` (por lo general siempre las colocamos al principio de nuestro script)

```
library(tidyverse) # para transformar los datos y graficarlos
library(eph) # para levantar los datos de la encuesta
library(gt) # para hermosear nuestras tablas
library(gtExtras) # para hermosear todavía más nuestras tablas
```

4.3 Obtener datos

Vamos a arrancar descargando los datos de la EPH con la función `get_microdata()`. En este caso vamos a descargar los resultados de la base de individuos por cada trimestre de 2023.

> TIP: Para revisar los parámetros de una función siempre podemos consultar la ayuda de R posicionándonos sobre la función y presionando la tecla F1.

```
ind_2023 <- get_microdata(
    year = 2023,
    period = 1:4,
    type = "individual",
    vars = "all")
```

En caso de que no quieras o puedas usar este paquete podes descargar las bases de microdatos en formato .txt en el sitio del INDEC en la sección Bases de datos y descargar EPH continua

4.4 Transformación de lo datos: armado de indicadores

Primero creamos una tabla con los siguientes indicadores de resumen (tomamos de base las definiciones mencionadas anteriormente):

- Población
- Ocupados
- Desocupados
- Población Economicamente Activa (PEA)
- Ocupados demandantes
- Subocupados (demandantes, no demandantes y total)

Estos niveles nos van a permitir calcular las tasas de forma sencilla.

> Para obtener cualquier tasa necesitamos crear un numerador y un denominador. Por ejemplo: Tasa de empleo = $Ocupados/Poblacin$.

Al ser una muestra en la EPH trabajamos con el total expandido por su ponderador, en este caso lo encontramos en la variable **PONDERA**.

4.5 Variables de la Base EPH

Para hacer la tabla vamos a trabajar con las siguientes variables de la base EPH:

- PONDERA: Es el factor de expansión para cada unidad de análisis
- ESTADO: Indica la condición de actividad y puede tomar los valores:
  - 0 = Entrevista individual no realizada (no respuesta al cuestionario individual)
  - 1 = Ocupado
  - 2 = Desocupado
  - 3 = Inactivo
  - 4 = Menor de 10 años
- PP03J: Es la pregunta: *Aparte de este/os trabajo/s, ¿estuvo buscando algún empleo / ocupación /actividad?*, que toma los valores:
  - 1 = Sí
  - 2 = No
  - 9 = Ns/Nr
- INTENSI: Refiere a la intensidad de la carga horaria laboral y cuyas categorías son:
  - 1 = Subocupado por insuficiencia horaria
  - 2 = Ocupado pleno
  - 3 = Sobreocupado
  - 4 = Ocupado que no trabajó en la semana
  - 9 = Ns/Nr

4.6 Tabla resumen

Con la función `summarise` del paquete `{dplyr}` podemos realizar distintas operaciones de resumen sobre las variables de nuestro análisis con la base que estamos trabajando:

```
tabla_1 <- ind_2023 |>
  group_by(TRIMESTRE) |>
  summarise(Poblacion          = sum(PONDERA),
            Ocupacion          = sum(PONDERA[ESTADO == 1]),
            Desocupacion       = sum(PONDERA[ESTADO == 2]),
            PEA                = Ocupacion + Desocupacion,
            Ocupacion_demandate = sum(PONDERA[ESTADO == 1 & PP03J ==1]),
            Suboc_demandante   = sum(PONDERA[ESTADO == 1 & INTENSI ==1 & PP03J==1]),
            Suboc_no_demand    = sum(PONDERA[ESTADO == 1 & INTENSI ==1 & PP03J %in% c(2,9)]),
            Subocupacion       = Suboc_demandante + Suboc_no_demand,
           'Actividad'                  = round(PEA/Poblacion*100,1),
           'Empleo'                     = round(Ocupacion/Poblacion*100,1),
           'Desocupación'               = round(Desocupacion/PEA*100,1),
           'Ocupación demandante'       = round(Ocupacion_demandate/PEA*100,1),
           'Subocupación'               = round(Subocupacion/PEA*100,1),
           'Subocupación demandante'    = round(Suboc_demandante/PEA*100,1),
           'Subocupación no demandante' = round(Suboc_no_demand/PEA*100,1)
  )

tabla_1
```

```
# A tibble: 4 x 16
  TRIMESTRE Poblacion Ocupacion Desocupacion      PEA Ocupacion_demandate
      <int>     <int>     <int>        <int>    <int>                <int>
1         1  29335155  13191361       973787 14165148              2096890
2         2  29398853  13107639       872119 13979758              2264700
3         3  29463819  13396202       813661 14209863              2159872
4         4  29528610  13517937       822050 14339987              2226764
# i 10 more variables: Suboc_demandante <int>, Suboc_no_demand <int>,
#   Subocupacion <int>, Actividad <dbl>, Empleo <dbl>, Desocupación <dbl>,
#   `Ocupación demandante` <dbl>, Subocupación <dbl>,
#   `Subocupación demandante` <dbl>, `Subocupación no demandante` <dbl>
```

