Modelación Estadística MA1001B.4

ANÁLISIS DE DATOS DE COVID-19 EN MÉXICO

Diego Perez Hernandez A00572893
Miguel Ángel Chávez Robles A01620402
Camila Cusicanqui Padilla A00571258
Nancy Lesly Segura Cuanalo A01734337

INTRODUCCIÓN

SARS-CoV-2

El covid-19 es un virus que en este año ha causado más de un millón de muertes. Durante este reto se desarrolló un análisis de la base de datos del gobierno de México que contiene información detallada sobre el virus en el país. Se realizó con el fin de conocer más acerca de su comportamiento y así generar conclusiones que nos ayuden a combatir la problemática.

Tabaquismo

El Covid-19 es una infección que ataca directamente a los pulmones. De acuerdo a la OMS las personas con problemas de tabaquismo tienen más probabilidades de desarrollar síntomas más graves en caso de padecer covid que los que no fuman, pues sus pulmones están muy débiles. Esto lo comprobaremos realizando el análisis de esta comorbilidad.

ETAPA 1 Exploración de la base de datos e identificación de las variables

Elección de programa para manipulación de datos Python **ETAPA 2**

ETAPA 3 Limpieza de datos extremos (Datos alejados a más de 3 desviaciones estándar de la media)

Graficación de las variables de interés. **ETAPA 4**

ETAPA 5 Propuestas de distribuciones

Pruebas de bondad de ajuste **ETAPA 6**

ETAPA 7 Intervalos de confianza

Conclusiones finales **ETAPA 8**

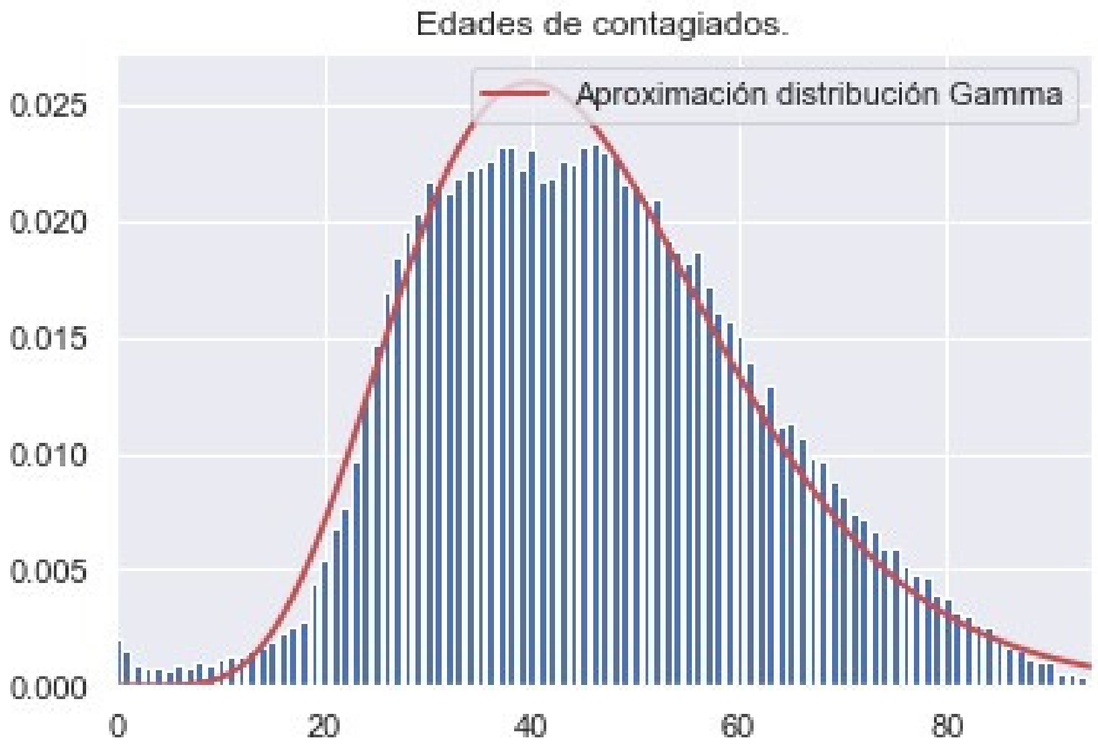
GRÁFICA 1.1 Edades de contagiados

CONTAGIOS POR EDAD

En esta gráfica se observan las personas contagiadas de acuerdo a sus edades.

