# Análisis de Supervivencia de Covid-19 en la población colombiana.

Camilo Oberndorfer Mejía
coberndorm@eafit.edu.co

Miguel Valencia Ochoa
mvalenciao@eafit.edu.co

Abelino Sepúlveda Estrada
asepulvede@eafit.edu.co

Pedro Botero Aristizábal
peboteroa@eafit.edu.co

3 de diciembre de 2020

**Resumen.** El análisis de supervivencia es una herramienta de análisis de datos utilizada fundamentalmente para estudios e investigaciones epidemiológicas, cuyo objetivo es determinar la supervivencia de ciertas poblaciones bajo un determinado tiempo. En este trabajo se mencionarán conceptos y algunas técnicas utilizadas en análisis de supervivencia para así determinar la probabilidad de supervivencia de la población colombia por Covid-19.

## Índice

# 1. Introducción.

La pandemia causada por el SARS-CoV-2, es sin duda la mayor problemática que el área de la salud se ha enfrentado en el ultimo siglo. El nivel al que el mundo se ha globalizado es la principal razón por la que este virus se ha logrado expandir de tal manera.

Esto ha traido una nueva serie de problemas dado que al ser cada país indepiendente, con su propia cultura, sistema médico y propias políticas. El virus ha tenido interacciones radicalmente diferentes dependiendo de su localización en el mundo. Por lo que no es suficiente con que se analice su comportamiento en solo un estado, sino que si se quieren tomar las medidas apropiadas para disminuir el impacto de esta pandemia, se debe conocer más sobre el comportamiento del virus en su región. Con esto en mente, se decidió hacer el análisis de supervivencia por covid-19 en la población colombiana, tomando en cuenta factores de riesgo, llegando a calcular cuáles son las variables que más afectan la supervivencia de los pacientes con la enfermedad.

# 2. Datos utilizados.

Los datos utilizados para nuestro trabajo fueron proporcionados por el portal web colombiano de Salud y Protección social. El Dataset cuenta con la información de la población colombiana contagiada por el virus, donde, se cuenta con al menos 1.316.806 reportes. Donde cada reporte cuenta con 25 características, entre las cuales podemos evidenciar fechas como la de inicio de síntomas, de diágnostico, de muerte por el virus y de recuperación, y otras características como sexo, edad y municipio de procedencia.

# 3. Implementación.

## 3.1. Análisis de Supervivencia.

El análisis de supervivencia es una herramienta imprescindible para estudios e investigaciones tanto clínicas como epidemiológicas. Ahora bien, el Análisis de Supervivencia es una herramienta de análisis de datos basada en analizar los estados de cambio de cierta poblacion frente al tiempo. En nuestro caso, será utilizada para el análisis de las muertes de las personas contagiadas en un tiempo determinado. [2]

Para entender un poco más sobre este tema, veamos los siguientes términos:

### 3.1.1. Función de Supervivencia.

Esta función denotada por S(t), es utilizada para para calcular la propabilidad de que una persona sobreviva en un tiempo t. Sea F(t) su función de distribución y T una variable aleatoria definida en el intervalo [0,$\infty$) entonces $S(t)$ está dada por:

$$S(t) = P(T > t) = 1 - F(t)$$

$$S(t) = \int_t^{\infty} f(x)\, dx = 1 - F(t)$$

Donde la función de supervivencia es monótona decreciente, es decir,

$$S(u) \leq S(t)$$

Si $u > t$. Además,

$$S(t) \leq 1$$

Cuando $t = 0$ la probabilidad de supervencia es 1, es decir, $S(0) = 1$ [2]

### 3.1.2. Estimador de Kaplan-Meier

Es otro modelo utilizado para el cálculo de la probabilidad de supervivencia. Es un estimador no paramétrico, que admite una representación gráfica. Supongamos que se cuentan con N observaciones de tiempos distintos, es decir,

$$t_1 < t_2 < ... < t_n$$

La formulación de este estimador está dada por:

$$S(t) = P(T > t)$$

$$S(t) = \prod_{t_i < t} \frac{n_1 - d_i}{n_i}$$

Donde $S(t)$ es la probabilidad de estar vivo en el tiempo $t_i$, $n_i$ es la población en riesgo en $t_i$ y $d_i$ representa el número de muertes en $t_i$. También podemos escribirla de una manera simplificada como:

$$S(t_i) = S(t_i - 1)(1 - \frac{d_i}{n_i})$$

$S(t_i - 1)$ es la probabilidad de supervivencia en $t_i - 1$ [5]

