PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

# Tarea

## Análisis de datos en R

Fecha de entrega: 11/08/2023 hasta las 23:59 hrs.

**Aspectos formales:**

- Deberá entregar una carpeta comprimida (`.zip`) que contenga: a) su reporte con las respuestas (`.pdf` o `.docx`), b) su proyecto (`.Rproj`) y c) el código de análisis (`.R`) **Este código debe ser reproducible**.
- Los aspectos de formato son flexibles, pero respetando la formalidad de un documento de trabajo.
- Todo el procesamiento, análisis, tablas y visualizaciones de datos deben ser realizado en `R`. En el código debe comentar el paso a paso de su tarea.
- Fecha de entrega: 11/08/2023 hasta las 23:59 hrs.
- Debe enviar su evaluación al profesor (jdconejeros@uc.cl)
- La tarea se puede realizar en parejas o individual.

## Descripción

El objetivo de la tarea es practicar las principales herramientas de `R` y `RStudio` para la manipulación de objetos, lectura de tablas y análisis. Se espera que puedan utilzar las herramientas vistas a lo largo de las sesiones.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

## Ejercicio 1

a. Genere en su `enviroment` los siguientes vectores:

`vector_1`: Contiene una secuencia de 9 numeros que inicia en 1 y termina en 5.

`vector_2`: Contiene una secuencia de números impares que van desde el 1 hasta el 20 de dos en dos.

`vector_3`: Contiene el número `2023` repetido 10 veces.

`vector_4`: Contiene 5 repeticiones del número `1` y 5 repeticiones del número `0`.

`vector_5`: Contiene la suma del `vector_1` y del `vector_2`.

Para el `vector_5` revise el `Warning message` que indica R y explique cuál la coerción implícita aplicada. Además, explique la diferencia entre un vector, una matris de datos y una lista. **(2 puntos)**

b. Genere una matriz con los vectores definidos en **a**. (*Hint:* recuerde que la información se pega por columnas). Debiera llegar al siguiente resultado:

```
      vector_1 vector_2 vector_3 vector_4 vector_5
 [1,]      1.0        1     2023        1      2.0
 [2,]      1.5        3     2023        1      4.5
 [3,]      2.0        5     2023        1      7.0
 [4,]      2.5        7     2023        1      9.5
 [5,]      3.0        9     2023        1     12.0
 [6,]      3.5       11     2023        0     14.5
 [7,]      4.0       13     2023        0     17.0
 [8,]      4.5       15     2023        0     19.5
 [9,]      5.0       17     2023        0     22.0
[10,]      1.0       19     2023        0     20.0
```

Explique la coerción implícita que se aplica en la columna `vector_1`. ¿Por qué sería importante saber esto? ¿Cuál es la diferencia entre mantener este objeto como una matriz a transformarlo en un data-frame? **(2 punto)**

c. El estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ es una especificación lineal para el vector de respuesta $\boldsymbol{Y}$ que tiene una distribución normal multivariada ($\boldsymbol{Y} \sim Normal(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$). Dicho estimador es de la forma:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \sim Normal(\boldsymbol{\beta}, (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2)$$

Donde $\boldsymbol{X}$ es el vector que predice $\boldsymbol{Y}$[1]. En otras palabras, el estimador $\hat{\boldsymbol{\beta}}$ corresponde a la multiplicación entre la inversa del vector $\boldsymbol{X}$ traspuesto multiplicada por el vector $\boldsymbol{X}$, y el vector $\boldsymbol{X}$ traspuesto multiplicado por el vector $\boldsymbol{Y}$. El estimador $\hat{\boldsymbol{\beta}}$ se interpretará como la asociación entre el predictor $\boldsymbol{X}$ y mi vector de respuesta $\boldsymbol{Y}$. Este estimador será fundamental a la hora de evaluar el impacto en políticas públicas.

