Integrantes:

David Melo Valbuena - 2230232
Martín García Chagüezá - 2230436
Juan Andrés Ruiz Muñoz - 2230557
Jose Daniel Carrera Bolaños - 2215335

Colab:
001_StatisticalDistributionSolutions.ipynb

Pregunta 1: Documentación del ejercicio: Distribuciones Binomial y Exponencial
En este ejercicio se analizaron dos distribuciones de probabilidad: la distribución binomial y la distribución exponencial. Para cada una de ellas, se generó una muestra aleatoria de tamaño 150, se construyó un histograma que se comparó con su función de probabilidad teórica, y se calcularon algunos estadísticos clave, como el promedio, los cuartiles y la desviación estándar.
1. Distribución Binomial
La distribución binomial se define por dos parámetros:

- n: número de ensayos (en este caso, n = 10).
- p: probabilidad de éxito en cada ensayo (en este caso, p = 0.5).

Se generó una muestra de 150 valores utilizando la función np.random.binomial(n, p, 150) de NumPy, con lo cual se obtuvo un histograma de frecuencias que fue comparado con la función de masa de probabilidad teórica (PMF) de la binomial, la cual describe la probabilidad de obtener exactamente k éxitos en n ensayos.
Resultados:

- Promedio de la muestra: 4.87.
- Cuartiles: [4, 5, 6].
- Desviación estándar: 1.58.

Estos resultados concuerdan con las expectativas teóricas de una distribución binomial con parámetros n = 10 y p = 0.5, donde el valor esperado es 5 y la desviación estándar teórica es 1.58.
2. Distribución Exponencial
La distribución exponencial se caracteriza por el parámetro λ (lambda), que es la tasa de ocurrencia de un evento. En este caso, se utilizó un valor de λ = 0.2, lo que corresponde a un promedio teórico de 5 (es decir, 1/λ).
La muestra aleatoria fue generada utilizando la función np.random.exponential(1/lam, 150) de NumPy. Se construyó un histograma de densidad que fue comparado con la función de densidad de probabilidad teórica (PDF) de la distribución exponencial.
Resultados:

- Promedio de la muestra: 5.03.
- Cuartiles: [1.33, 3.75, 7.49].
- Desviación estándar: 4.72.

Estos valores son consistentes con los esperados teóricamente, dado que el promedio esperado para una distribución exponencial con $\lambda = 0.2$ es de 5, y la desviación estándar esperada es también 5.

Comparación de los resultados:

Para la distribución binomial, tanto el promedio como la desviación estándar obtenidos de la muestra se acercan a los valores teóricos. Los cuartiles también son coherentes con la simetría esperada de una binomial con $n = 10$ y $p = 0.5$.

Para la distribución exponencial, el promedio y la desviación estándar de la muestra son consistentes con los valores teóricos. Los cuartiles reflejan la naturaleza asimétrica de esta distribución, donde una mayor proporción de los datos se concentra en valores bajos.

Pregunta 2: Análisis Estadístico del Sistema de Almacenamiento de Datos.

Introducción

En esta sección analizamos el rendimiento de un sistema de almacenamiento de datos en la nube, utilizado por una gran empresa de comercio electrónico para procesar continuamente transacciones de ventas. El análisis se enfoca en evaluar las probabilidades asociadas a los tiempos entre llegadas de transacciones y otros comportamientos del sistema, utilizando herramientas estadísticas como la distribución exponencial y la distribución de Poisson. El objetivo es obtener información clave que permita mejorar la gestión de estos tiempos, optimizando la eficiencia del sistema.

Resumen de los hallazgos

A continuación se presenta un resumen de los principales hallazgos obtenidos a partir del análisis de los datos:

1. Probabilidad de que el tiempo entre dos transacciones consecutivas sea menor o igual a 2 segundos: El resultado obtenido es aproximadamente 0.487, lo que indica que es casi igualmente probable que el tiempo entre dos transacciones sea menor o igual a 2 segundos como que sea mayor. **Esto sugiere que el sistema tiene tiempos de respuesta relativamente rápidos.**

2. Probabilidad de que el tiempo entre dos transacciones consecutivas sea mayor a 10 segundos: La probabilidad obtenida es aproximadamente 0.036, lo que significa que es muy poco probable que el sistema espere más de 10 segundos entre transacciones. **Esto sugiere que las transacciones suelen ocurrir en intervalos mucho más cortos, lo cual es positivo para la eficiencia del sistema.**

