República Bolivariana de Venezuela
Ministerio del Poder Popular para la Educación
Universidad de Carabobo — Facultad de Ciencias y Tecnología
Naguanagua, Edo. Carabobo

MODELADO DE TRÁFICO DE RED CON POISSON Y EXPONENCIAL

Autor: Jesús Hernández
Materia: Probabilidad — 6to Semestre
Profesor: Mirba Romero
18 de octubre de 2025

Informe Técnico

Índice

Introducción

Modelamos un tráfico de red a partir de un CSV con marcas de tiempo, tamaños de paquete y protocolo. Se analiza:

- Variable discreta: número de paquetes por segundo, con hipótesis Poisson(λ).
- Variable continua: tiempo entre llegadas, con hipótesis Exponencial(λ).
- Variables conjuntas: Protocolo (TCP/UDP) × Tamaño (Pequeño/Grande), con tabla de probabilidades y pruebas de independencia.

Datos y Metodología

- Columnas: Timestamp (fecha-hora), Packet_Size (bytes), Protocol (6=TCP, 17=UDP).
- Preprocesamiento: orden por Timestamp; agregación por segundo para conteos; diferencias consecutivas para interarribos.
- Estimaciones: λ_discreto como promedio de paquetes/seg en el rango; λ_expon = 1 / media(interarribos).
- Gráficos: histograma de conteos con PMF Poisson; histograma de interarribos con PDF Exponencial. Mapa de calor para conjuntas.

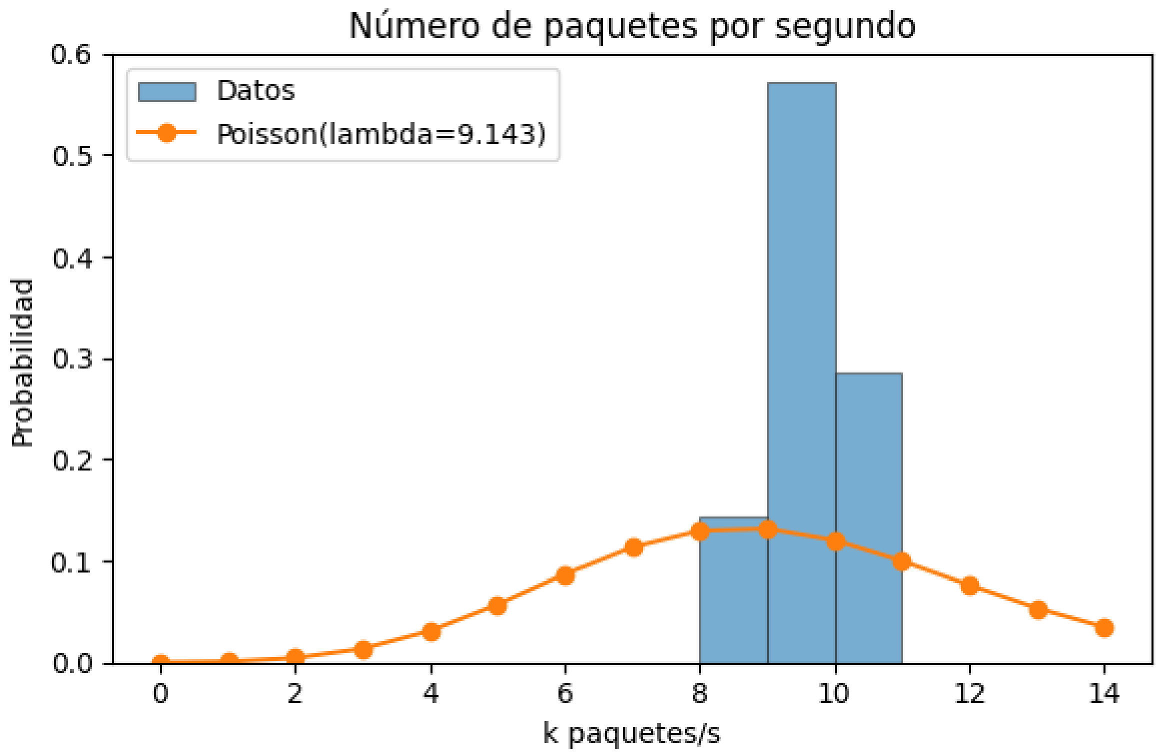
Decisiones de análisis (consolidadas):

- Discreta (Poisson): se usa la ventana 1..7 definida por el usuario (no se rellenan extremos con ceros). En esa ventana λ = 64/7 ≈ 9.142857.
- Continua (Exponencial): interarribos sobre los timestamps originales; si hay muchos Δ=0 por resolución, se reporta y se puede estimar λ ignorando Δ=0.
- Conjunta: tamaño discretizado a Pequeño (≤500) y Grande (>500). En este dataset no hay "Grandes".

Resultados

Variable Discreta (Poisson)