Utilizando la tabla de datos `simce2m2016_extracto.csv` calcule manualmente el estimador $\hat{\boldsymbol{\beta}}$ que nos permite validar la siguiente hipótesis: *"Los resultados en el SIMCE están negativamente asociados. Un resultado positivo en la prueba de lectura implica un resultado negativo en la prueba de matemáticas"*[2]. ¿Qué puede decir de su hipótesis? ¿La relación entre el resultado de ambas pruebas es negativo o positivo? Justifique. **(2 puntos)**

[1] Más detalles en: https://sochinf.cl/prevenir-sospechar-diagnosticar-virus-hanta-campana-verano-2023/
[2] Para construir la variable año utilice la fecha de notificación del caso: `fecha_notificacion`.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

## Ejercicio 2

La infección por Hantavirus, cuya mayor expresión de gravedad es el Síndrome Cardiopulmonar (SCPH), es una zoonosis endémica de Chile causada por el virus Andes (ANDV). El ANDV es un virus de genoma ARN segmentado y con envoltura lipídica, miembro del género Orthohantavirus y la familia Hantaviridae. Este virus tiene como reservorio natural el roedor Oligoryzomys longicaudatus conocido como "ratón colilargo o cola larga", cuyo hábitat se encuentra distribuido desde el valle de Copiapó en la III Región a Campos de Hielo Sur (50° S)[3].

El ministerio de salud le encarga a usted realizar un análisis descriptivo de la situación histórica de los casos de hantavirus y entregar recomendaciones para una mejor atención y control. Para esto usted trabajara con la tabla de datos `Hantavirus_chile.xlsx` que cuenta con el registro histórico de casos identificados de esta infección. A partir de esto se le pide realizar las siguientes tareas:

a. Importe su tabla de datos e indique cuántas infecciones por Hantavirus se han registrado a la fecha. Explique cuál es la unidad de análisis de estos datos ¿La tabla esta compuesta por personas únicas o hay duplicados? Explique que sería un duplicado en este caso. Explore si hay casos pérdidos en sus datos y concluya. **(1.0 punto)**

b. Construya dos tablas con los porcentajes de 1) infecciones de hantavirus por año desagregado por sexo y 2) infecciones de hantavirus por año desagregado por grupo etarios[4]. ¿Qué le podría comentar al Ministerio de Salud respecto a sus resultados? **(1.0 punto)**

Debería llegar a algo de este estilo (solo es una referencia):

Table 1: Porcentaje de infectados por hantavirus por año y sexo

| Año | Sexo | N | % |
|---|---|---|---|
| 1995 | mujer | 1 | 100.0 |
| 1996 | hombre | 3 | 100.0 |
| 1997 | hombre | 21 | 77.8 |
| 1997 | mujer | 6 | 22.2 |
| 1998 | hombre | 28 | 73.7 |
| 1998 | mujer | 10 | 26.3 |
| 1999 | hombre | 18 | 72.0 |
| 1999 | mujer | 7 | 28.0 |
| 2000 | hombre | 23 | 82.1 |
| 2000 | mujer | 5 | 17.9 |

*Fuente:* Elaboración propia.

Puede trabajar con las funciones `mutate()`, `group_by()`, `summarise()`.

[3] Más detalles en: https://sochinf.cl/prevenir-sospechar-diagnosticar-virus-hanta-campana-verano-2023/
[4] Para construir la variable año utilice la fecha de notificación del caso: `fecha_notificacion`.

Table 2: Porcentaje de infectados por hantavirus por año y edad

| Año | Sexo | N | % |
|---|---|---|---|
| 1995 | 20-24 | 1 | 100.0 |
| 1996 | 20-24 | 1 | 33.3 |
| 1996 | 25-29 | 1 | 33.3 |
| 1996 | 30-34 | 1 | 33.3 |
| 1997 | 0-4 | 1 | 3.7 |
| 1997 | 10-14 | 3 | 11.1 |
| 1997 | 15-19 | 1 | 3.7 |
| 1997 | 20-24 | 2 | 7.4 |
| 1997 | 25-29 | 2 | 7.4 |
| 1997 | 30-34 | 5 | 18.5 |

*Fuente:* Elaboración propia.

c. Construya una tabla con el número de casos por región (`region_residencia`) agrupando cada 5 años[5]. Indique la región con el mayor cantidad de casos en el tiempo y realice un zoom para identificar las comunas más críticas (mayor número de casos totales) para esa región ¿Qué input relevante le podría indicar al Ministerio de Salud? ¿Dónde podríamos tener una mayor vigilancia? **(2.0 puntos)**

Debería llegar a algo de este estilo (solo es una referencia):

Table 3: Porcentaje de infectados por hantavirus por región y año

| Región | Período | N |
|---|---|---|
| Region del Maule | 1995 - 1999 | 2 |
| Region del Maule | 2000 - 2004 | 29 |
| Region del Maule | 2005 - 2009 | 28 |
| Region del Maule | 2010 - 2014 | 32 |
| Region del Maule | 2015 - 2022 | 47 |

*Fuente:* Elaboración propia.

