Análisis estadístico del Covid-19 en México

José Manuel Ito Guzmán, Franco Quintanilla Fuentes, Ian Timothy Henry Suárez, Julio César López López.
A00227051, A00826953, A01701578, A01741427

I. INTRODUCCIÓN

La modelación estadística es una forma matemática y formalizada, de aproximarse a la realidad, y hacer predicciones con base en los resultados obtenidos. Además, las mejores decisiones se basan en el análisis objetivo de información confiable.

Para el desarrollo de este reto, realizaremos el análisis de la base de datos sobre los casos de Covid-19 en México, tomando en cuenta datos oficiales, reportados por la Secretaría de Salud del Gobierno de México en la página web del Gobierno de México, hasta el 29 de junio de 2020.

Para realizar el análisis de datos se utilizarán como herramienta, el lenguaje de programación R, el software en línea SALT y los tópicos de estadística y probabilidad vistos en el curso.

II. DESCRIPCIÓN DE LA PROBLEMÁTICA

Actualmente, se está sufriendo una pandemia a nivel mundial, ocasionada por el Covid-19. El Covid-19 es una enfermedad infectocontagiosa viral emergente con elevada mortalidad, que surgió en los últimos meses (Serra, 2020).

Al ser un virus emergente, que recientemente comenzó a afectar a los humanos, se desconocen las reacciones que pueden existir en diferentes grupos de personas. Es por ello que nuestro objetivo es utilizar la modelación estadística para comprender el comportamiento del virus e identificar qué factores son importantes a tomar en cuenta para tomar decisiones. En otras palabras, buscamos identificar a las poblaciones más vulnerables y los patrones que ha manifestado el Covid-19 en México.

Además, enfocaremos nuestra investigación en los riesgos en personas que padecen la comorbilidad del asma, ya que nos interesa saber si una enfermedad relacionada con la inflamación de las vías respiratorias (Álvarez, Álvarez, Fernández y Cal, 1995), puede agravar los riesgos y mortalidad por Covid-19 en las personas. Del mismo modo, determinaremos cuál grupo de la población de género resulta más propensa al contagio.

Cabe destacar que durante la presente investigación, se realizará un análisis estadístico para poder responder las hipótesis planteadas y poder llegar a una conclusión sobre la problemática.

III. PREGUNTAS DE INVESTIGACIÓN

1. Al analizar grandes bases de datos, es posible encontrar inconsistencias en ellas. ¿Existen valores extremos en la base de datos que se analizó para el desarrollo de este reto? ¿Son válidos para el análisis?

2. Distintos grupos de la población pueden presentar distintos síntomas y reacción al Covid-19 ¿Qué grupos de la población tienen mayor vulnerabilidad frente al Covid - 19?

3. El asma es una enfermedad pulmonar que se caracteriza por la inflamación de las vías respiratorias (Álvarez, Álvarez, Fernández y Cal, 1995), por lo cual es posible que la recuperación o fallecimiento de los pacientes con Covid-19 se encuentren relacionados con su padecimiento. ¿Qué porcentaje de las personas dentro de la base de datos que tiene asma falleció?¿Cómo se compara esto con el porcentaje total de personas fallecidas?

4. ¿Cual de los grupos de hombres o mujeres son más vulnerables a contagiarse por Covid-19?

IV. METODOLOGÍA UTILIZADA

Para el desarrollo de esta investigación se siguió una secuencia de fases para llevar en orden la resolución de los diferentes cuestionamientos. Tales fases se enlistan y explican a continuación:

Fase 1: Se exploró la base de datos y se realizó un resumen general. También se analizaron las distintas clasificaciones y sus resultados. El objetivo de esta fase fue familiarizarnos con la base de datos y obtener una noción más completa de los datos proporcionados por la Secretaría de Salud.

Fase 2: En esta fase se analizó la distribución de casos y muertes en los distintos grupos de edad y géneros. También se encontró una gran cantidad de muertes registradas no atribuidas al Covid-19, para realizar análisis posteriores, se decidió excluir este grupo de la población. También se excluyeron los datos relacionados a otras comorbilidades distintas al asma.