Y ahora nos vamos a quedar solamente con las columnas de las tasas seleccionándolas con la función `select`

```
tabla_2 <- tabla_1 |>
  select(1,Actividad:'Subocupación no demandante')

tabla_2
```

```
# A tibble: 4 x 8
  TRIMESTRE Actividad Empleo Desocupación `Ocupación demandante` Subocupación
      <int>     <dbl>  <dbl>        <dbl>                  <dbl>        <dbl>
1         1      48.3   45            6.9                   14.8          9.4
2         2      47.6   44.6          6.2                   16.2         10.6
3         3      48.2   45.5          5.7                   15.2         10.3
4         4      48.6   45.8          5.7                   15.5         10.5
# i 2 more variables: `Subocupación demandante` <dbl>,
#   `Subocupación no demandante` <dbl>
```

Ya tenemos los datos que necesitamos pero parece que la tabla estuviera "acostada". Para que se ponga de pie o estire a lo largo vamos a usar la función `pivot_longer`

```
tabla_2 <- tabla_2 |>
  pivot_longer(cols = -c(TRIMESTRE), #Las columnas que queremos estirar en este caso son toda
               names_to = "Tasas", # El nombre que le queremos poner
               values_to = "Valor") # el valor en cuestión que hay que repartir


tabla_2
```

```
# A tibble: 28 x 3
   TRIMESTRE Tasas                      Valor
       <int> <chr>                      <dbl>
 1         1 Actividad                   48.3
 2         1 Empleo                      45
 3         1 Desocupación                 6.9
 4         1 Ocupación demandante        14.8
 5         1 Subocupación                 9.4
 6         1 Subocupación demandante      6.3
 7         1 Subocupación no demandante   3.1
 8         2 Actividad                   47.6
 9         2 Empleo                      44.6
10         2 Desocupación                 6.2
# i 18 more rows
```

Ahora vamos a crear una variable que combina el año y trimestre, desechando año y trimestre como variables por separado y vamos a "estirar" la tabla colocando al 1er trimestre de 2023 al final, tal cual se presenta en el reporte de INDEC

```
tabla_2 <- tabla_2  |>
  mutate(trimestre = paste(TRIMESTRE,"Trimestre")) |> # creamos la variable que concatena las
  select(-TRIMESTRE) |>
  pivot_wider(names_from = c("trimestre"), values_from = Valor) # estiramos la variable que

 tabla_2
```

```
# A tibble: 7 x 5
  Tasas                  `1 Trimestre` `2 Trimestre` `3 Trimestre` `4 Trimestre`
  <chr>                          <dbl>         <dbl>         <dbl>         <dbl>
1 Actividad                       48.3          47.6          48.2          48.6
2 Empleo                          45            44.6          45.5          45.8
3 Desocupación                     6.9           6.2           5.7           5.7
4 Ocupación demandante            14.8          16.2          15.2          15.5
5 Subocupación                     9.4          10.6          10.3          10.5
6 Subocupación demandan~           6.3           7.4           6.8           6.8
7 Subocupación no deman~           3.1           3.2           3.5           3.7
```

> Las funciones `pivot_longer` y `pivot_wider` nos ayudan a modificar la posición del dato (estirar y alargar) para saber más sobre ellas te recomendamos la lectura de esta viñeta donde vas a encontrar varios ejemplos.

5 Comunicación del dato

5.1 Tabla con gt

Ahora vamos a darle un poco de estilo y color con el paquete de "gramática de tablas" `{gt}`

```
tabla_3 <- tabla_2 |>
  gt(rowname_col = 'Tasas') |>
  # ponemos un título y subtitulo
  tab_header(
    title = md("**Cuadro 1: Principales tasas del mercado de trabajo. Total 31 aglomerados u
    subtitle = "Primer trimestre 2023-cuarto trimestre 2023") |>
  # centramos las columnas del cuerpo
  cols_align(
    align = "center",
```

```
    columns = 2:5
  ) |>
  tab_stubhead(label = md('**Tasas**')) %>%
   tab_stub_indent(rows=6:7, indent = 5) %>%
  # cambiamos las etiquetas de las columnas
  tab_source_note('Fuente: INDEC, Encuesta Permanente de Hogares.') |>
  tab_footnote("Datos ponderados.") %>%
  tab_options(
     heading.align = "left",
     heading.background.color = "#2B5597",
     footnotes.background.color = "#cedcf1",
     column_labels.font.weight = "bold"
  )

tabla_3
```

Cuadro 1: Principales tasas del mercado de trabajo. Total 31 aglomerados urbanos.

