Alumnos: Mendoza Saenz de Buruaga Imanol
Villalon Pineda Luis Enrique

Ejercicio 1

Problema:

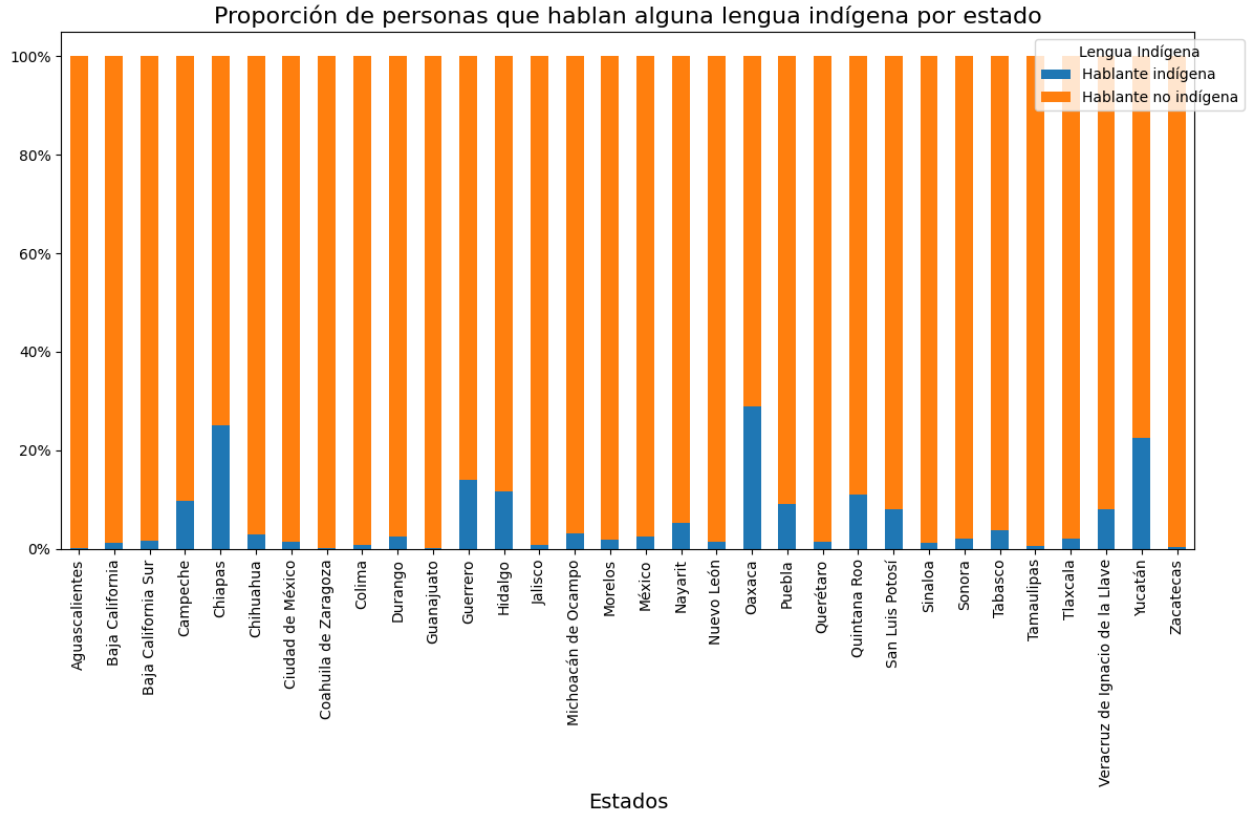
Censo de Población y Vivienda 2020. Cada 10 años, el Instituto Nacional de Estadística y Geografía (INEGI) levanta el Censo de Población y Vivienda con el objetivo de conocer diversas características de los habitantes de México y sus viviendas a nivel nacional, estatal, municipal, por localidad, por grupos de manzanas y hasta por manzana.

Solucion:

Tras extraer la base de datos y aplicar un preprocesamiento para seleccionar únicamente los registros y columnas relevantes, obtuvimos un conjunto de datos con dimensiones 32×7 y un tamaño de 0.0087490 MB.

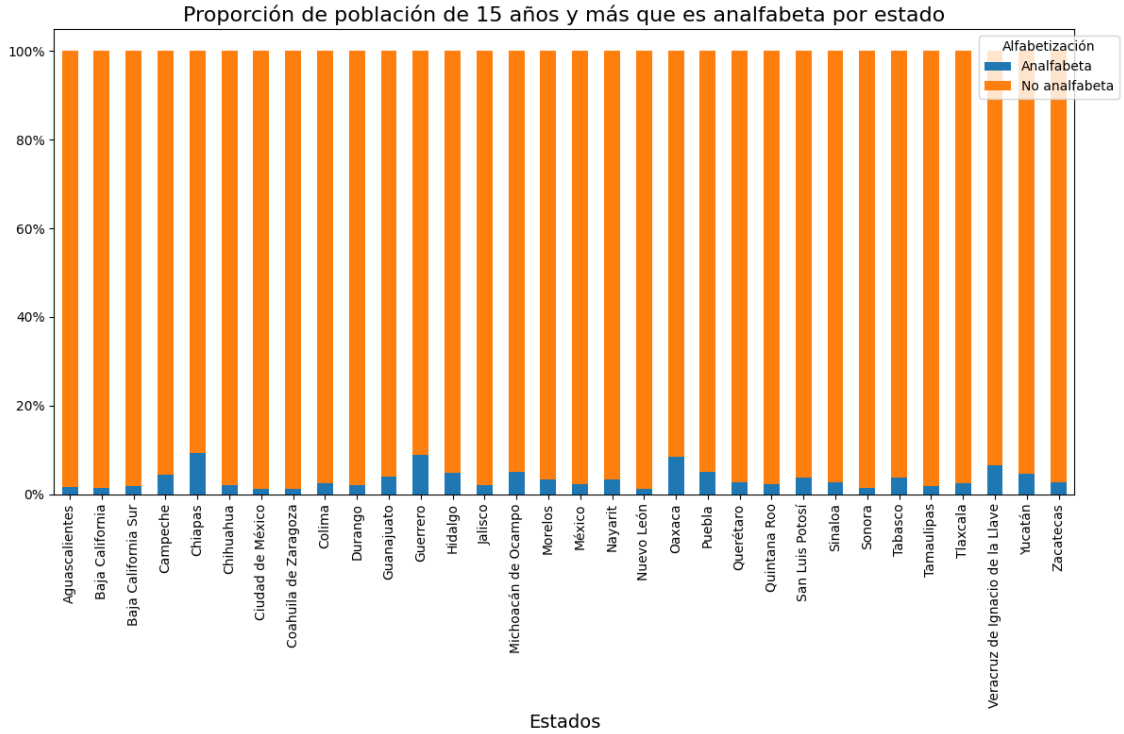
Al sumar las poblaciones de cada entidad, obtuvimos un total de 126,014,024 personas, coincidiendo con los datos reportados por el INEGI para el año 2020.

