PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Tarea 2

Taller de análisis de datos I

Fecha de entrega: 15/08/2023 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: 17/07/2023 hasta las 23:59 hrs. **No se recibirán evaluaciones después de la fecha de entrega**
- Debe enviar su evaluación al profesor (jdconejeros@uc.cl) y a la ayudante del curso (samadariaga@uc.cl)
- La tarea deberá ser realizada en parejas o individual.
- Sea breve, simple y con un lenguaje directo en sus interpretaciones.

Tutoría el lunes 07/08/2023 (19:00 - 20:30 hrs): tendremos una tutoría voluntaria en que la ayudante entregara tips y resolvera dudas de la tarea. Dudas al correo del profesor y/o la ayudante. También pueden recurrir a cualquier apoyo disponible en el web, además de la literatura del curso.

Descripción

El objetivo de la tarea 2 es practicar las principales herramientas de `R`, `RStudio` y `dplyr` para la manipulación tablas y análisis. Se espera que puedan utilzar las herramientas vistas en el laboratorio 1, 2 y 3 del curso.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Ejercicio 1 (5 puntos)

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)[1].

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. **(0.5 puntos)**

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[2]. ¿Qué le podría comentar al Ministerio de Salud respecto a sus resultados? **(1.5 puntos)**

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.

[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`.

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.

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

c. Construya una tabla con el número de casos por región (`region_residencia`) agrupando cada 5 años[3]. 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? **(1.5 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

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

conveniente. ¿Qué puede decir respecto a los tiempos de notificación de casos de Hantavirus en Chile? Sea breve. Puede apoyarse de la función `difftime()` u otra que estime conveniente. **(1 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)**

Ejercicio 2 (2 puntos)

En el ejercicio 1 se le pidió realizar una extracción desde la API del Banco Mundial y realizar análisis a partir de su extracción. En este caso, usted trabajará con la siguiente extracción de la API que corresponde al ingreso per cápita (GDP) para América Latina y el Caribe solo para el año 2020.

```
library(WDI)
gdp <- WDI(
    country = "all",
    indicator = "NY.GDP.PCAP.PP.KD",
    start = 2015,
    end = 2020,
    extra = TRUE,
    cache = NULL,
    latest = NULL,
    language = "es") %>%
  filter(region=="Latin America & Caribbean" & year==2020)
```

Usted debería tener una tabla con las siguientes columnas:

```
glimpse(gdp)
```

```
Rows: 22
Columns: 13
$ country           <chr> "Argentina", "Aruba", "Barbados", "Bolivia", "Chile"~
$ iso2c             <chr> "AR", "AW", "BB", "BO", "CL", "CO", "CR", "CU", "CW"~
$ iso3c             <chr> "ARG", "ABW", "BRB", "BOL", "CHL", "COL", "CRI", "CU~
$ year              <int> 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020, 2020~
$ NY.GDP.PCAP.PP.KD <dbl> 19685.216, 33155.243, 13805.778, 7679.933, 22970.550~
$ status            <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
$ lastupdated       <chr> "2023-06-29", "2023-06-29", "2023-06-29", "2023-06-2~
$ region            <chr> "Latin America & Caribbean", "Latin America & Caribb~
$ capital           <chr> "Buenos Aires", "Oranjestad", "Bridgetown", "La Paz"~
$ longitude         <chr> "-58.4173", "-70.0167", "-59.6105", "-66.1936", "-70~
```

```
$ latitude          <chr> "-34.6118", "12.5167", "13.0935", "-13.9908", "-33.4~
$ income            <chr> "Upper middle income", "High income", "High income",~
$ lending           <chr> "IBRD", "Not classified", "Not classified", "IBRD", ~
```

Además usted sabe que el promedio del GDP para los países en el 2020 es de:

```
mean(gdp$NY.GDP.PCAP.PP.KD, na.rm = TRUE)
```

```
[1] 16154.51
```

A continuación se realiza el siguiente flujo empaquetado en una función llamada `muestreo`:

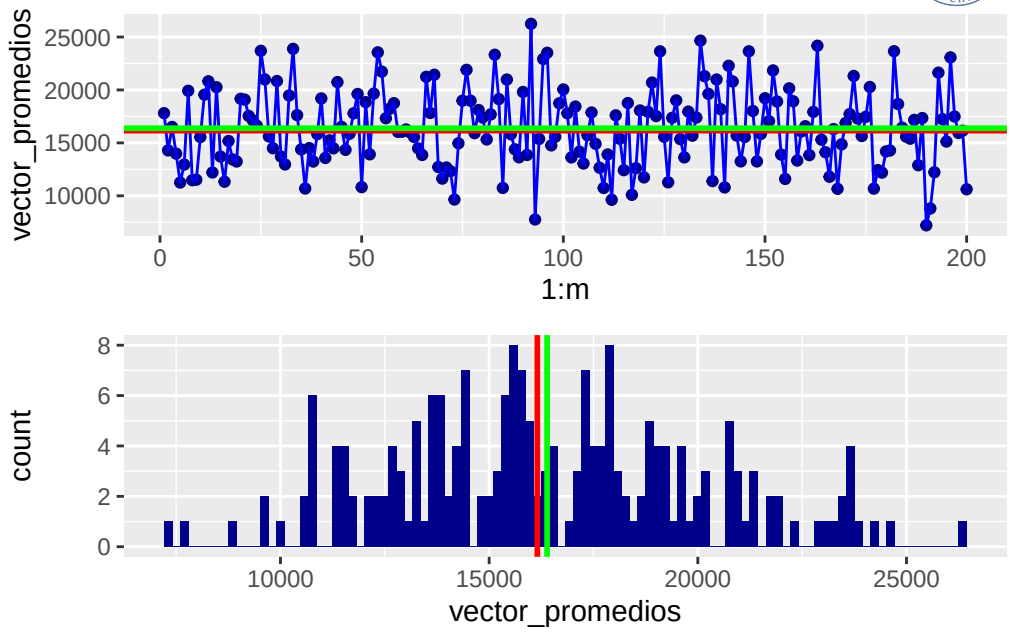
```
muestreo <- function(v, m, n, replace = TRUE) {
  library(dplyr)
  library(ggplot2)
  library(patchwork)

  vector_promedios <- c()
  for(i in 1:m) {
    muestra <- sample(x = v,
                      size = n,
                      replace = TRUE)
    vector_promedios[i] <- mean(muestra, na.rm=TRUE)
  }

  promedio <- mean(vector_promedios, na.rm=TRUE)

  g1 <- ggplot(data = NULL, aes(x = 1:m, y = vector_promedios)) +
    geom_point(color = "darkblue") + geom_line(color = "blue") +
    geom_hline(yintercept = mean(v, na.rm=TRUE), color = "red", lwd = 1) + #Promedio real
    geom_hline(yintercept = promedio, color = "green", lwd = 1) #Gran media


  g2 <- ggplot(data = NULL, aes(x = vector_promedios)) +
    geom_histogram(bins = 100, fill="darkblue") +
    geom_vline(xintercept =  mean(v, na.rm=TRUE), color = "red", lwd = 1) +
    geom_vline(xintercept =  promedio, color = "green", lwd = 1)

  grafico <- g1/g2

  print(paste0("Promedios de ", m, " muetras de tamaño ", n, ":"))
  print(vector_promedios)
```

```
  print(paste0("Promedio de promedios (Gran media): ", promedio))
  print(paste0("Promedio 'verdadero': ", mean(v, na.rm=TRUE)))
  print(grafico)
}
```

