Análisis Descriptivo del Taller de Distribuciones de Muestreo y Pruebas de Hipótesis

Universidad del Norte
Análisis de Datos en Ingeniería

06 de junio de 2024

Introducción

En este taller se analiza un sistema de entrega de paquetes en una ciudad utilizando el concepto de *Crowdshipping*. El objetivo es observar experimentalmente el cumplimiento del teorema del límite central, analizar el riesgo de una prueba de hipótesis y afianzar el uso de herramientas informáticas para el análisis y manipulación de datos. Los datos se generaron mediante un simulador que modela la operación del sistema de entrega de paquetes, y se analizaron los tiempos de llegada de paquetes y viajeros de la multitud.

Procedimiento y Análisis

Paso 0: Generación de Datos Simulados

Se corrió la simulación para generar los archivos de salida necesarios, asegurando obtener al menos 500,000 observaciones de cada variable. Esto se logró ajustando los parámetros de tiempo de simulación en 14 horas por dia y número de réplicas en 30 dias en el archivo `masterfile.csv`.

Paso 1: Lectura de Archivos CSV

Se leyeron los archivos `outputs_parr.csv` y `outputs_carr.csv` utilizando la función `read_csv` de pandas para almacenar los datos en data frames. Estos archivos contienen los tiempos simulados de llegada de paquetes con 588227 datos y viajeros de la multitud con 881217 datos, respectivamente.

Paso 2: Separación de Origen y Destino

Se extrajeron los campos de origen y destino del texto en cada archivo CSV utilizando las funciones `extract` de pandas. El data frame resultante para cada archivo contiene tres columnas: origen, destino y tiempo de llegada.

Paso 3: Cálculo del Tiempo entre Llegadas

Se creó un nuevo campo en cada data frame que calcula el tiempo entre llegadas de cada paquete y viajero. Este campo se obtuvo ordenando el data frame por tiempo de llegada y calculando la diferencia entre el tiempo de llegada de un paquete o viajero y el anterior.

Paso 4: Análisis de la Distribución del Tiempo entre Llegadas

Se analizaron las distribuciones de la variable *tiempo entre llegadas* para paquetes y viajeros. Se calcularon estadísticos descriptivos (media, Q1, Q2, Q3, desviación estándar, coeficiente de asimetría y curtosis), estos datos se encuentran en segundos (s) y se graficaron histogramas de frecuencias relativas y diagramas de cajas.

Estadístico	**Paquetes**	**Viajeros**
Media	0.0684	0.0576
Q1	0.018	0.018
Mediana (Q2)	0.0468	0.0396
Q3	0.0936	0.0792
Desviación Estándar	0.0684	0.0576
Coeficiente de Asimetría	1.98	2.002
Curtosis	5.84	5.99