#### 3.1.3. Riesgo proporcional de Cox.

El objetivo principal de este método es notar cómo diferentes factores en un conjunto de datos impactan en el evento de interés. Donde la función $h(t)$ está dada por:

$$h(t) = h_0(t)exp(b_1x_1, ..., b_nx_n)$$

Donde, $h(t)$ es la función de hazard, $x_1, ..., x_n$ son las covariables, $b_1, ..., b_n$ son los factores que impactan las covariables. [1]

#### 3.1.4. Hazart Ratio

Es el riesgo relativo de que ocurra una muerte. Denatada por $HT$ de la siguiente manera:

$$HT = exp(b_i)$$

Donde $b_i$ es el coeficiente de qué tanto afecta la covariable [1]

# 4. Metodologia

Ya teniendo conocimiento de lo que es el análisis de supervivencia, la función de supervivencia y los métodos con los que se puede calcular y las censuras, ya podemos empezar con la aplicación de éstos con nuestro Dataset y empezar a calcular la probabilidad de supervivencia de la población colombiana por Covid-19.

## 4.1. Reestructuración del Dataset

Como se había mencionado anteriormente, contamos con un Dataset de la población colombiana contagiada por el virus, en la cual cada caso reportado cuenta con 25 características, pero determinamos que muchas de estas características no eran relevantes para el análisis, así que reestructuramos el Dataset y nos quedamos con (12) características, como las que son la fecha de inicio de síntomas, fecha de de diagnóstico, sexo, ciudad de procedencia y entre otras.

### 4.2. Censuras en el Dataset

Al analizar los datos, se eligieron según el criterio de que para hacer el análisis, se debía tener si las personas ya se habian recuperado o habían fallecido. Dado a que nos encontramos con unos casos que seguían activos después de la finalización del estudio, se censuraron, y en su total se llegaron a censurar 96 personas.

## 5. Implementación en Phyton

Para hacer el análisis de supervivencia, se hizo la implementación en Python. La cual se basó principalmente en el uso de las librerías Pandas y Lifelines. Con la primera, se realizó la lectura de los datos para posteriormente con lifelines analizar cada caso de acuerdo al tiempo y al estado del paciente.

### 5.1. Variables

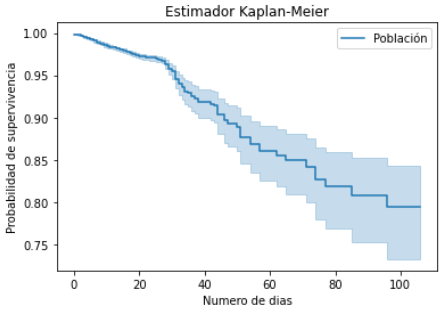
Para hacer el análisis, se estudiaron las características de los casos, es decir, se analizaron las poblaciones según su edad, departamento y sexo, en las cuales se evalúa la correlación entre éstas y las muertes. Además, se quizo hacer el análisis de la población en general, en el cual, se incluían todas las edades, ambos sexos y todos los departamentos de las poblaciones.

## 6. Resultados

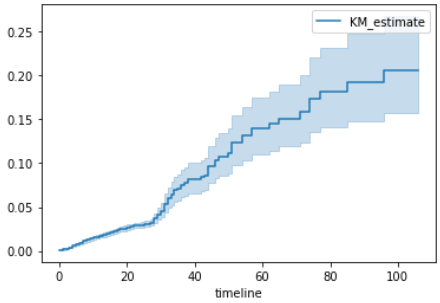
Después de haber hecho la implementación del análisis en Python y Excel se obtuvieron los siguentes resultados:

### 6.1. Resultados de Kaplan-Meier

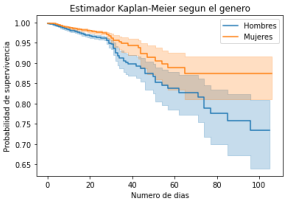
En la implementación de Python se le hizo el análisis de las personas contagiadas el 24/07/2020 donde se contaba con un población 12.489 personas. Los resultados se ven el las siguientes gráficas:



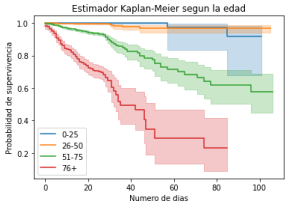

Gráfica 1. Función de Supevivencia



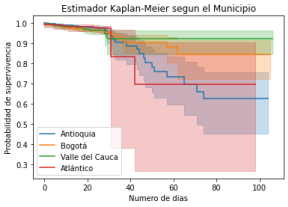

Gráfica 2. Función de Riesgo




Gráfica 3. Análisis de supervivencia por Género




Gráfica 4. Análisis de supervivencia por Edades




Gráfica 5. Análisis de supervivencia por Departamentos

### 6.1.1. Análisis de los resultados Kapla-Meier

Como se puede observar en la Gráfica 1, obtuvimos que la probabilidad de que una persona con covid siga viva para el dia 110 despues de la aparición de sus síntomas es de alrededor de un 80 %, con un posible error del 7 %.