Incluir información sobre las localidades de cada municipio no sería recomendable, ya que haría los datos demasiado específicos, complicando el cálculo y manejo de las métricas en los ejercicios.

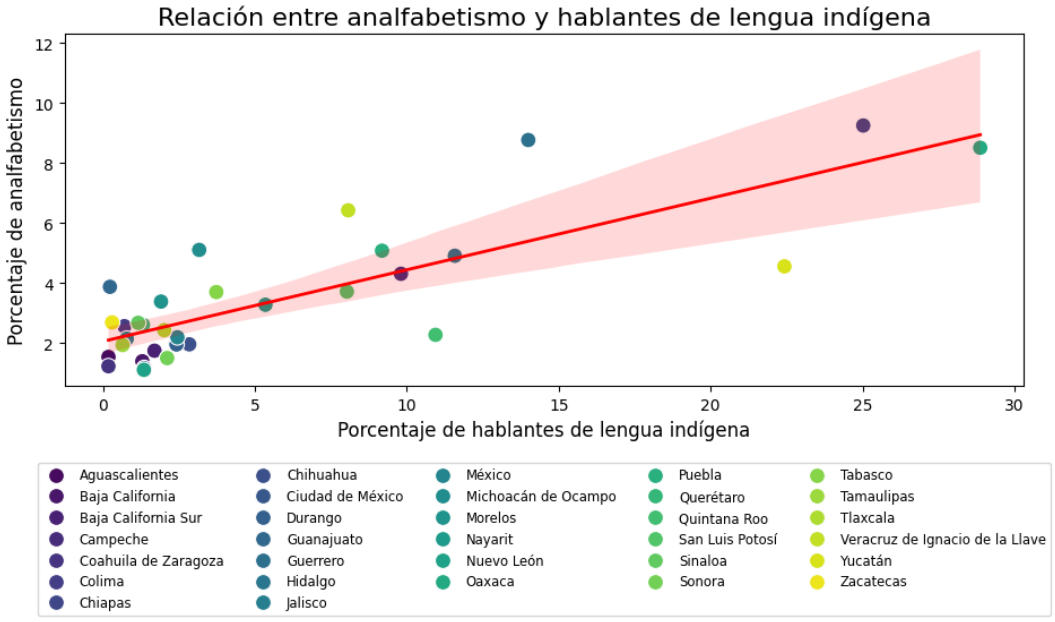


Los tres estados con más hablantes de lengua indígena con respecto a su población son Chiapas, Oaxaca y Yucatán. Estos estados tienen una gran diversidad cultural y lingüística debido a una presencia significativa de comunidades indígenas que mantienen vivas sus lenguas y tradiciones. Las políticas y esfuerzos para preservar y promover las lenguas indígenas por parte del estado han tenido cierto éxito. Aunque la preservación de las lenguas es

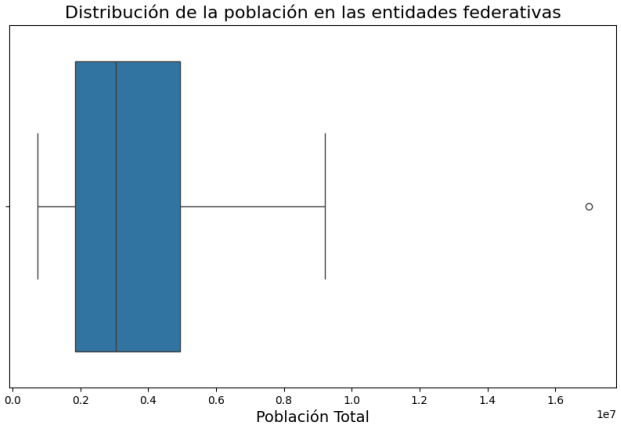
algo bueno, también puede estar asociada con desafíos socioeconómicos, ya que estas zonas montañosas de alguna manera se encuentran aisladas del resto de regiones en México