Fase 3: Para esta fase se realizó una estimación del porcentaje de contagiados que fallecían, con un intervalo de confianza del 95%. Después, se realizó una prueba de hipótesis con relación a la población asmática. Se buscó determinar si los datos sustentan la hipótesis de que el riesgo de fallecimiento aumenta si la persona padece de asma. Asimismo, se realizó otro análisis para determinar si el número de días promedio desde que se presentan los síntomas hasta que se fallece en personas que padecían de asma y Covid-19 era mayor a los que sólo padecían de Covid-19.

V. ANÁLISIS DE DATOS

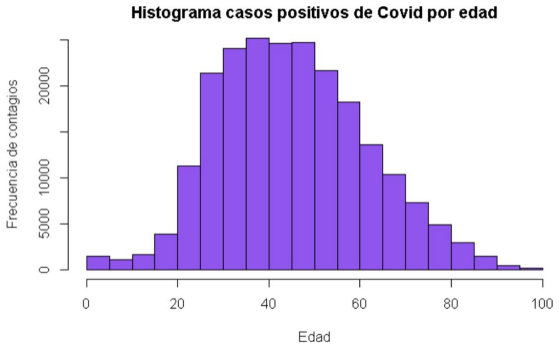
Para nuestro análisis de datos se utilizó el lenguaje de programación R. Con dicho programa se pudo analizar más de un millón de registros en la base de datos.

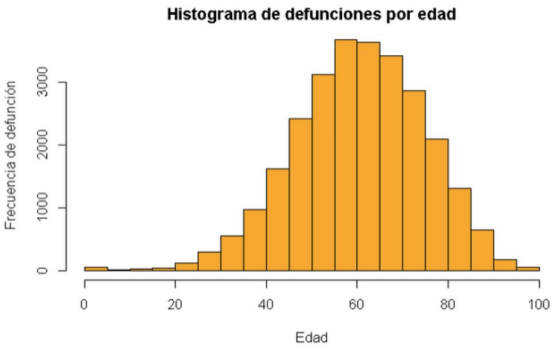
Nuestro análisis siguió el siguiente procedimiento. Primero, se revisaron las inconsistencias de la base de datos y se seleccionaron las variables que consideramos relevantes para nuestro análisis. Luego, se buscó el número total de contagiados y definimos los grupos vulnerables por sexo y edad. Después, se calculó un intervalo de confianza para determinar el porcentaje de personas contagiadas que fallece. El siguiente paso fue realizar una prueba de hipótesis para verificar si el asma es un factor que contribuya al riesgo de fallecimiento en las personas contagiadas. Finalmente, se buscó si existe una diferencia significativa entre el número de días que pasan desde que se presentan los síntomas hasta que se fallece en personas que padecen de asma y Covid-19, y en personas que padecen únicamente de Covid-19.

Después de analizar un poco la base de datos llegamos a varias conclusiones. Identificamos los valores nulos <NA>. Todos los valores nulos de la base se distribuyen en dos columnas; “DIAS.SINTOMAS.A.DE“ y “REVISAR“. Podemos eliminar la columna “REVISAR“, ya que no cuenta con ningún registro. Por otro lado, la variable “DIAS.SINTOMAS.A.DE“ es importante para nuestro análisis. Los valores nulos representan a todos los registros de la base de datos que no fallecieron por la enfermedad, es decir, solo tiene valores reales para los registros en los que alguien fallece.

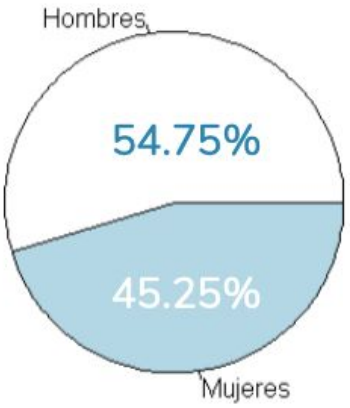
Para analizar los grupos vulnerables por edad, se descartaron valores extremos. Consideramos únicamente a las personas que tenían menos de 100 años. Se analizó la cantidad de personas contagiadas por edad y la cantidad de defunciones por edad para identificar a los grupos más vulnerables.