Primer trimestre 2023-cuarto trimestre 2023

Tasas	1 Trimestre	2 Trimestre	3 Trimestre	4 Trimestre
Actividad	48.3	47.6	48.2	48.6
Empleo	45.0	44.6	45.5	45.8
Desocupación	6.9	6.2	5.7	5.7
Ocupación demandante	14.8	16.2	15.2	15.5
Subocupación	9.4	10.6	10.3	10.5
Subocupación demandante	6.3	7.4	6.8	6.8
Subocupación no demandante	3.1	3.2	3.5	3.7

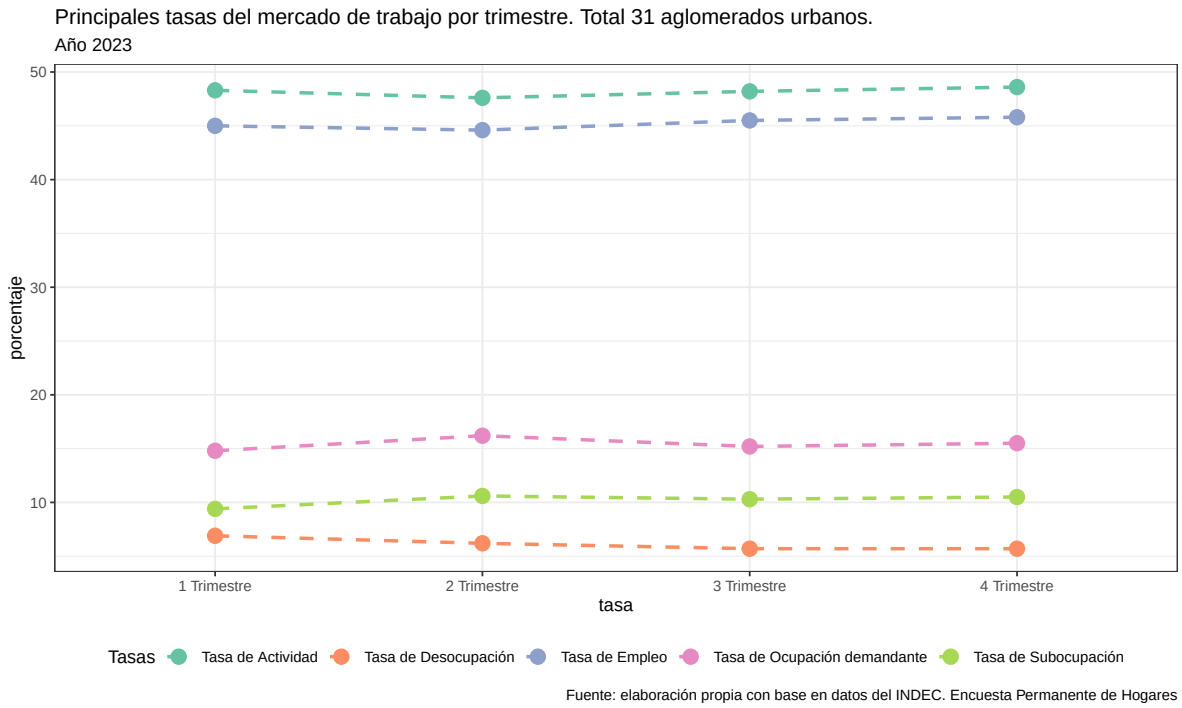
Datos ponderados.
Fuente: INDEC, Encuesta Permanente de Hogares.

5.2 Gráfico con ggplot2

```
subtabla <- tabla_2 %>%
  pivot_longer(cols = c(2:5),names_to = "Trimestre", values_to = "value") %>%
  mutate(Tasas =  paste0("Tasa de ",Tasas)) %>%
  filter(!Tasas %in% c("Tasa de Subocupación demandante", "Tasa de Subocupación no demandante
```

```r
grafico1 <- ggplot(subtabla, aes(Trimestre, value, colour =  Tasas, group = Tasas)) +
  geom_point(size = 4) +
  geom_line(linetype = 2,linewidth = 1) +
  theme_bw() +
  theme(legend.position = "bottom") +
  scale_color_brewer(palette = "Set2") +
    labs(
    title = "Principales tasas del mercado de trabajo por trimestre. Total 31 aglomerados urb
    subtitle = "Año 2023",
    y = "porcentaje",
    x = "tasa",
    caption = "Fuente: elaboración propia con base en datos del INDEC. Encuesta Permanente de
  )

grafico1
```



6 Licencia

Este taller está a disposición del público bajo licencias de código abierto. Toda la documentación y los materiales publicados están disponibles bajo una licencia **CC**

BY-SA-NC.

A continuación, puedes leer más sobre la licencia de Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional.

Figura 1: Licencia de Creative Commons