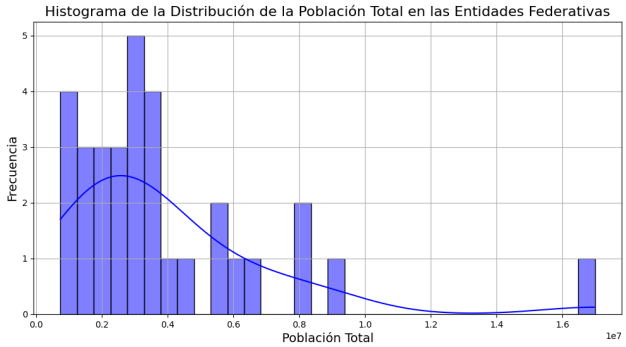


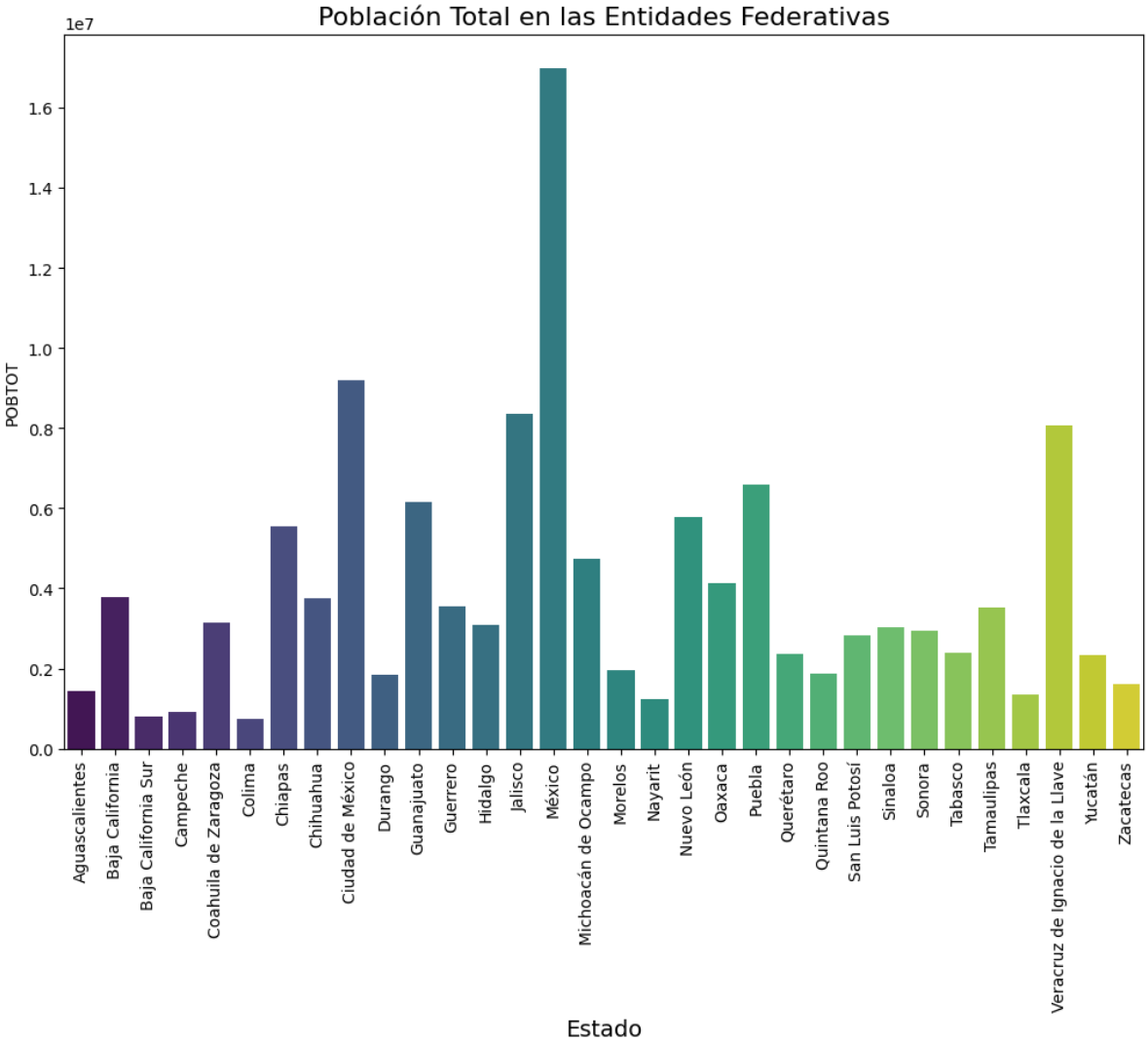
Aunque el analfabetismo en los estados de Chiapas, Guerrero y Oaxaca no superan el 10 % en su respectivo total de población, sigue siendo un indicador de desafíos educativos persistentes. Estos estados tienen tasas de analfabetismo más altas en comparación con otras regiones del país, lo que refleja desigualdades en el acceso a la educación y la falla del estado en proveer un mejor sistema de integración en estas regiones.



Con la gráfica podemos concluir que si existe una correlación lineal entre ambas variables. Esto debido a que, en muchas regiones con alta población indígena, el acceso a la educación puede ser limitado debido a la falta de infraestructura escolar o a la distancia a los centros educativos además de la escasez de maestros capacitados para enseñar en lenguas indígenas. Esto puede resultar en tasas más altas de analfabetismo.







En atención a las gráficas se pudo observar que existen ciertos valores extremos los cuales producen que nuestra gráfica se incline hacia la derecha, por lo que la media truncada (3321136.458) junto con la mediana (3054892) representan buenas medidas para describir a la población de México. Se calculo la desviación estándar con los valores normalizados obteniendo un valor de 20.15 % entre las poblaciones, se usó esta medida ya que representa se buena manera el comportamiento de los datos debido a sus características. Podemos observar que el histograma parece tener una distribución Poisson pues tiene la cola derecha muy pesada y la cola izquierda mas suave.

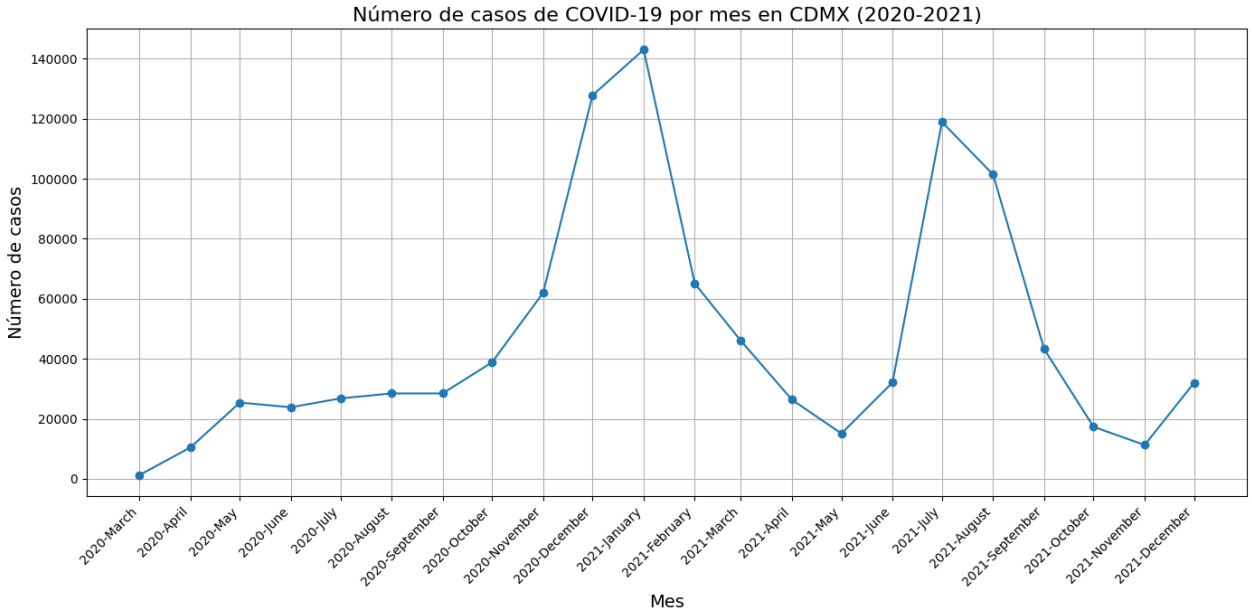
Ejercicio 2

Problema:

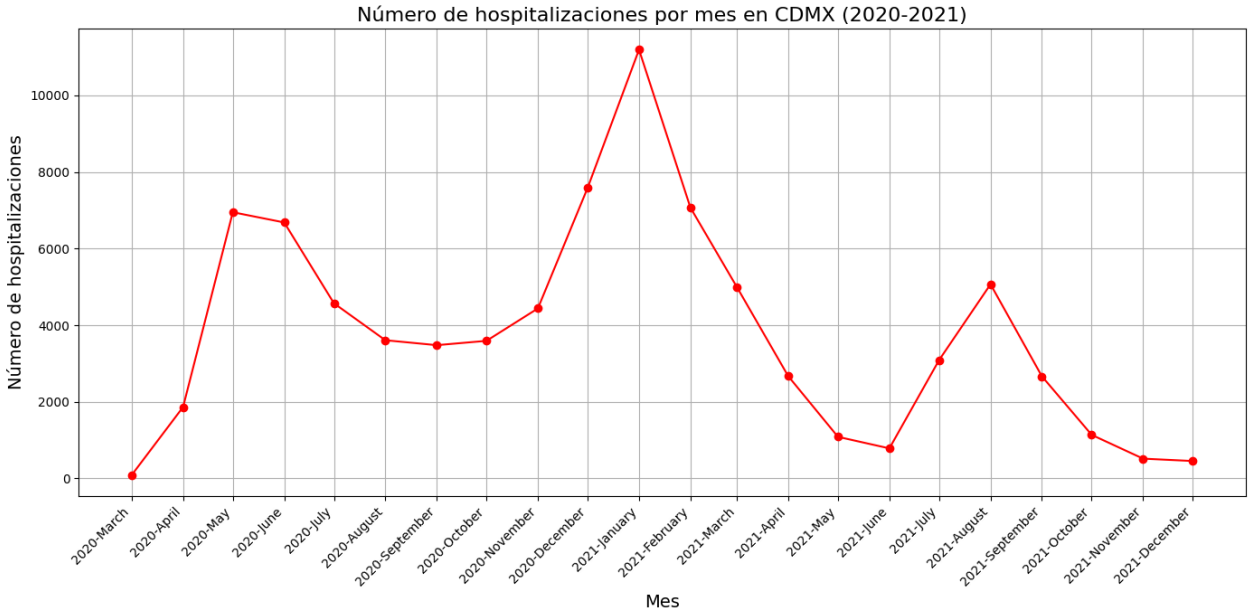
Descarga y filtrado de datos. Descarga los datos de la pandemia a nivel nacional, en la sección "Bases de datos históricas Influenza, COVID-19 y otros virus respiratorios", descarga Cierre Datos Abiertos Históricos para los años 2020 y 2021.

Solución:

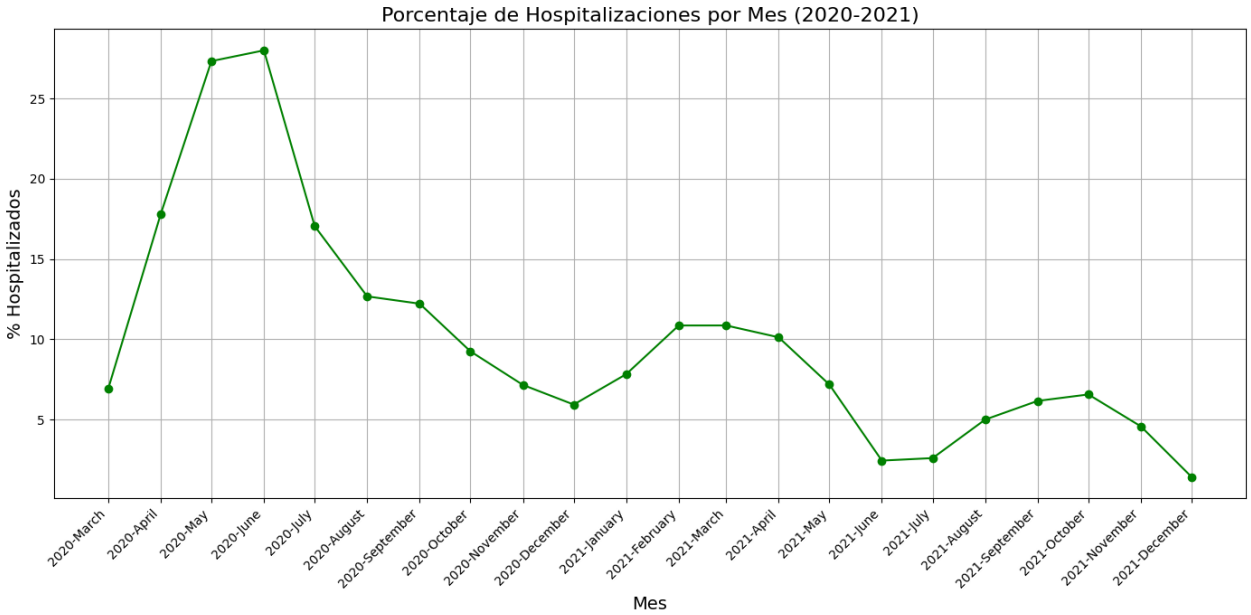
Primero visualizamos loas casos de Covid en la Cuidad de México , para poder hacer un análisis



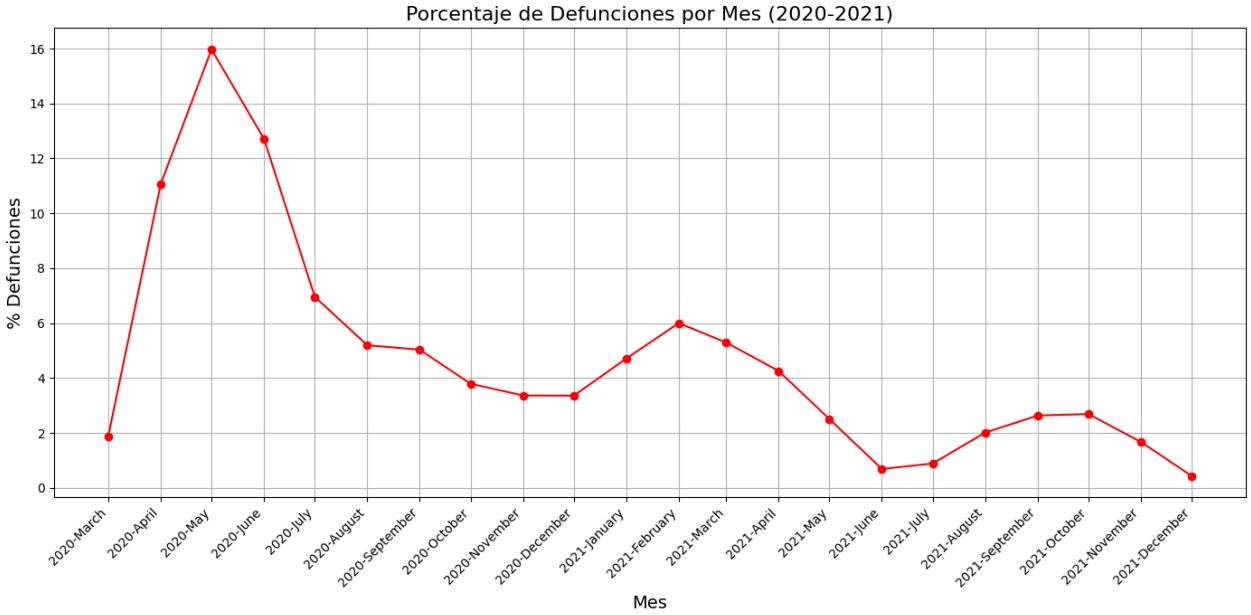
En esta gráfica se puede observar dos picos de casos en los meses de diciembre del 2020 y julio de 2021, esto muy probablemente se deba a que corresponden a periodos vacacionales para las familias mexicanas y tendieron a no seguir las medidas de prevención impuestas por el gobierno. Durante estas fechas, es común que las personas realicen reuniones familiares, viajes y actividades sociales lo cual facilitó la propagación del virus. Estos picos también podrían estar relacionados con la presencia de variantes más contagiosas que se presentaron en estos periodos.