Table 1: Estadísticos descriptivos de los tiempos entre llegadas

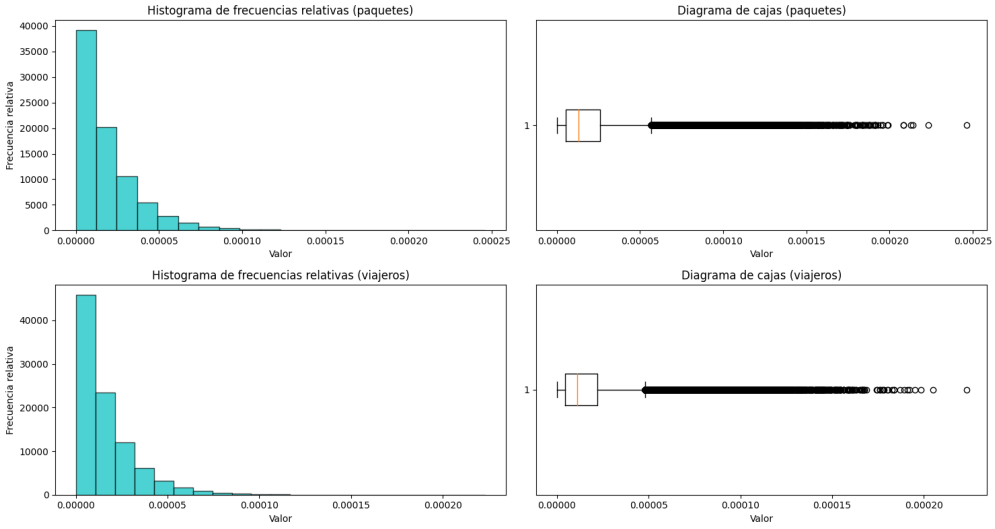

Figure 1: Distribución del tiempo entre llegadas de paquetes y viajeros

Los estadísticos descriptivos mostraron que ambas distribuciones presentaban un sesgo positivo y mucho apuntamiento, sugiriendo que seguían una distribución exponencial. La comparación entre ambas distribuciones mostró que los tiempos entre llegadas de paquetes eran más variables que los tiempos entre llegadas de viajeros.

Paso 5: Prueba de Bondad de Ajuste

Se planteó la hipótesis de que los tiempos entre llegadas seguían una distribución exponencial. Para verificar esta hipótesis, se realizó una prueba de bondad de ajuste Kolmogorov-Smirnov utilizando la función `kstest` de Scipy. Los resultados indicaron que los tiempos entre llegadas de viajeros se ajustaban mejor a la distribución exponencial que los tiempos entre llegadas de paquetes, teniendo parametros $\lambda = 1.8701 \times 10^{-5}$ y $\lambda = 1.588 \times 10^{-5}$ respectivamente.

Variable	**Estadístico K-S**	**Valor p**
Paquetes	0.000821	0.8220
Viajeros	0.00120	0.156

Table 2: Resultados de la prueba de Kolmogorov-Smirnov

Paso 6: Creación de Data Frames de Muestras Aleatorias

Desde los data frames originales, se extrajeron aleatoriamente 10000 muestras de tamaño 50 para cada variable. Estas muestras se organizaron en nuevos data frames con las muestras como columnas y las observaciones como filas. En

ambos casos quedanron valores de la poblacion original que no fueron tomados por lo que es probable que haya posibles errores o discrepancias a los valores esperados.

Paso 7: Análisis de las Medias Muestrales y Prueba de Hipótesis

Se calcularon las medias muestrales para cada muestra, por cada variable y se describió la distribución del conjunto de medias muestrales. Se comparó esta distribución con la distribución de frecuencias de los datos originales, observándose que las medias muestrales seguían una distribución más cercana a la normal esto nos permite observar el efecto del teorema del límite central.