El análisis lo puede realizar con las funciones `mutate()`, `filter()`, `group_by()`, `summarise()`.

d. Una preocupación importante del MINSAL es comprender la dinámica entre el tiempo en que aparecen los primeros síntomas y la notificación de los casos a las autoridades. Construya una variable nueva que represente el número de días entre la notificación y la aparición de los primeros sintomas. Luego realice un análisis descriptivos (medidas de tendencia central, dispersión y posición) de su variable de interés. Puede presentar una figura si es que lo estima conveniente. ¿Qué puede decir respecto a los tiempos de notificación de casos de Hantavirus

[5] Considere como primer bloque de tiempo 1995 - 1999 y último bloque de tiempo: 2015 - 2022.

en Chile? Sea breve. Puede apoyarse de la función `difftime()` u otra que estime conveniente. **(0.5 punto)**

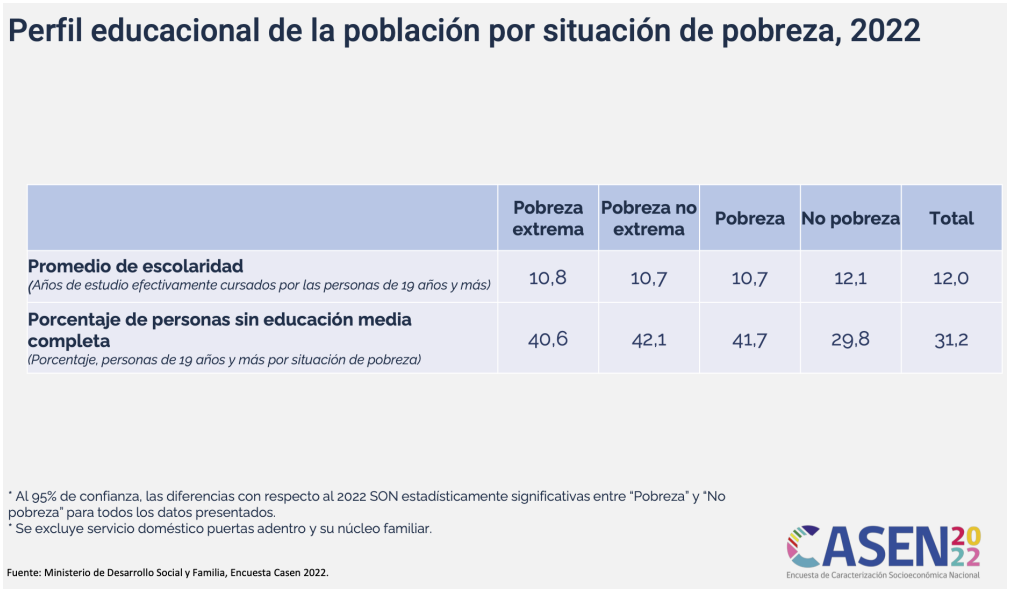
e. A partir de sus resultados y su experiencia, ¿qué medidas podría proponer al Ministerio con el objetivo de tener un mejor control y reducir los casos de Hantavirus en el país? Sea breve. **(0.5 puntos)**

f. Utilice un gráfico de líneas para describir la serie temporal con el número de casos de Hantavirus según año-mes. Presente una primera serie general para todos los datos y luego otra figura con las series desagregadas por grupo etario. Utilice la variable `fecha_notificacion` para construir sus series. ¿Qué puede información puede extraer de esta figura? **(1.0 puntos)**

## Ejercicio 3

A continuación se le pide replicar y analizar los siguientes resultados obtenidos de estudios reales:

a. **Resultado 1: (2.0 puntos)**



**Perfil educacional de la población por situación de pobreza, 2022**

| | Pobreza extrema | Pobreza no extrema | Pobreza | No pobreza | Total |
|---|---|---|---|---|---|
| **Promedio de escolaridad** *(Años de estudio efectivamente cursados por las personas de 19 años y más)* | 10,8 | 10,7 | 10,7 | 12,1 | 12,0 |
| **Porcentaje de personas sin educación media completa** *(Porcentaje, personas de 19 años y más por situación de pobreza)* | 40,6 | 42,1 | 41,7 | 29,8 | 31,2 |