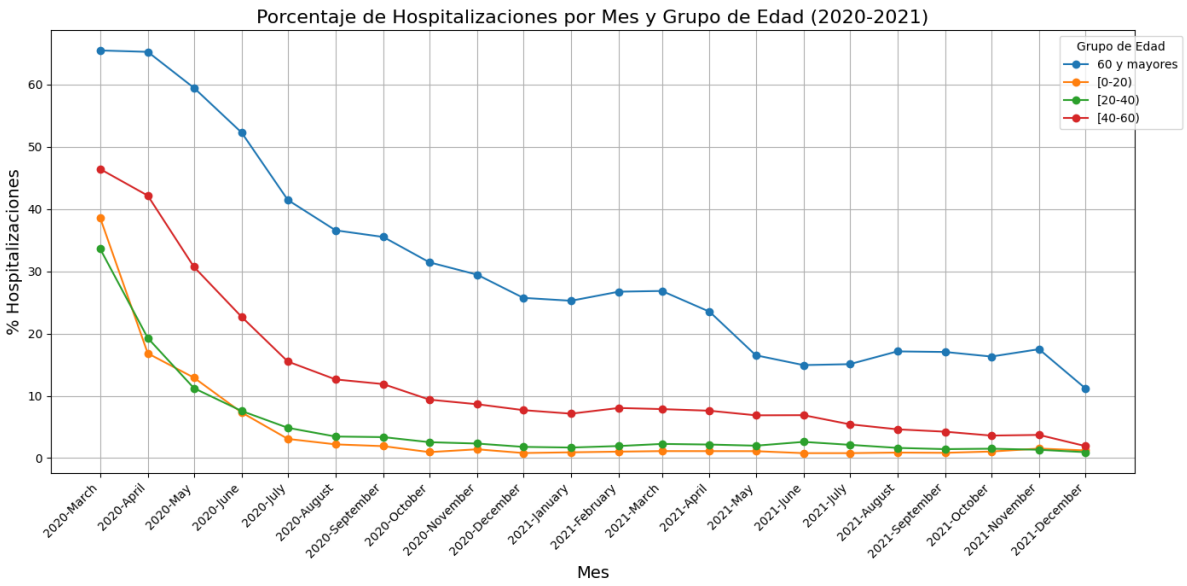
Los picos corresponden a los respectivos aumentos en la gráfica de casos, aunque a inicios de la pandemia existe otro pico, el cual probablemente sea debido a que la población se internaba en el hospital por miedo o sospechas sin necesariamente tener confirmada la enfermedad. Durante las primeras etapas de la pandemia, el pánico sobre el virus llevó a muchas personas a buscar atención médica ante cualquier síntoma similar a los de COVID-19. Esto pudo generar un aumento en las hospitalizaciones y, en consecuencia, un pico en los casos reportados, incluso si no todos estaban confirmados. Además, en ese momento, los protocolos de atención y las medidas de prevención aún estaban en sus primeras etapas, lo que pudo contribuir a una mayor saturación de los servicios de salud.



En esta gráfica se puede observar la proporción de ingresos al hospital en cuanto a casos registrados. Se puede concluir que, al principio de la pandemia, muchas personas se internaban al presentar síntomas o tener un caso sospechoso, lo que refleja el miedo y la incertidumbre inicial ante un virus desconocido. Sin embargo, con el paso del tiempo, las personas comenzaron a optar por tratarse por su cuenta o simplemente permanecer en casa para no contagiar a más personas. Este cambio en el comportamiento puede atribuirse a varios factores, como una mayor comprensión de la enfermedad, la disponibilidad de información sobre cómo manejar los síntomas leves en casa, y el deseo de evitar la saturación de los hospitales.



Se puede observar que existían muchos casos de defunciones al inicio de la pandemia respecto a los casos detectados. Esto se debió principalmente a que, en las primeras etapas, no se sabía cómo tratar la enfermedad de manera efectiva, lo que resultaba en un manejo clínico limitado y menos eficaz. Además, los hospitales estaban completamente saturados, lo que dificultaba brindar atención adecuada y oportuna a todos los pacientes. Este fenómeno subraya la importancia de la preparación y la capacidad de respuesta de los sistemas de salud ante emergencias sanitarias, así como la necesidad de invertir en infraestructura y recursos médicos para enfrentar futuras crisis de manera más eficiente.



Veamos que entre más grande era el grupo de edad, se presentaban más hospitalizaciones.Esto se debe a que, con el aumento de la edad, las personas tienden a tener un sistema inmunológico más debilitado que aumentan el riesgo de complicaciones por enfermedades infecciosas. Por otro lado, entre los dos grupos más jóvenes no existe una diferencia significativa en su porcentaje de hospitalizaciones, lo que sugiere que, en general, las personas más jóvenes tienen un sistema inmunológico más robusto y menos probabilidades de desarrollar cuadros graves que requieran hospitalización. Sin embargo, es importante destacar que, aunque los jóvenes tienen menores tasas de hospitalización, no son inmunes a la enfermedad y más concretamente representan un factor de riesgo ya que pueden transmitir el virus a grupos más vulnerables.