Después de esto, se obtuvieron las siguientes distribuciones; el primer histograma nos muestra la concentración de contagios por edad, el segundo histograma nos muestra la concentración de fallecimientos por edad.





Además, se obtuvo la distribución respecto al género en los casos de personas contagiadas. Se obtuvo que el 54.75% fueron hombres, y el 45.25% fueron mujeres, esto se puede ver a continuación en una gráfica de pastel.



Para calcular el porcentaje de personas contagiadas de Covid-19 que fallece, necesitamos obtener un intervalo de confianza, en nuestro caso establecimos un intervalo del 95%. Primero, debemos obtener una estimación con base en los datos. Como tenemos una muestra muy grande, podemos usar el Teorema del Límite Central y usar un estadístico que tenga una distribución normal estándar para calcular el margen de error.

$$\hat{p} \pm z_{\alpha} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Nos percatamos de que 220,657 individuos dieron positivo en la prueba de Covid-19, de los cuales fallecieron 27,121. Esto nos da una estimación del 12.29% de las personas contagiadas que fallecen. Después de hacer el análisis estadístico, podemos decir con una certeza del 95% que el porcentaje de fallecimientos en las personas contagiadas de Covid-19 se encuentra entre 12.15% y 12.43%. Este intervalo de confianza tiene un margen de error de aproximadamente 0.14%.

Para determinar si el asma constituye un factor de riesgo en el porcentaje de defunciones, realizamos una prueba de hipótesis. Nuestra hipótesis alternativa consiste en que el porcentaje de personas que fallece por Covid-19 que tiene asma es mayor al porcentaje de personas que fallece que tiene únicamente Covid-19. Para nuestro análisis tomamos en consideración únicamente dos poblaciones, los individuos que sólo padecían de asma y Covid-19 y los individuos que sólo padecían de Covid-19. Eliminamos los registros que contaban con otras comorbilidades. De esta forma eliminamos cualquier interferencia que pueda modificar los resultados de nuestro análisis. A continuación se plantean nuestras hipótesis.

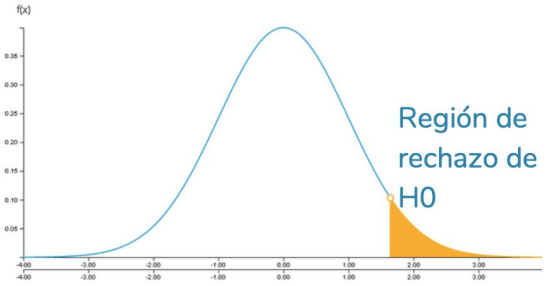
$$H_0 : p_1 \leq p_2$$
$$H_1 : p_1 > p_2$$

Donde p1 representa la proporción de fallecidos que padecían de asma y Covid-19 y p2 la proporción de fallecidos que sólo padecían de Covid-19. Utilizamos el siguiente estadístico Z para realizar nuestra prueba de hipótesis. Como es una muestra muy grande, podemos decir que Z tiene una distribución normal estándar.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} - \frac{p_2(1-p_2)}{n_2}}}$$

Los valores calculados $\hat{p}_1$ y $\hat{p}_2$ fueron 0.56% y 0.86% respectivamente. Consisten en la proporción de fallecidos en cada una de nuestras poblaciones.

Establecimos un nivel de significancia del 95%. Para poder rechazar la hipótesis nula, nuestra Z calculada debe ser mayor a la Z en alpha. Se puede observar la región de rechazo en la siguiente imágen.



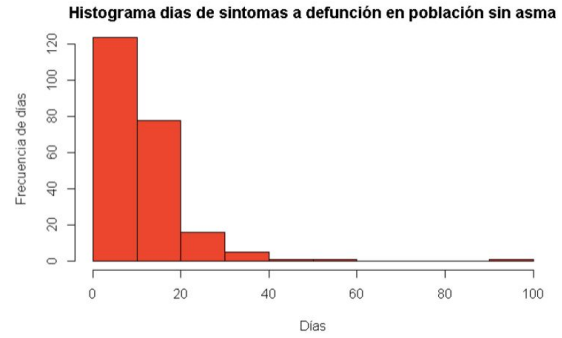
Obtuvimos los siguientes resultados.