3. Probabilidad de que no llegue ninguna transacción en un período de 30 segundos (respaldo): La probabilidad obtenida es extremadamente baja, aproximadamente 0.000045, lo que implica que es muy poco probable que el sistema quede inactivo durante un período de

respaldo de 30 segundos. **Esto indica que el sistema está constantemente procesando transacciones y que los tiempos muertos son muy raros.**

4. Tiempo esperado entre dos transacciones consecutivas: El tiempo esperado es de 3 segundos. Esto sugiere que, en promedio, el sistema recibe una nueva transacción cada 3 segundos, **lo que indica una alta actividad y una eficiente capacidad de procesamiento.**

5. Probabilidad de que ocurran al menos 5 transacciones en un intervalo de 10 segundos: La probabilidad es de aproximadamente 0.244, lo que significa que es relativamente poco probable que ocurran al menos 5 transacciones en tan solo 10 segundos. **Sin embargo, esta probabilidad no es insignificante, lo que sugiere que el sistema puede manejar ráfagas de datos con cierta frecuencia.**

Conclusiones:

A partir del análisis realizado, podemos concluir que el sistema de almacenamiento de datos en la nube para la empresa de comercio electrónico está **bien diseñado para manejar grandes cantidades de transacciones con tiempos de espera cortos entre ellas**. El tiempo esperado entre transacciones consecutivas es de solo 3 segundos, lo que es un buen indicador de eficiencia. Además, la probabilidad de que **el sistema permanezca inactivo durante un periodo de respaldo es casi nula, lo que muestra su fiabilidad.**

Es importante destacar que la probabilidad de que ocurran ráfagas de datos de al menos 5 transacciones en 10 segundos no es insignificante, lo que indica que el sistema debe estar preparado para manejar tales eventos. El uso de técnicas estadísticas avanzadas, como las distribuciones exponencial y de Poisson, ha permitido obtener información valiosa para optimizar la eficiencia del sistema y garantizar su buen funcionamiento.

Pregunta 3:

a Gráficas PDF y CDF de la temperatura
Las gráficas muestran cómo se distribuyen las temperaturas de los servidores. La gráfica de la PDF (Función de Densidad de Probabilidad) revela que las temperaturas cercanas a los 35°C son las más probables, con una disminución hacia los extremos (temperaturas menores a 30°C o mayores a 40°C son poco frecuentes). La CDF (Función de Distribución Acumulada) nos ayuda a entender la probabilidad acumulada de que la temperatura no exceda un cierto valor. La transición rápida de baja a alta probabilidad en el rango de 30°C a 40°C indica que la mayor parte del tiempo los servidores están dentro de este rango.

b Probabilidad de que el sistema de enfriamiento se active (> 40°C)
- Resultado: La probabilidad de que el sistema de enfriamiento se active cuando la temperatura supere los 40°C es 0.0228 o 2.28%.

- Conclusión: Esta baja probabilidad indica que el sistema de enfriamiento no se activará con frecuencia, ya que los servidores rara vez alcanzan temperaturas superiores a 40°C. Esto sugiere que el sistema está diseñado de manera eficiente, manteniendo las temperaturas bajo control en la mayoría de los casos.

c Porcentaje de tiempo que los servidores operan entre 30°C y 38°C
- Resultado: Los servidores operan en el rango óptimo de temperatura entre 30°C y 38°C el 86.22% del tiempo.

- Conclusión: Un porcentaje elevado de tiempo (más del 85%) en este rango muestra que los servidores operan de manera eficiente y en condiciones seguras la mayor parte del tiempo. El sistema de refrigeración está bien optimizado para mantener las temperaturas dentro de estos límites, minimizando el riesgo de sobrecalentamiento o enfriamiento excesivo.

d Temperatura para el 10% superior de los eventos de temperatura
- Resultado: La temperatura que corresponde al 10% superior de los eventos es 38.20°C.

- Conclusión: Las temperaturas por encima de 38.20°C representan el 10% más alto de los eventos de temperatura. Este valor puede usarse como umbral para generar alertas tempranas y tomar medidas preventivas antes de que las temperaturas alcancen niveles críticos, como el punto de activación del sistema de enfriamiento a los 40°C.

e Probabilidad de que el promedio de 5 servidores exceda los 37°C
- Resultado: La probabilidad de que el promedio de temperatura de 5 servidores supere los 37°C es 0.0368 o 3.68%.

- Conclusión: Aunque la probabilidad de que un solo servidor supere los 37°C es baja, existe una pequeña probabilidad de que el promedio de 5 servidores lo haga. Esto sugiere que, aunque raro, es posible que varios servidores simultáneamente experimenten temperaturas superiores a los 37°C, lo que podría requerir atención adicional o ajustes en la refrigeración para estos casos.

f Probabilidad de que más de 25 de 1000 servidores superen los 41°C simultáneamente
- Resultado: La probabilidad de que más de 25 de los 1000 servidores superen los 41°C al mismo tiempo es extremadamente baja: 0.000000 o prácticamente 0%.