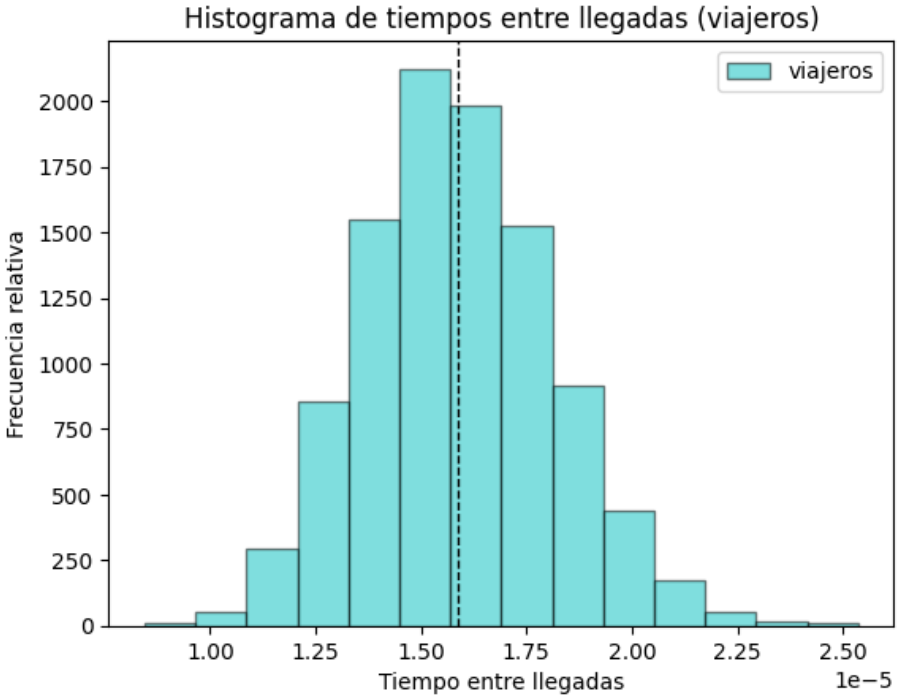

Figure 2: Histograma de frecuencias relativas de las M medias muestrales de la poblacion de Viajeros

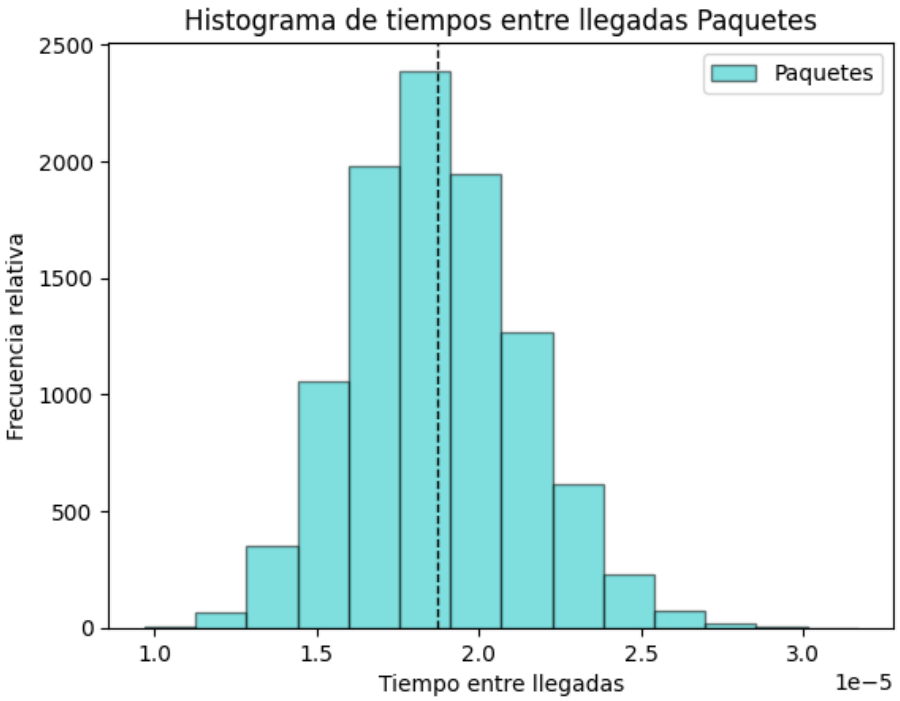

Figure 3: Histograma de frecuencias relativas de las M medias muestrales de la poblacion de Paquetes en el sistema

Se realizó una prueba de hipótesis para inferir si el tiempo promedio entre llegadas era de 2 segundos. Se estableció un nivel de significancia del 5% y se ejecutaron las 10000 pruebas de hipótesis. Los resultados mostraron que, en el escenario donde llegan en promedio 25,198 paquetes o viajeros en un día de operación de 14 horas, la hipótesis nula se rechazaba en aproximadamente el 100% de las pruebas para el caso de los paquetes y el de los viajeros, por lo que es valido afirmar que la media de llegada de paquetes o viajeros al sistema no es 2 segundos, ademas este mismo comportamiento fue observado al realizar nuevamente las 10000 pruebas de hipotesis con el valor de media dado por el escenario b. por lo que se reafirma la idea de que el valor de la media de la poblacion debe ser inferior en una gran maginitud a 2 segundo.