DISTRIBUCIÓN GAMMA

ESTIMACIÓN DE PARÁMETROS A PARTIR DEL MÉTODO DE MOMENTOS



GRÁFICA 1.2 Edades de contagiados con ajuste de distribución gamma

Para estimar los parámetros de esta distribución se utilizó el método de momentos, donde se realizó un sistema de ecuaciones entre los momentos muestrales y los de la distribución.

ALFA

$$\hat{\alpha} = \frac{\bar{X}^2}{\frac{\sum X_i^2}{n} - \bar{X}^2} = 5.772$$

BETA

$$\hat{\beta} = \frac{\frac{\sum X_i^2}{n} - \bar{X}^2}{\bar{X}} = 7.8891$$

PRUEBA: BONDAD DE AJUSTE

Distribución de Probabilidad Gamma

Hipotesis Nula: La variable aleatoria, número de contagios por edad, tiene una distribución gamma con alfa =5.772, beta=7.8891

Tenemos dos parámetros de estimación, alfa y beta, debemos restar 1 grado por parámetro a los grados de libertad.
La cantidad de intervalos es k=89 entonces gl=k-2-1=86. Ahora, podemos evaluar la Chi cuadrada de la muestra y comparar la Chi cuadrada del valor critico dado por alfa.

$$X^2 = \sum_{i=1}^{k} \frac{(FO_i - FE_i)^2}{FE_i}$$

Hipotesis Alternativa: La variable aleatoria tiene una distribución de probabilidad diferente.

Alfa es 5%

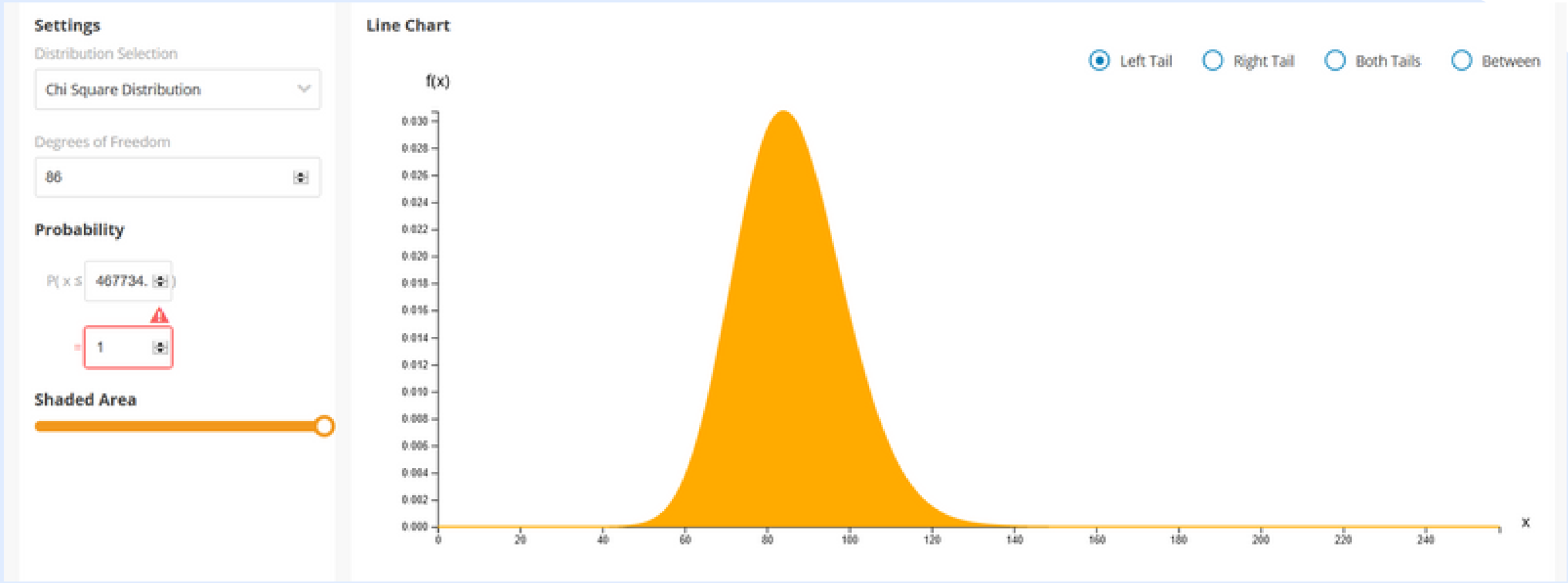
Zona de rechazo:

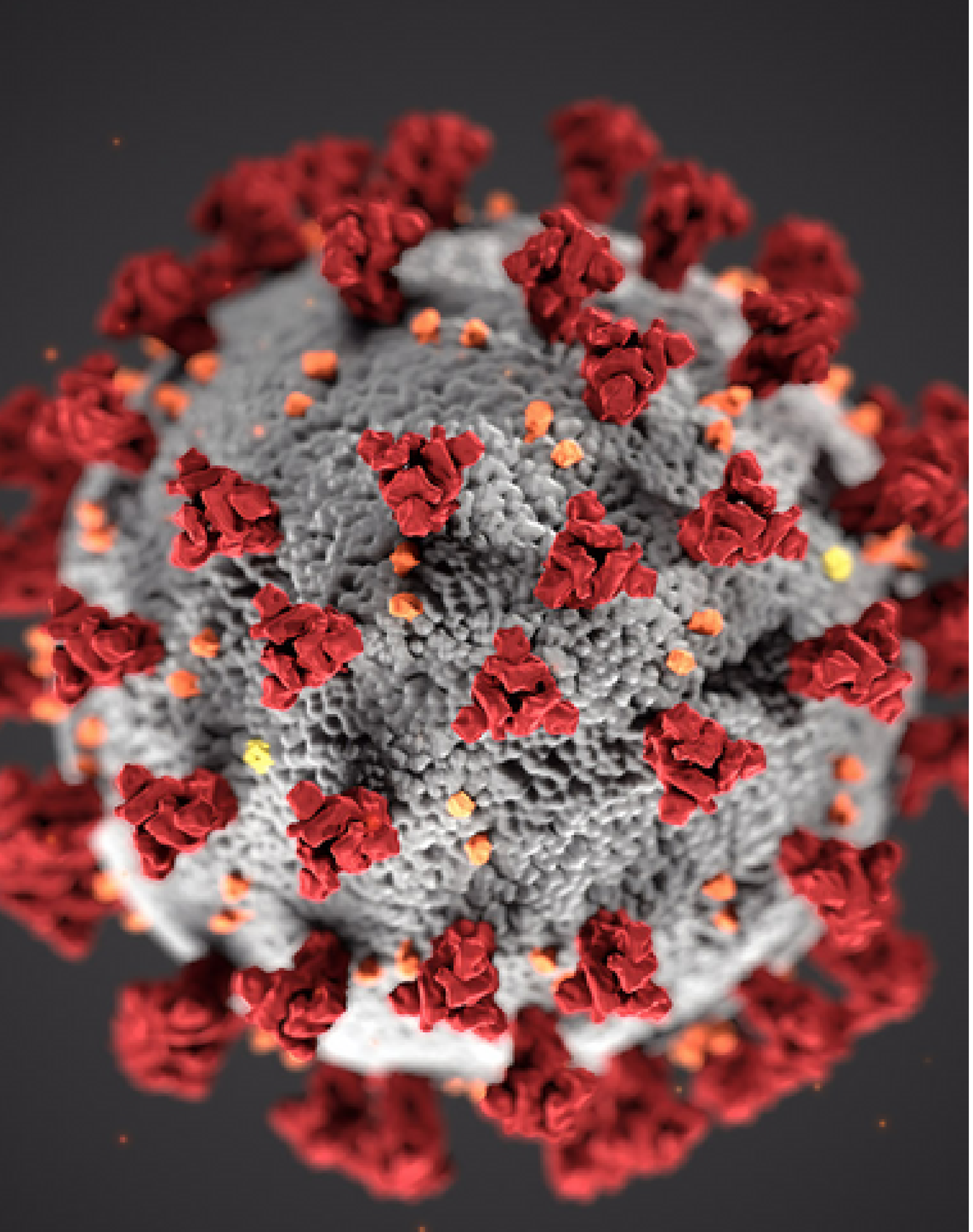
$$P(X^2 > 108.6479)$$

Zona de no rechazo:

$$P(X^2 \leq 108.6479)$$

EL VALOR DE LA CHI CUADRADA DE LA MUESTRA





Valor P

El valor p tiene como relación la fiabilidad del estudio, cuyo resultado será más fiable cuanto menor sea la p. Es la probabilidad de obtener un valor semejante si se realiza el experimento en las mismas condiciones[1]

$$p = 1 - P(X^2 \leq 467734.03944) = \quad 0$$

$Valor\ p > \alpha \rightarrow No\ se\ tiene\ evidencia\ para\ rechazar\ H_o$
$Valor\ p \leq \alpha \rightarrow Se\ tiene\ evidencia\ para\ rechazar\ H_o$

$$0 \leq \alpha$$

Se rechaza H0, se concluye que la distribución Gamma no es adecuada con un 95% de confianza

CONCLUSIÓN

Dado que la Chi cuadrada de la muestra es es mayor al Chi cuadrada del valor crítico, hay suficiente evidencia muestral para rechazar la Ho; podemos inferir que la variable aleatoria, el número de contagios por edad, tiene una distribución diferente y un valor p de 0 con 95% de confianza.

FALLECIDOS POR EDAD

Aquí se presenta una gráfica de los fallecimientos de acuerdo a las edades.

DISTRIBUCIÓN NORMAL

GRÁFICA

ESTIMACIÓN DE PARÁMETROS

Para estimar los parámetros de esta distribución se utilizaron los estimadores insesgados de miu y de sigma

MU

$$\mu = \bar{X} = 61.30$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

SIGMA

$$\sigma = 14.4$$

PRUEBA: BONDAD DE AJUSTE

Distribución de Probabilidad Normal

Hipotesis Nula: La variable aleatoria, número de días entre los síntomas y defunciones, tiene una distribución NORMAL con mu=61.30 y sigma=14.4

Hipotesis Alternativa: La variable aleatoria tiene una distribución de probabilidad diferente a la normal

Alfa es 5%

Grados de libertad, gl=k-2-1= 86
Ahora, podemos evaluar la Chi cuadrada de la muestra y comparar la Chi cuadrada del valor critico dado por los parámetros establecidos.

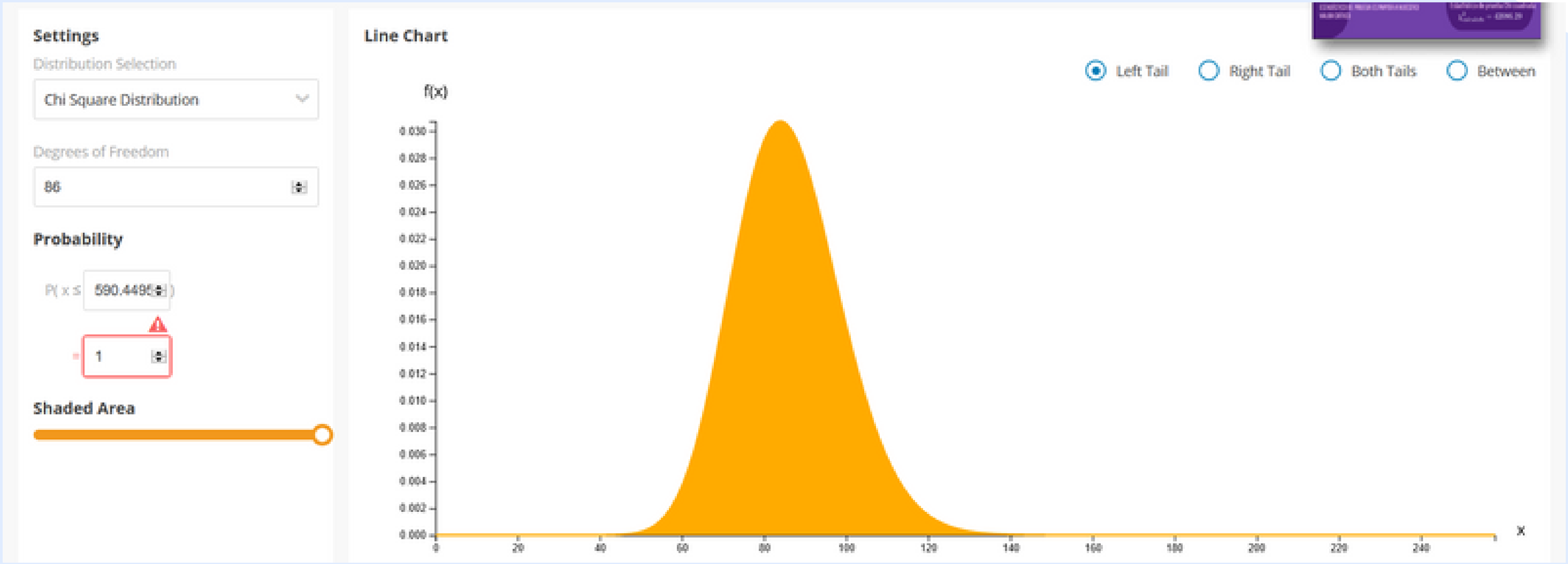
Zona de rechazo:

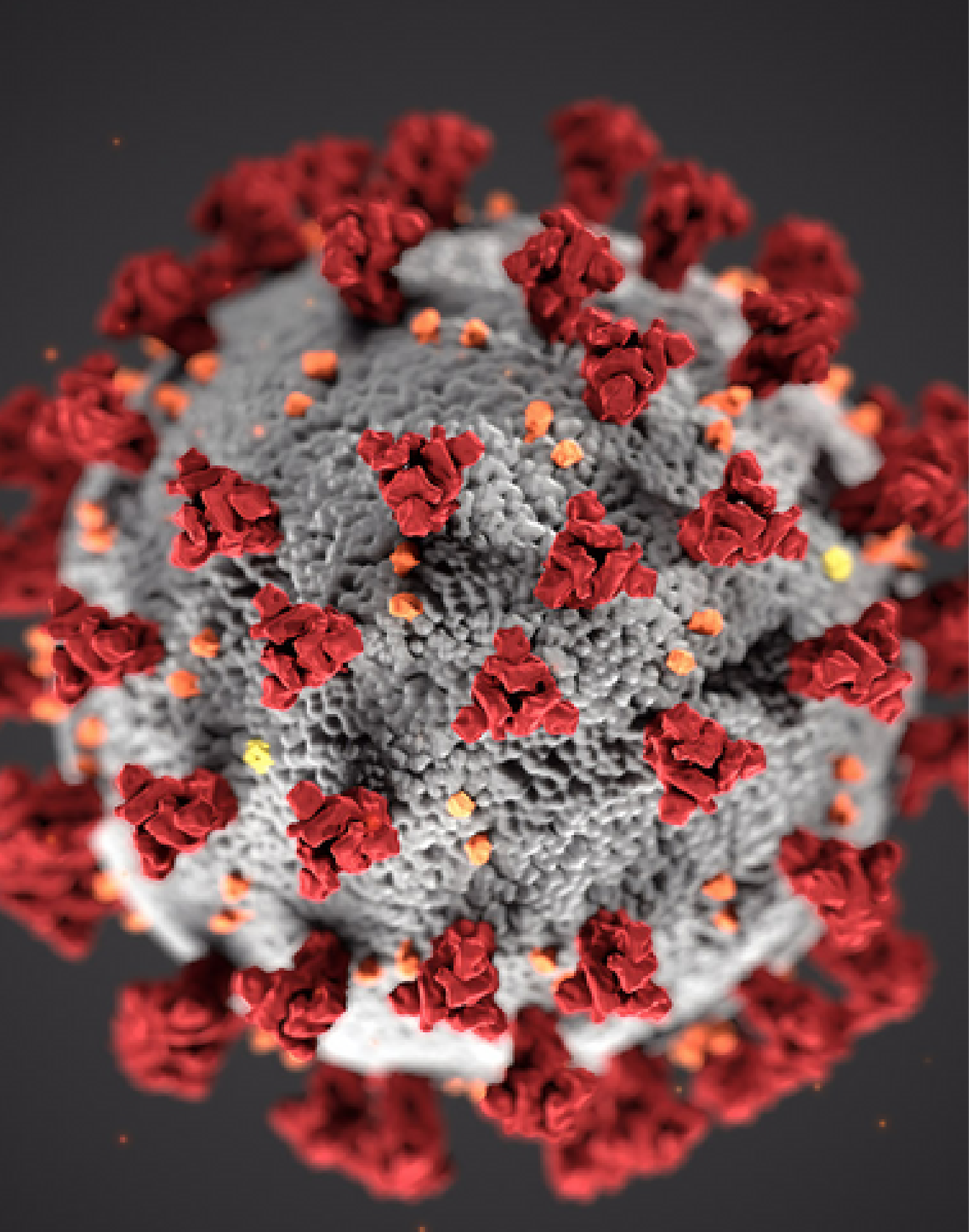
$$P(X^2 > 108.6479)$$

Zona de no rechazo:

$$P(X^2 \leq 108.6479)$$

Chi cuadrada de la muestra con gl=89 poner valor chi de muestra





Valor P

Sacamos el valor p como previamente mencionado

$$p = P(X^2 \geq 590.4495) = 1 - P(X^2 \leq 590.4495) = 0$$

$Valor\ p > \alpha \rightarrow No\ se\ tiene\ evidencia\ para\ rechazar\ H_o$
$Valor\ p \leq \alpha \rightarrow Se\ tiene\ evidencia\ para\ rechazar\ H_o$

