Tarea 1

Fecha de entrega: 31-agosto

Nota: Deberán subir a *Canvas* un archivo de texto con sus respuestas. Pueden utilizar el formato de su preferencia (e.g. LaTeX, Word u hojas escritas a mano y escaneadas). Además, deberán subir otro archivo que genere todos sus resultados de las preguntas prácticas. Puede ser un archivo de Excel, R-script o Do-File.

En esta tarea exploraremos algunos aspectos del COVID-19 utilizando conceptos básicos de estadística. Encontrarás los archivo BaseCOVIDp y BaseCOVIDm en *Canvas* (ambos disponibles en formato .dta y .csv). La primera base se llama BaseCOVIDp y contiene la información recopilada por el *Center for Systems Science and Engineering (CSSE) at Johns Hopkins University* hasta el 29 de julio de 2020 sobre la pandemia de COVID-19 para 182 países. Además se agregaron a esta base datos de la población total por país provenientes del Banco Mundial e información sobre pruebas realizadas provista por las autoridades sanitarias de los diferentes países. La segunda base se llama BaseCOVIDm y es simplemente una muestra de 100 países de la BaseCOVIDp elegidos de forma aleatoria[1].

Las variables que integran las bases de datos son:

Variable	Descripción
Country	Nombre del país.
ISO_code	Combinación de tres letras asignada a cada país.
Confirmed	Número total de casos confirmados a la fecha de corte.
Deaths	Número total de muertes a la fecha de corte.
Active	Número total de casos activos a la fecha de corte
Recovered	Número total de personas recuperadas a la fecha de corte.
Population	Población del país (año 2019).
Tests	Número de individuos con pruebas realizadas.

1. Utilizando la base BaseCOVIDm realiza lo siguiente [Tip R: Para cargar la base de datos en formato .csv puedes utilizar la función `read.csv`. Para cargar la base en formato .dta puedes utilizar el paquete `haven` y la función `read_dta`("nombre_dataframe_stata")]:

   (a) (*2.5 puntos*) Calcula la tasa de positividad[2] acumulada para COVID-19 de cada país (POS_i), dicha variable se define como $POS_i = \frac{Confirmed_i}{Tests_i}$. Calcula la media y varianza muestral. Calcula un intervalo de confianza al 95 % para la media. [Tip R: Puedes utilizar paquete `dplyr` y la función `mutate` para crear nuevas variables].

[1] Considera que en las bases de datos hay países que no cuentan con información para todas las variables.
[2] La tasa de positividad corresponde a la proporción de pruebas que resultaron positivas.

(b) (*5 puntos*) Dado el número reducido de pruebas que México ha hecho, te planteas la hipótesis de que México debe tener una tasa de positividad mayor que la media mundial. Plantea la prueba de hipótesis relevante para evaluar esta afirmación. Reporta el *valor-p* e indica qué concluyes con respecto a la prueba hipótesis planteada.

(c) (*5 puntos*) Durante finales del año 2002 comenzó en diversas provincias de China el brote de SARS-CoV, que guarda similitudes con el SARS-CoV-2, virus responsable de la pandemia actual. Nos interesa comparar la tasa de fatalidad (CFR) de ambas pandemias. Por esto crearás la variable $CFR_i = \frac{Deaths_i}{Confirmed_i}$, la cual es la tasa de fatalidad del país i. Un artículo periodístico indica que la tasa de fatalidad media del SARS-CoV es cuatro veces mayor que la del SARS-CoV-2. Plantea la prueba de hipótesis relevante para evaluar dicha afirmación y construye el estadístico t que necesitarías para realizar dicha prueba. Asume que la media de la tasa de fatalidad del SARS-CoV es conocida (de acuerdo a la OMS) e igual a $9.6\,\%$[3]

(d) (*7.5 puntos*) Dada la media de la tasa de fatalidad del SARS-CoV ($9.6\,\%$), ¿a partir de que nivel de confianza, el intervalo de confianza relevante para nuestra prueba de hipótesis ya no incluiría el valor de $\frac{0.096}{4}$? Describe la relación entre tu respuesta y el *valor-p*.

2. Estamos interesados en conocer el porcentaje de la población total de cada país que ha contraido el virus (PPI), por ello creamos la variable $PPI_i = \frac{Confirmed_i}{Population_i}$.