Variable	**Media Muestral**	**Rechazo de H0 (%)**
Paquetes	2	100
Viajeros	2	100

Table 3: Resultados de las pruebas de hipótesis

Paso 8: Conteo de Paquetes y Viajeros por Origen-Destino

Se contó el número de paquetes y viajeros observados para cada par origen-destino y se guardaron los resultados en nuevos data frames. Estos resultados se escribieron en archivos CSV separados, tanto para el experimento 1 y 2, encontrandose 2450 parejas origen-destino diferentes.

Origen-Destino	**Paquetes**	**Viajeros**
Arabellapark - Bohmerwaldplatz	252	368
Westpark - Thalkirchen	222	311
Westpark - Westfriedhof	225	350
...	...	...

Table 4: Conteo de paquetes y viajeros por origen-destino (muestra)

Paso 9: Data Frame con Paquetes por Ubicación y Mapa de Calor

Se crearon data frames mostrando el número total de paquetes y viajeros que salieron y llegaron a cada ubicación. Utilizando las coordenadas de las estaciones de tren subterráneo de Múnich (U-Bahn), se construyeron mapas de calor que ilustran la distribución geográfica del flujo de paquetes y viajeros. Estos gráficos se guardaron en archivos .html separados.

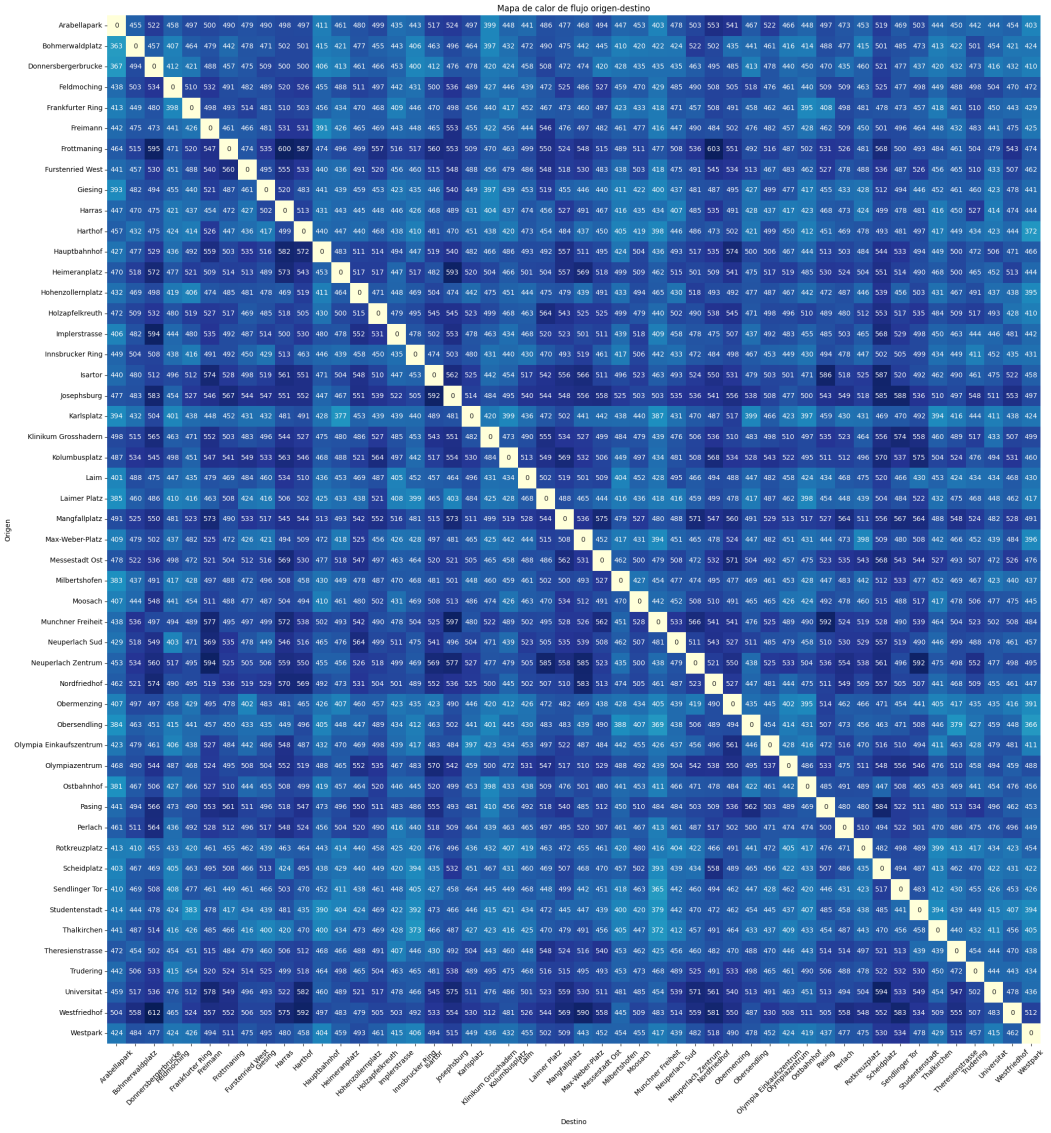

Figure 4: Mapa de calor de paquetes por ubicación

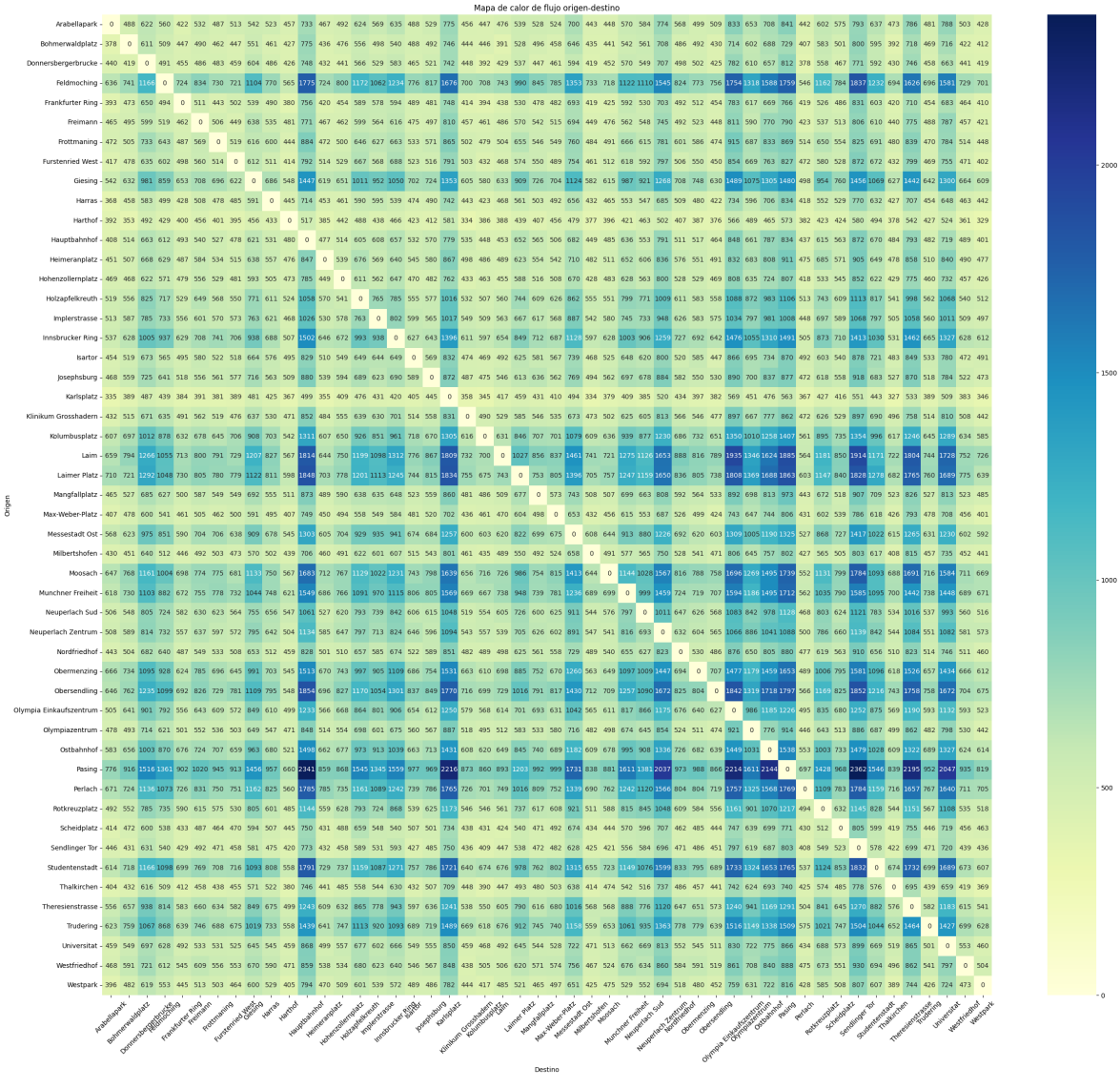

Figure 5: Mapa de calor de viajeros por ubicación

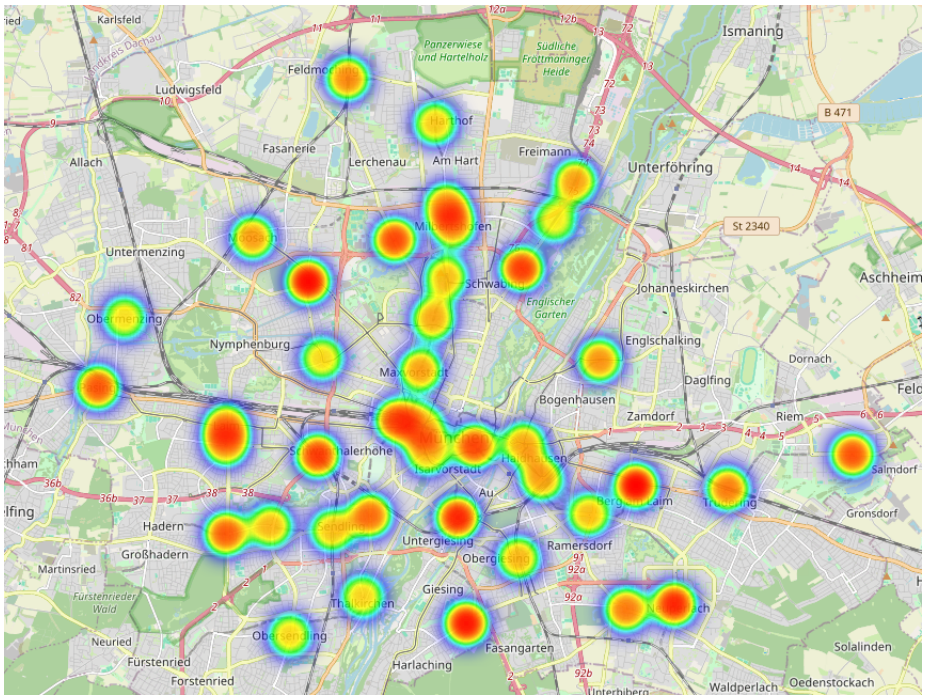

Figure 6: Mapa de calor geografico de paquetes (saliendo) por ubicación

Conclusiones

El análisis realizado permitió observar el cumplimiento del teorema del límite central y evaluar el riesgo en pruebas de hipótesis. Las distribuciones de los tiempos entre llegadas de paquetes y viajeros presentaron diferencias significativas, lo que sugiere la necesidad de estrategias distintas para optimizar el sistema de *crowdshipping*. La capacidad del sistema para suplir la demanda fue analizada y se concluyó que, en general, los viajeros pueden suplir adecuadamente la demanda de entrega de paquetes, aunque con variabilidad en los tiempos entre llegadas.

Archivos Adjuntos

- `reporte.pdf`
- `notebook.ipynb`
- `outputs/distribucion_tiempos.png`
- `outputs/paquetes_o-d.csv`

- outputs/viajeros_o-d.csv
- outputs/mapa_calor_paquetes.png
- outputs/mapa_calor_viajeros.png