Ejercicio 3

Problema:

Requerimos ver como que funciona el Teorema Central del Limite. Considera una una v.a. $X \sim \mathcal{U}(0, 10)$, calcula su esperanza (μ) y varianza (σ^2). Fija $n = 5$ por lo que se requiere generar una m.a. $(X_1, X_2, \ldots, X_n) = (x_1^1, x_2^1, \ldots, x_n^1)$ de X.Calcular su media observada. Repite los dos pasos anteriores $m = 10,000$ veces para obtener $\bar{x}_n^1, \bar{x}_n^2, \ldots, \bar{x}_n^m$.Realiza un histograma de las medias observadas $\bar{x}_n^1, \bar{x}_n^2, \ldots, \bar{x}_n^m$ y además ajusta una distribución normal, repite toda la simulación para $n = 30, 100, 200$ y presenta las cuatro gráficas. Ahora considera una v.a. $X \sim$ Bernoulli(θ), calcula su esperanza (μ) y varianza (σ^2). Fija $\theta = 0.05, n = 5$ y realiza los exactamente los mismos pasos (ademas para $\theta = 0.3$)

Solución:

Primero calculemos la Esperanza y la varianza de una Uniforme de 0 a 10:

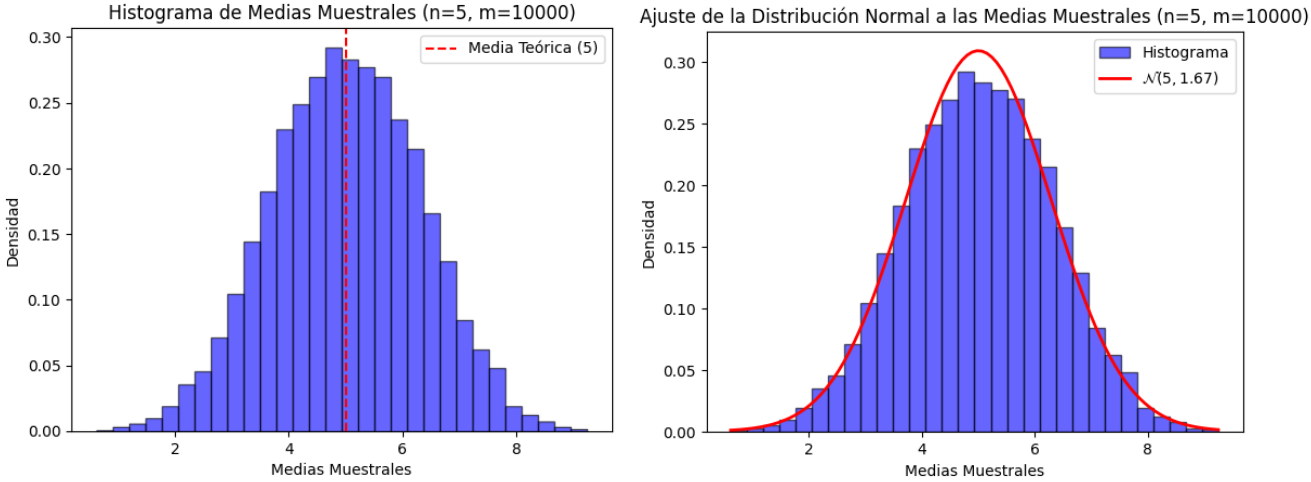
- Esperanza (μ): Sabemos que $X \sim \mathcal{U}(0, 10)$ por lo que podemos obtenerla mediante $\mu = \frac{a+b}{2}$ Ahora bien como tenemos que a $= 0$ y b $= 10$, sustituyendo tenemos que : $\mu = \frac{0+10}{2} = 5$

- Varianza (σ^2): Al tratarse de una uniforme sabemos que : $\sigma^2 = \frac{(b-a)^2}{12}$ Por lo que sustituyendo los valores tenemos que $\sigma^2 = \frac{(10-0)^2}{12} \approx 8.333$

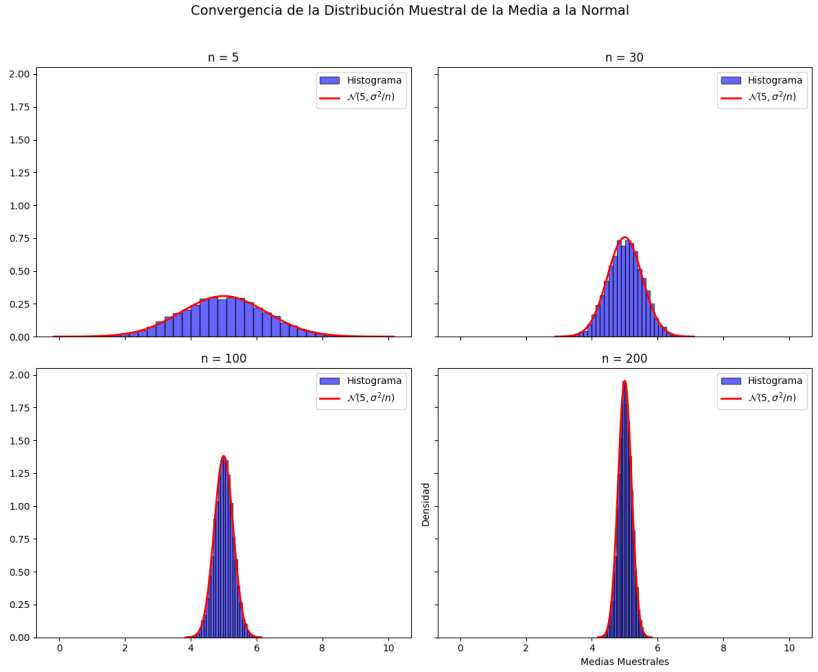
Ahora bien generando una muestra aleatoria nos dio que:

$$[2.1312344, 6.782021312.30704057, 7.53052761, 4.41506886]$$

Para calcular la media ocupamos la fórmula de: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Por lo que usando los datos anteriores nos dio $\bar{x} = 3.56104$. Realizamos esto 10,000 veces por lo que el histograma nos queda así y ajustando la normal al histograma:



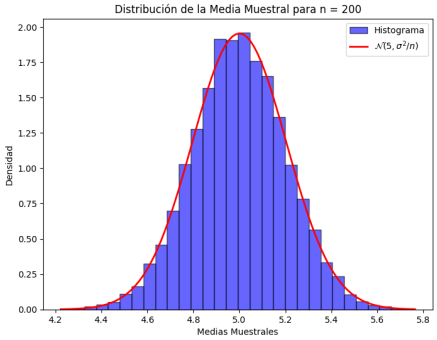
El mismo proceso lo realizamos para n distintas, de modo que nos queda así:



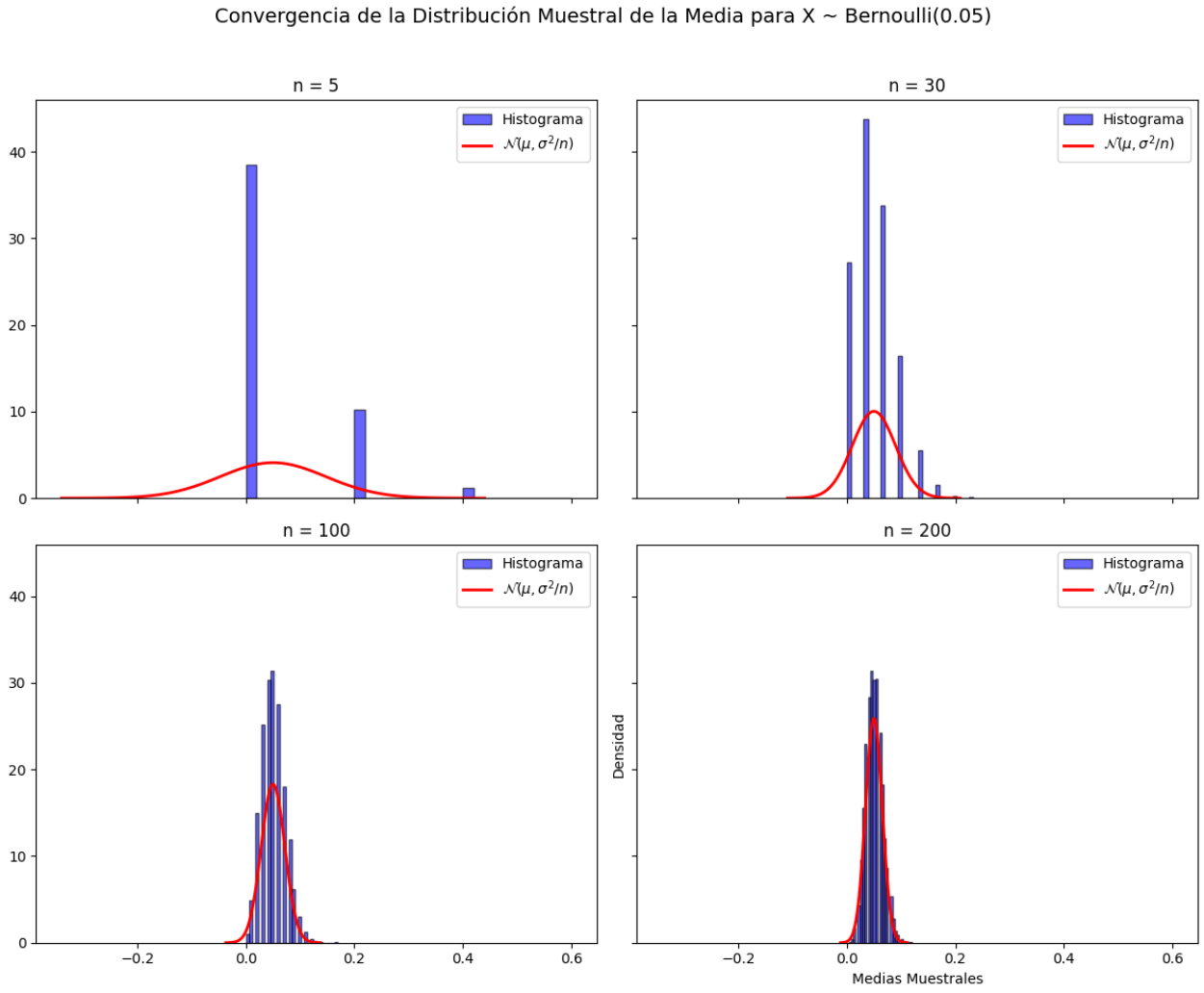
Ya podemos ver que conforme crecen las n va teniendo más a una normal, que es justo lo que queríamos ver en el ejercicio, pues el Teorema del Límite Central establece que:

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

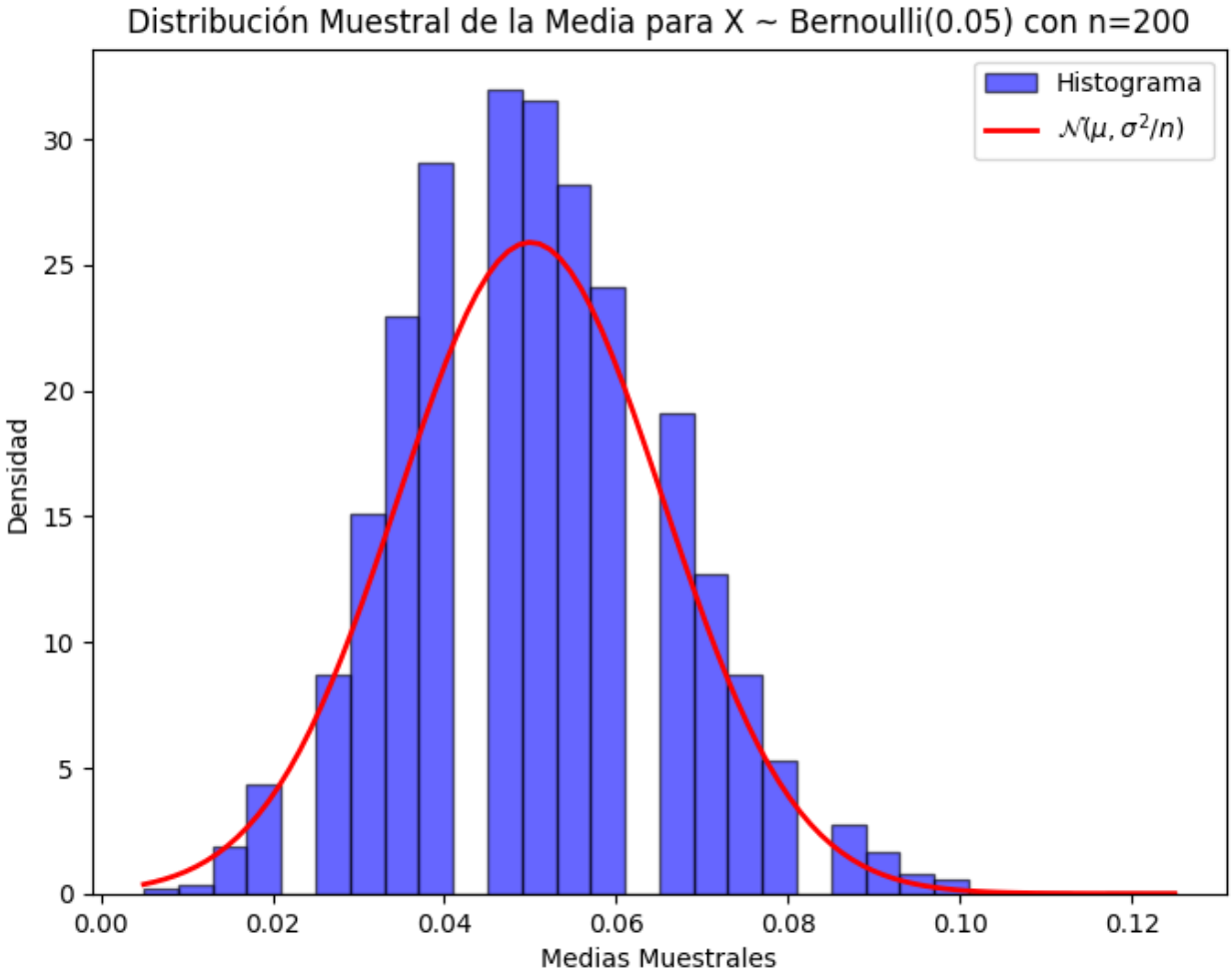
Por lo que agregamos la gráfica de la última reescalada y vemos que justamente se va acercando a lo que queremos ver.