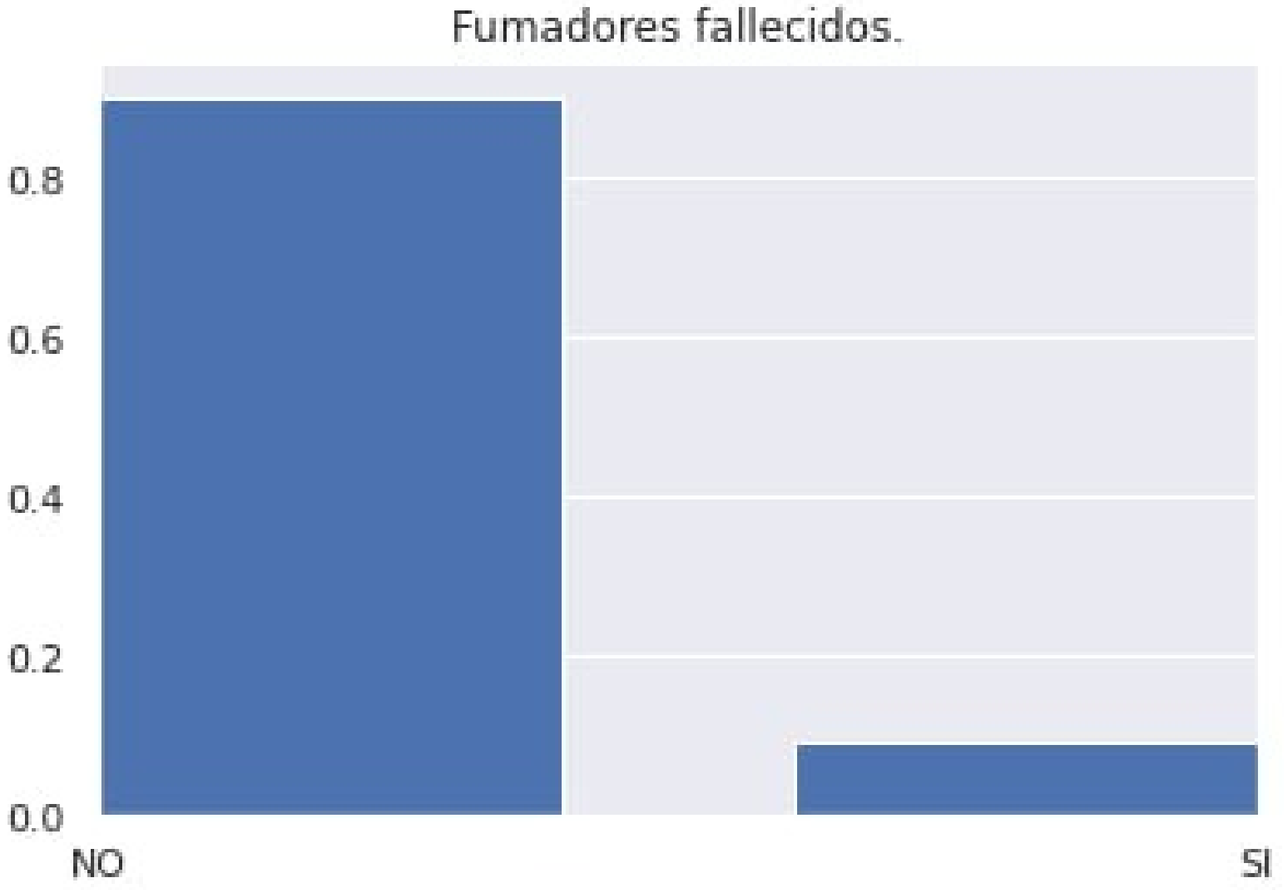
$$0 \leq \alpha$$

Hay suficiente evidencia estadística para rechazar H0

CONCLUSIÓN

Dado que la Chi cuadrada de la muestra es mayor al Chi cuadrada del valor crítico, hay suficiente evidencia muestral para rechazar la Ho; podemos inferir que la variable aleatoria, el número de contagios por edad, tiene una distribución DIFERENTE y su valor p es 0 con 95% de confianza.

FALLECIDOS Y TABAQUISMO



DISTRIBUCIÓN BINOMIAL

Se calculó la proporción de defunciones con tabaquismo en la gráfica al lado.

$$\hat{p} = 0.09229876$$

Estimación del parámetro

Se propone una distribución binomial dado que solo existe 2 resultados posibles.

$$\hat{p} = \frac{X}{n} = \frac{3339\ fumadores\ muertos}{36176\ defunciones}$$

X ES UN VARIABLE ALEATORIA QUE INDICA EL NUMERO DE POSIBLES EXITOS

N=DATOS TOTALES

$$\hat{p} = 0.09229876$$

$$\hat{q} = 0.9077012$$

INTERVALO DE CONFIANZA

Es un tipo de estimado hecho del dato observado, propone un rango de valores para la proporción poblacional, el parámetro verdadero.

$$P\left(\hat{p} - Z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}\hat{q}}}{n} < p < \hat{p} + Z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}\hat{q}}}{n}\right) = 1 - \alpha$$

$$P(0.08931602 < p < 0.0952815) = 0.95$$

PARA ENCONTRAR EL LIMITE SUPERIOR E INFERIOR

$$\hat{p} + Z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}\hat{q}}}{n} = 0.0952815$$