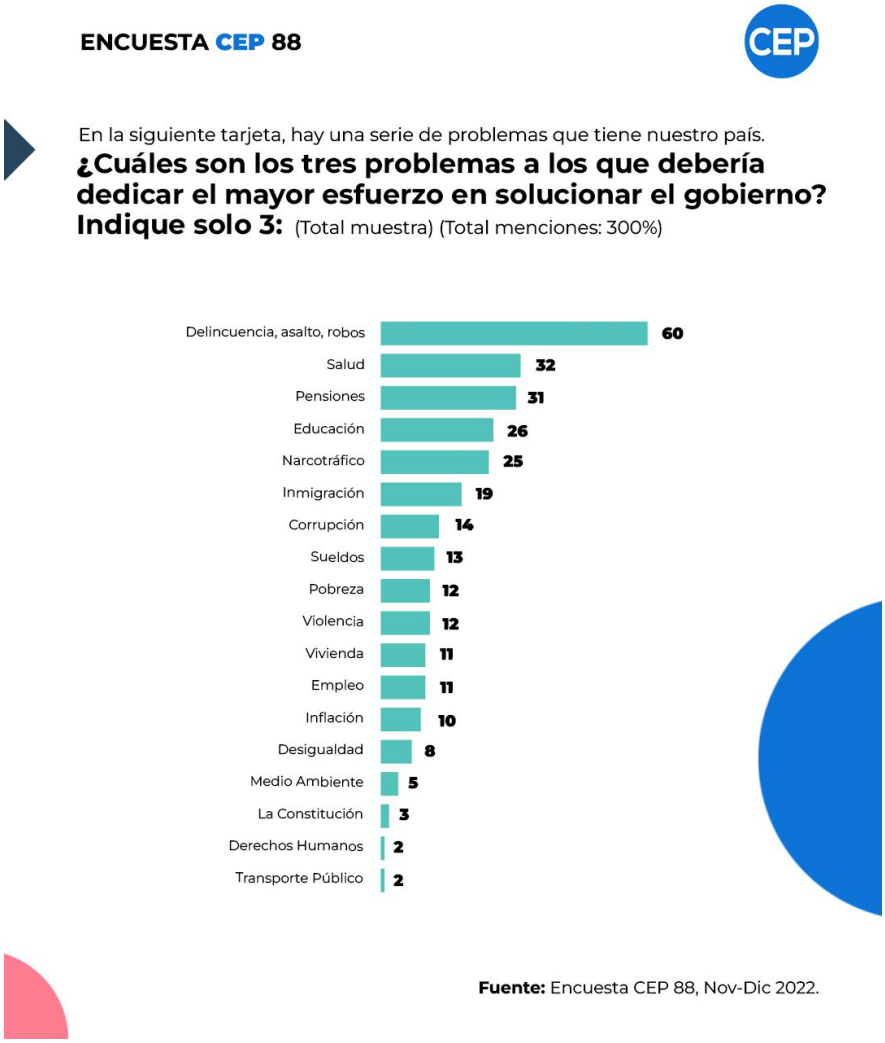
* Al 95% de confianza, las diferencias con respecto al 2022 SON estadísticamente significativas entre "Pobreza" y "No pobreza" para todos los datos presentados.
* Se excluye servicio doméstico puertas adentro y su núcleo familiar.

Fuente: Ministerio de Desarrollo Social y Familia, Encuesta Casen 2022.

CASEN 2022
Encuesta de Caracterización Socioeconómica Nacional

- Reporte de resultados: Resultados pobreza por ingresos, CASEN 2022
- Fuente de datos: Click aquí

b. **Resultado 2: (2.0 puntos)**




- Reporte de resultados: Encuesta CEP N° 88, Noviembre-Diciembre 2022
- Fuente de datos: Click aquí

c. A partir de sus intereses, trabajo actual o curiosidad, genere un breve análisis incluyendo una pregunta de interés, gráfico, una tabla y una interpretación. **(2.0 puntos)**