(a) (*5 puntos*) Utilizando la base BaseCOVIDp construye un histograma de la variable PPI y calcula el primer cuartil de su distribución. [Tip R: Puedes utilizar el paquete `ggplot2` y la función `geom_hist` para elaborar un histograma. La función `quantile` te permite estimar el percentil que desees.]

(b) (*5 puntos*) Utilizando la base BaseCOVIDm construye un histograma de dicha variable y calcula el primer cuartil.

(c) (*5 puntos*) ¿Cuál es la relación que existe entre los histogramas de los incisos anteriores? ¿Ex-ante esperabas que los histogramas se parecieran?

(d) (*10 puntos*) Utilizando el método Bootstrap, genera 1000 submuestras del tamaño de la muestra original ($n = 100$) partiendo de la base BaseCOVIDm. Para cada submuestra, calcula el primer cuartil y grafica un histograma de los $1,000$ cuartiles estimados. [Tip R: Puedes construir un for loop en el que en cada ciclo generes una submuestra con la función `sample` y a dicha submuestra le calcules el primer cuartil. Considera que la función `sample` se aplica sobre vectores, en

[3]Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003

este caso el vector sería la columna de tu dataframe que corresponde a la variable PPI; la opción `replace` indica si la muestra se obtiene con reemplazo].

(e) (*10 puntos*) Repite el inciso anterior pero ahora utilizando 1000 submuestras de tamaño 70. Utilizando una gráfica compara el histograma del inciso anterior con el histograma que se produce en este inciso. ¿Qué similitudes y diferencias notas?

(f) (*7.5 puntos*) Utilizando los resultados de los dos incisos anteriores construye un intervalo de confianza del 99 % para el primer cuartil, grafícalos y ubica en dicha gráfica el valor del primer cuartil poblacional.

3. A lo largo de los meses que ha durado la pandemia han surgido diversas preguntas. Una de ellas se enfoca en la relación que existe entre el porcentaje de personas a las que se les han realizado pruebas $\left(PT = \frac{Tests}{Population}\right)$ y el porcentaje de personas contagiadas $\left(PPI = \frac{Confirmed}{Population}\right)$. Utilizando los datos de BaseCOVIDm contestaremos las siguientes preguntas:

(a) (*2.5 puntos*) Realiza un diagrama de dispersión (*scatterplot*) utilizando la variable PT en el eje X y PPI en el eje Y. [Tip R: Puedes utilizar el paquete `ggplot2` y la función `geom_point` para graficar un *scatterplot*].

(b) (*5 puntos*) Sea PT_{mex} el valor de la variable PT para México. Explora cuántos países tienen una valor PT 0.005 mayores o menores al valor de PT_{mex}. Utilizando sólo estos países, grafica un histograma de la variable *PPI*. Ubica en dicho gráfico el valor de la media de PPI (solo utilizando este conjunto de paises) y ubica el valor de la variable PPI corresponidente a México. ¿Está México por encima o por debajo de la media de estos países?

(c) (*5 puntos*) Utilizando el método de mínimos cuadrados ordinarios, estima la siguiente regresión:

$$PPI_i = \beta_0 + \beta_1 \ PT_i + U_i \tag{1}$$

Grafica la recta que resulta de esta regresión junto con el diagrama de dispersión y resalta el punto que corresponde a México en dicha gráfica. ¿Está México por encima o por debajo de la recta? ¿Cómo se relaciona esta respuesta con la pregunta del inciso anterior? [Tip R: Puedes utilizar el paquete `ggplot2` y la función `geom_smooth` para agregar a un *scatterplot* la recta del ajuste de MCO, para esto último debes indicarle a la función el método "lm"].

(d) (*7.5 puntos*) Describe brevemente (menos de 200 palabras) qué similitudes y diferencias tienen ambas estrategias.

(e) (*10 puntos*) Describe brevemente (menos de 200 palabras) qué diferencia teórica hubiera hecho utilizar la base BaseCOVIDp para contestar las preguntas. No busco que describan la diferencia en cuanto a los números y los resutlados de volver a

contestar los incisos anteriores, sino la diferencia conceptual de usar una u otra base de datos.

4. (*10 puntos*) Considerando que en el mundo hay un total de 195 países y que la base poblcional que se proporcionó (BaseCOVIDp) tiene información de únicamente 182, ¿qué países pueden estar subrepresentados en la base muestral?, ¿cómo afecta esto a la definición de nuestros parámetros poblacionales?, ¿cómo se vería reflejado esto en nuestro estimador y en el valor estimado muestral? Límite de palabras: 300.