En la Gráfica 2, donde se compara la supervivencia entre hombres y mujeres, se llega a la conclusión de que en la mayoria de los casos posibles (incluyendo el caso promedio), las mujeres tienen una mayor probabilidad de supervivencia con el tiempo, y aunque la probabilidad de supervivencia hasta el primer mes es similar entre hombres y mujeres, de este punto en adelante se aumenta la diferencia entre ambas hasta casi un 15 % en el caso promedio.

En las Gráficas 3 y 4. Se compara la supervivencia por edades y la supervivencia por Departamento respectivamente, en la primera, se ve cómo la edad es un factor que afecta considerablemente la supervivencia con respecto al tiempo desde el dia uno hasta los últimos días considerados en el estudio, mostrando cómo a mayor edad, menor probabilidad de supervivencia, habiendo una menor diferencia entre los jóvenes y jóvenes adultos que entre los adultos y los adultos mayores. En la Gráfica 4 se muestra la probabilidad de supervivencia según el departamento que se habita, en este caso, Atlántico, al tener tan pocos pacientes dentro de esta población, el rango de error es muy alto, por lo que es posible que este tenga una alta probabilidad de supervivencia desde despues del dia 30, como puede tener la peor probabilidad de supervivencia.

## 6.2. Resultados de Riesgo proporcional de Cox

| | coef | exp(coef) | se(coef) | coef lower 95% | coef upper 95% | exp(coef) lower 95% | exp(coef) upper 95% | z | p | -log2(p) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Sex** | -0.39 | 0.68 | 0.1 | -0.58 | -0.2 | 0.56 | 0.82 | -3.99 | <0.005 | 13.9 |
| **Edad** | 0.09 | 1.09 | 0 | 0.08 | 0.09 | 1.08 | 1.1 | 30.67 | <0.005 | 683.72 |

Tabla 1. Regresión de Cox

| | coef | z-test | p-value |
|---|---|---|---|
| **Sex** | -3.99 | <0.005 |
| **Edad** | 30.67 | <0.005 |
| **Código DIVIPOLA departamento** | 0.52 | 0.6 |

Tabla 2. z-test y p-value

### 6.2.1. Análisis de resultados de Regresión de Cox

Al implementar la regresión de Cox, se notó que tanto la edad como el sexo, tenian alta incidencia sobre la función de supervivencia por lo que hace sentido tomarlos como casos aparte,(tomese que $p<0.05$) mientras que el departamento no afecta mucho la probabilidad de supervivencia ($p>0.05$). Además, también podemos observar que en el sexo y la edad $z > 2$

# 7. Conclusiones

Como se esperaba, la edad es el factor que mas influye en la supervivencia por el tiempo de las personas, que esto es un comportamiento común con muchas enfermedades similares. Lo que no se esperaba, y todavía no tiene razón médica, es que los hombres son mas propensos a morir por SARS-CoV-2 en un plazo de 110 días después de la aparición de los sintomas. Tampoco se esperaba que los municipios afectaran tan poco a la supervivencia.

Es importante mencionar que en este trabajo no se buscó encontrar la distribucion que siguen las funciones de supervivencia, por tanto, no se pudo hallar la ecuación de la función de riesgo.

Nuestro informe fue basado en [3] y [4]

# 8. Referencias.

## Referencias


[1] Jadwiga Borucka. "Methods of handling tied events in the Cox proportional hazard model". En: *studia oeconomica posnaniensia* 2.2 (2014), págs. 91-106.

[2] Humberto Gutiérrez Pulido, Vara Salazar y col. "Control estadıstico de calidad y seis sigma/Humberto Gutiérrez pulido, coautor Román de la Vara Salazar." En: (2004).

[3] Guillermo Salinas-Escudero y col. "A survival analysis of COVID-19 in the Mexican population". En: *BMC public health* 20.1 (2020), págs. 1-8.

[4] Pratik Shukla. *Survival Analysis with Python Tutorial — How, What, When, and Why.* urlhttps://medium.com/towards-artificial-intelligence/survival-analysis-with-python-tutorial-how-what-when-and-why-19a5cfb3c312. Accedido 03-12-2020. 2020.

[5] Freddy Tineo. "Estimación de Kaplan Meier bootstrap de la Curva de Supervivencia". En: *Monografıa. Universidad Nacional Mayor de San Marcos, FCM* (2005).