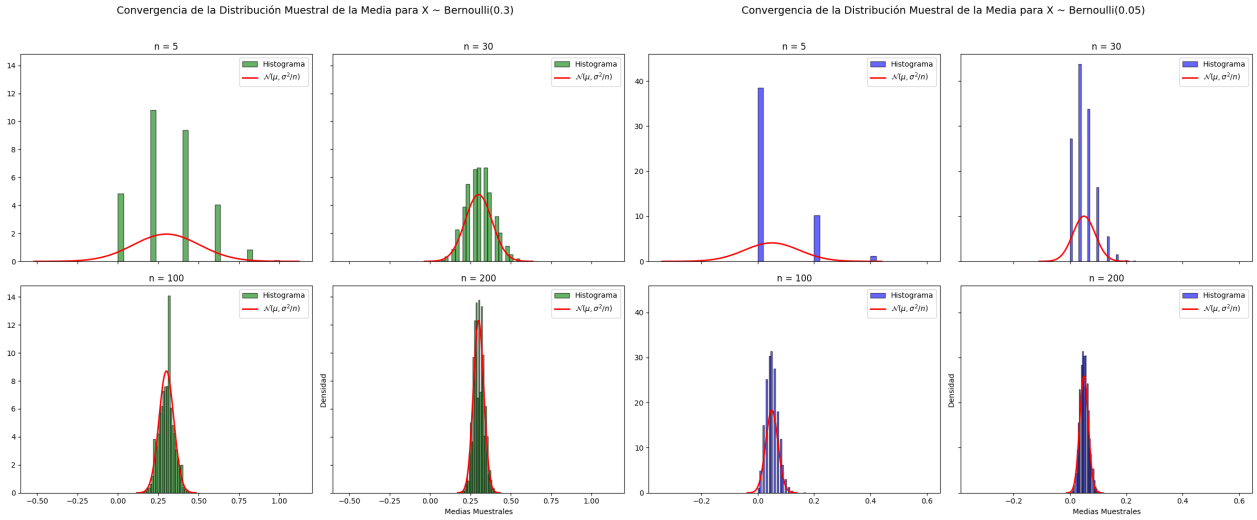
Realizando los mismos pasos con la Bernoulli, obtenemos los siguientes histogramas



Igual escalando la última, vemos que igual se va cumpliendo el teorema del limite central



Podemos ver que a diferencia de la uniforme , esta tarda mas tiempo en converger, es decir, aun necesita mas n's para que se vea que se pueda ver el teorema del límite central.
Pero ahora comparemos con la Bernoulli de $\theta = 0.05$ contra $\theta = 0.3$



Veamos que $\theta = 0.3$ converge más rápido al teorema del límite central , lo que nos puede

indicar que entre mas cercano theta a 1, converge más rápido.
Ahora bien, notemos que en el caso de la uniforme no necesitamos una n muy grande para que se vea la convergencia a una normal estándar, pero, en cambio, con la Bernoulli necesitamos muestras mas grandes para poder ver como converge la distribución Bernoulli a una Normal, Por lo que podemos decir que la convergencia depende de la función de distribución que utilicemos.

Ejercicio 4

Problema:
Sea $X_1, X_2, \ldots$ una sucesión infinita de variables aleatorias y sea θ una variable aleatoria latente. Supongamos que, condicionadas en θ, las variables X_i son independientes e idénticamente distribuidas según la densidad $f(x \mid \theta)$, es decir,

$$X_1, X_2, \cdots \mid \theta \overset{\text{i.i.d.}}{\sim} f(x \mid \theta).$$

Demuestra que la covarianza entre cualesquiera dos variables X_i y X_j es siempre no negativa, es decir,

$$\text{Cov}\,(X_i, X_j) \geq 0, \qquad \text{para } i \neq j.$$

Solución:
Recordemos que:

$$\text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$$

Usando la esperanza total:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[\mathbb{E}[X_i X_j|\Theta]]$$

Ahora bien, como X_i y X_j son independientes condicionados a Θ:

$$\mathbb{E}[X_i X_j|\Theta] = \mathbb{E}[X_i|\Theta]\mathbb{E}[X_j|\Theta]$$

$$\Rightarrow \mathbb{E}[X_i X_j] = \mathbb{E}[\mathbb{E}[X_i|\Theta]\mathbb{E}[X_j|\Theta]]$$

Veamos que X_i y X_j son idénticamente distribuidas condicionadas a Θ, por lo que:

$$\mathbb{E}[X_i|\Theta] = \mathbb{E}[X_j|\Theta]$$

Sea $\mu(\Theta) = \mathbb{E}[X_i|\Theta]$, entonces:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[\mu(\Theta)^2]$$

Veamos que:

$$\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i|\Theta]] = \mathbb{E}[\mu(\Theta)]$$

Por lo que la covarianza es:

$$\text{Cov}(X_i, X_j) = \mathbb{E}[\mu(\Theta)^2] - (\mathbb{E}[\mu(\Theta)])^2$$

Notemos que esto es la varianza de $\mu(\Theta)$:

$$\text{Cov}(X_i, X_j) = \text{Var}(\mu(\Theta))$$

Ahora bien, como la varianza de cualquier variable aleatoria es no negativa:

$$\Rightarrow \text{Cov}(X_i, X_j) = \text{Var}(\mu(\Theta)) \geq 0$$

$$\Rightarrow \text{Cov}(X_i, X_j) \geq 0$$