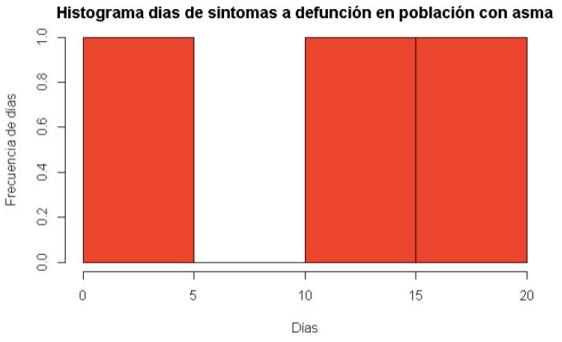
Z cal = -0.724
Z alpha = 1.644

Como la Z que calculamos con base en los datos fue menor que la Z en alpha, podemos decir que no hay evidencia suficiente para rechazar nuestra hipótesis nula, es decir, no hay evidencia suficiente para sostener la hipótesis de que la probabilidad de fallecer por Covid-19 es mayor si se padece la comorbilidad de asma.

Para verificar si hay una diferencia significativa entre el número de días que pasan desde que se presentan los síntomas hasta que se fallece en personas que padecen de asma y Covid-19 y en personas que padecen únicamente de Covid-19, tuvimos que realizar una segunda prueba de hipótesis. Sin embargo, nos percatamos de que no se puede realizar un análisis confiable, puesto que solamente tres personas con asma fallecieron por Covid-19. Con esta cantidad de datos no se puede justificar el Teorema del Límite Central.

A continuación se muestra la concentración de registros que hay para el número de días que pasan desde que se presentan los síntomas hasta que se fallece en nuestras dos poblaciones.





VI. CONCLUSIONES

La modelación estadística nos permitió realizar un análisis relevante con respecto a los factores de riesgo relacionados con el Covid-19. Dentro de este análisis, pudimos identificar las poblaciones que se veían más afectadas por este padecimiento, así como aproximarnos a una valoración correcta de los riesgos de los pacientes que poseían la comorbilidad del asma.

Es importante destacar que nuestra base de datos solo corresponde a una muestra muy grande de la población que presentó síntomas de Covid-19 en el país, por lo cual los resultados obtenidos corresponden a estimaciones estadísticas.

De esta manera, se encontró que los grupos poblacionales de edad más vulnerables, donde se encontró un alto nivel de defunciones, fue en personas de entre 50 y 70 años.

Por otro lado, también se encontró que, de acuerdo a la base de datos proporcionada, así como al análisis estadístico realizado, y debido a la escasez de datos

relacionados con nuestra investigación, no se encontraron suficientes evidencias para determinar si el asma como única comorbilidad, resultó o no resultó ser un factor importante en la letalidad y padecimiento del Covid-19.

VII. REFERENCIAS

Alvarez, R., Alvarez, R., Fernández, E., y Cal, F. (1995). Mediadores inflamatorios en el asma. Revista Cubana de Medicina General Integral, 11(2), 168-170. Recuperado de: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21251995000200011

DeVore, J. L. (2015). Probability and Statistics for Engineering and the Sciences (9th Revised ed.). Cengage Learning.

R Studio (4.0.2). (2020). [Computación estadística y gráficos]. https://rstudio.com/

SALT: Statistical Analysis and Learning Tool [**Software**] (2020). Recuperado de https://www.webassign.net/csalt/#/toolset/distributions

SALUD. (2020). Información referente a casos COVID-19 en México. Recuperado el 20 de octubre, de 2020, de SALUD. Sitio web: https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico

Serra, M. (2020). Las enfermedades crónicas no transmisibles y la pandemia por COVID-19. Recuperado el 20 de octubre, de 2020, de Hospital General Docente Enrique Cabrera. Sitio web: http://scielo.sld.cu/scielo.php?pid=S2221-24342020000200078&script=sci_arttext&tlng=pt