$$\hat{p} - Z_{\frac{\alpha}{2}}\frac{\sqrt{\hat{p}\hat{q}}}{n} = 0.08931602$$

CONCLUSIÓN

Se estima que la proporción, p, de fumadores fallecidos por defunciones podría encontrarse entre 0.08931602 y 0.0952815 con 95% de confianza. Dentro de este intervalo de confianza no se podría decir que la distribución de la probabilidad es homogénea ya que es menor a 0.5.

CONCLUSIÓN GENERAL

Se rechazó la hipotesis nula del número de contagios por edad y el número de días entre sintomas y defunciones con un valor p de 0 en cada caso. Calculamos los limites del intervalo de confianza con 95% de confianza.

Se puede observar la importancia del análisis estadístico ya que juegan un rol fundamental en la investigación. Nos permite saber cuales heramientas son adecuadas para un estudio particular ya que mientras más conozcamos sobre un fenómeno mejor podremos afrontarlo.

REFERENCIAS:

[1] Molina Arias, M. (2017). ¿Qué significa realmente el valor de p?. Pediatría Atención Primaria, 19(76), 377-381. Recuperado en 22 de octubre de 2020, de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1139-76322017000500014&lng=es&tlng=es.

[2] Jay L. Devore. (2016). Probabilidad y Estadística para Ingeniería y Ciencias. CENGAGE Learning, Edición: 9

[3] Datos abiertos Dirección General de Epidemiología Secretaría de Salud, Gobierno de México : https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico Consultado el 30 de junio de 2020.