Utilizando la función `muestreo()` podemos obtener los siguientes resultados:

```
muestreo(v=gdp$NY.GDP.PCAP.PP.KD, m=200, n=5, replace = TRUE)
```

```
[1] "Promedios de 200 muetras de tamaño 5:"
  [1] 17810.490 14292.508 16502.313 13985.085 11252.901 12953.488 19923.506
  [8] 11474.926 11518.237 15551.771 19554.568 20821.058 12203.462 20263.931
 [15] 13706.425 11331.524 15181.342 13432.894 13243.380 19165.471 19086.756
 [22] 17547.332 17138.806 16631.065 23698.583 20994.776 15610.650 14494.408
 [29] 20843.670 13705.703 12963.292 19478.771 23881.307 17621.054 14384.438
 [36] 10688.481 14508.620 13238.764 15834.599 19199.231 13577.472 15241.511
 [43] 14484.462 20741.620 16510.508 14343.452 15863.850 17794.526 19617.945
 [50] 10819.874 18848.080 13919.751 19652.058 23552.921 21729.294 17326.179
 [57] 18232.019 18752.380 16041.353 16055.335 16269.889 15999.116 15555.994
 [64] 14483.415 13858.433 21238.716 17826.166 21438.391 12719.221 11624.281
 [71] 12677.827 12328.583  9648.682 14962.407 18986.944 21902.799 18982.738
 [78] 15920.058 18105.810 17394.018 15321.329 17689.485 23328.800 19136.435
 [85] 10743.466 20985.073 15814.356 14392.869 13636.082 19811.473 13861.101
 [92] 26248.457  7763.386 15374.717 22927.358 23512.737 14762.387 15573.053
 [99] 18736.867 20049.373 17797.013 13628.829 18436.222 14167.197 13047.442
[106] 15727.275 17870.579 14913.861 12652.250 10743.041 13902.704  9623.416
[113] 17604.934 15387.281 12420.896 18766.300 10093.395 12608.627 18064.904
[120] 11740.130 17940.173 20700.353 17534.300 23649.411 15579.583 11274.322
[127] 17368.784 19016.565 15352.437 13632.306 17955.730 15678.901 17392.379
[134] 24659.181 21302.075 19626.578 11387.834 20989.016 18184.041 10793.485
[141] 22285.096 20799.704 15686.951 13253.523 15557.266 23653.609 18011.972
[148] 13249.277 15856.708 19218.110 17049.828 21844.844 18909.051 13896.702
[155] 11596.285 20155.400 18928.986 13332.665 16082.190 16573.643 13843.815
[162] 17928.332 24179.057 15300.903 14113.225 11798.633 16302.253 10655.964
[169] 14862.072 16879.401 17723.387 21324.984 17301.107 15645.381 17467.802
[176] 20279.190 10676.434 12428.985 12191.173 14191.262 14280.070 23660.510
[183] 18666.383 16439.802 15588.781 15394.902 17173.114 12886.135 17345.895
[190]  7211.666  8806.473 12224.951 21636.188 17239.688 15125.565 23077.527
[197] 17493.503 15936.181 16146.492 10615.338
[1] "Promedio de promedios (Gran media): 16388.0727723567"
[1] "Promedio 'verdadero': 16154.5099216273"
```



A partir de la función `muestreo()` y el análisis detallado resuelva lo siguiente:

a. Describa lo que realiza la función. ¿Cuál es el objetivo de aplicar este proceso sobre los datos? **(0.5 puntos)**

b. Explique en qué consisten los inputs `v`, `m`, `n` y `replace = TRUE` de la función propuesta. **(0.5 puntos)**

c. Interprete las dos figuras que se generan a partir de este flujo. **(0.5 puntos)**

d. ¿En qué se diferencian de forma conceptual la "gran media" de la "media verdadera"? Explique e interprete el resultado obtenido. **(0.5 puntos)**

Bonus (+0.5)

Utilizando la tabla de datos `Hantavirus_chile.xlsx`, genere cualquiera de los dos inputs solicitados a continuación (solo 1) y entregue una breve lectura de sus resultados. Sea breve, simple y con un lenguaje directo.

Input 1: Gráficos de líneas con la serie temporal para 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.

Input 2: Realice mapas de Chile con el número de casos de hantavirus por región o comuna (usted decide que es mejor). Estos mapas deben ser desagregados por períodos:

- Un mapa para el número de casos entre 1995 - 2004
- Un mapa para el número de casos entre 2005 - 2014
- Un mapa para el número de casos entre 2015 - 2022

Para esto se puede apoyar de la librería `sf` y `chilemapas`.