- Conclusión: La posibilidad de que 25 o más servidores superen los 41°C de manera simultánea es prácticamente inexistente. Esto demuestra que el sistema de refrigeración es muy eficaz en mantener las temperaturas controladas incluso cuando se considera un gran número de servidores, lo que garantiza la estabilidad del sistema en su conjunto.

Conclusión General

Las temperaturas de los servidores están bien reguladas. La probabilidad de que el sistema de enfriamiento se active es baja (solo el 2.28%), y los servidores operan en el rango óptimo de 30°C a 38°C el 86.22% del tiempo, lo que es una proporción muy alta. Los eventos de temperaturas extremas, como superar los 40°C o que más de 25 servidores alcancen los 41°C simultáneamente, son extremadamente raros. Esto indica que tanto el sistema de monitoreo como el de enfriamiento están funcionando eficientemente, garantizando condiciones de operación seguras y estables para los servidores.

Pregunta 4: Web_Server_Requests_G

Introducción

El análisis de la tasa de llegada de solicitudes a un servidor web es fundamental para garantizar su correcto funcionamiento, evitando posibles sobrecargas y mejorando la planificación de la infraestructura. Este documento aborda un análisis detallado del comportamiento del tráfico web en un servidor utilizando un conjunto de datos sintéticos que registra la tasa de llegada de solicitudes cada cinco minutos. A través de visualizaciones y el ajuste de una distribución gamma a los datos, se busca identificar patrones clave de actividad, calcular probabilidades críticas, y evaluar el riesgo de sobrecarga del servidor.

Resumen de los hallazgos

a) Patrones de tráfico web: picos de actividad y momentos de menor demanda

En el análisis de la tasa de llegada de solicitudes, se identificaron varios patrones importantes:

- Picos de actividad: Los picos de actividad más pronunciados ocurrieron alrededor de los *10 minutos* y *40 minutos*, con un valor promedio de solicitudes cercano a 3.96 en esos momentos.
- Momentos de menor demanda: Los intervalos de *25 minutos* y *50 minutos* presentaron una caída significativa en la tasa de llegada, bajando por debajo de 3.88 solicitudes.
- Variabilidad moderada: La tasa de llegada fluctuó entre *3.88 y 3.96 solicitudes*, mostrando una variabilidad moderada que sugiere una carga relativamente estable en el servidor, aunque con fluctuaciones notables.

b) Ajuste de una distribución gamma a los datos de la tasa de llegada

Se ajustó una distribución gamma a los datos para modelar la variabilidad en la tasa de llegada. Los parámetros ajustados fueron:

- Shape (α): 1.287
- Loc: 0 (*fijado a 0 para este análisis*)
- Scale (θ): 3.047

La mayor parte de los datos se concentró en tasas bajas, con una distribución de cola larga que representa intervalos ocasionales con tasas de llegada más altas.

c) Probabilidad de recibir más de 8 solicitudes en un intervalo

Se calculó una probabilidad del **11.58%** de que el servidor reciba más de 8 solicitudes durante el próximo intervalo. Aunque esto no es frecuente, se debe considerar la posibilidad de sobrecarga en picos de demanda.

d) Probabilidad de que la tasa de llegada supere 15 solicitudes por minuto

La probabilidad de que el servidor reciba más de 15 solicitudes en un solo intervalo es muy baja, aproximadamente **1.35%**, lo que indica que el servidor rara vez enfrenta una carga tan alta.

e) Probabilidad de superar la capacidad máxima de procesamiento del servidor

Dado que la capacidad máxima del servidor es de 20 solicitudes por intervalo, se calculó una probabilidad de **0.28%** de que el servidor reciba más de 20 solicitudes en el próximo intervalo, lo que indica que el riesgo de sobrecarga es extremadamente bajo.

Conclusiones

El análisis sugiere que, aunque existen fluctuaciones en la tasa de llegada de solicitudes, el servidor web se enfrenta a una carga mayormente estable con algunos picos ocasionales de actividad. El ajuste de la distribución gamma proporciona una buena representación de la variabilidad en los datos, y las probabilidades calculadas muestran que el riesgo de sobrecarga del servidor es bajo, pero no inexistente. Esto sugiere que el servidor está bien preparado para manejar la mayoría de las cargas, pero sería prudente monitorear los picos de actividad para ajustar la capacidad del servidor si fuera necesario.