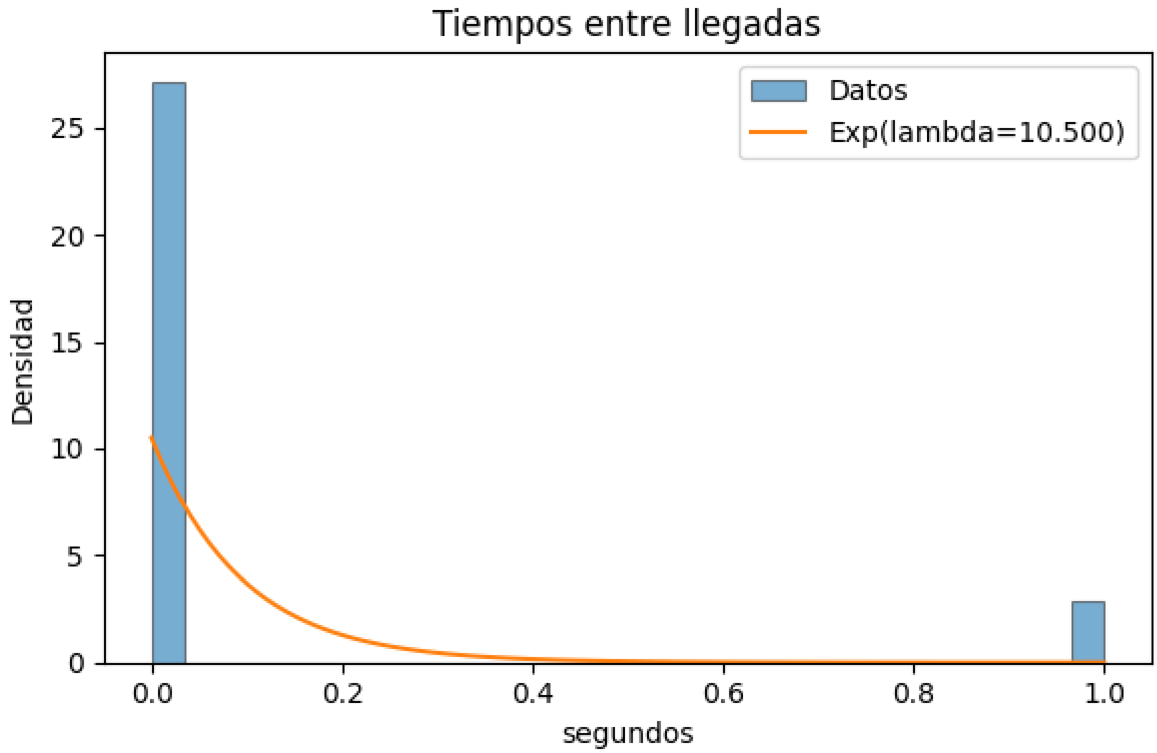
- λ por conteos (rango seleccionado) = **9.142857** 1/s.
- Índice de dispersión Var/Media ≈ **0.052083** (≈1 en Poisson).



- Figura:
  Interpretación: la curva Poisson(λ) superpuesta indica el grado de ajuste. Var/Media cercano a 1 sugiere congruencia con un proceso de Poisson; desviaciones notables indican sub/sobredispersión o muestra pequeña.

Variable Continua (Exponencial)

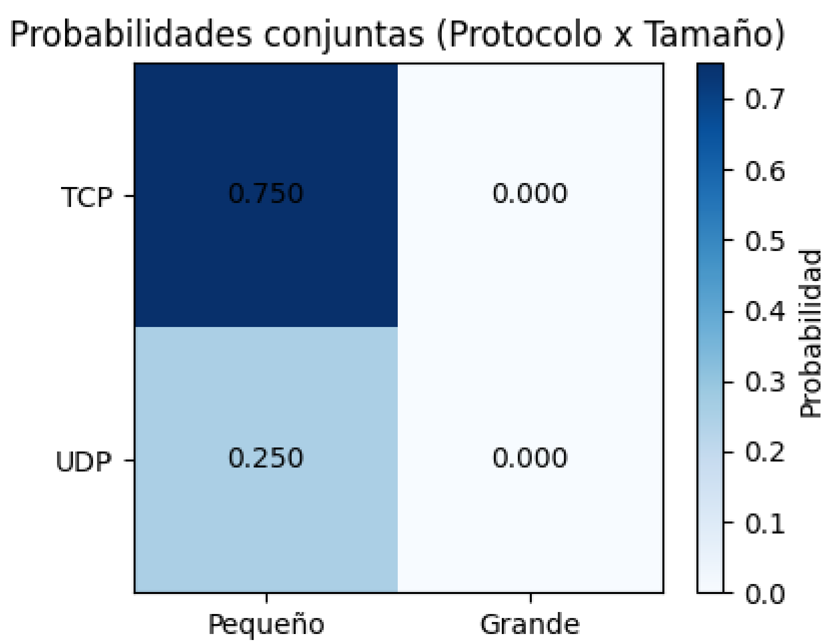
- λ por interarribos = **10.500000** 1/s (λ = 1/mean).
- KS test exponencial (opcional): estadístico = **0.904762**, p-valor = **0.000000**.



- Figura:
  Interpretación: si la PDF Exp(λ) describe razonablemente el histograma, el modelo exponencial es plausible. El p-valor del KS se interpreta con cautela cuando n es pequeño.

Variables Conjuntas (Protocolo × Tamaño)

- P(UDP) = **0.25 (25%)**; P(TCP) = **0.75 (75%)**; P(Pequeño) = **1.00 (100%)**;P(Grande) = **0.00 (0%)**; P(Grande | TCP) = **0.00 (0%)**.



- Figura (mapa de calor):
  Explicación matemática de independencia:
- Sean A = {protocolo es TCP} y B = {tamaño es Grande}. A y B son independientes si P(A ∩ B) = P(A)·P(B), equivalente a P(B|A)=P(B) cuando P(A)>0.

- A partir de la tabla conjunta estimamos P(TCP), P(Grande) y P(TCP,Grande); entonces comprobamos si P(TCP,Grande) ≈ P(TCP)·P(Grande) (o si P(Grande|TCP) ≈ P(Grande)).

Detección de Anomalías

- Regla: marcar anómalo un segundo si $k > \lambda + 3\sqrt{\lambda}$ (con $\lambda=\mu$ y $Var=\lambda$ en Poisson).
- Ver `summary_metrics.csv` (umbral y conteos) y `anomalies.csv` (segundos marcados).

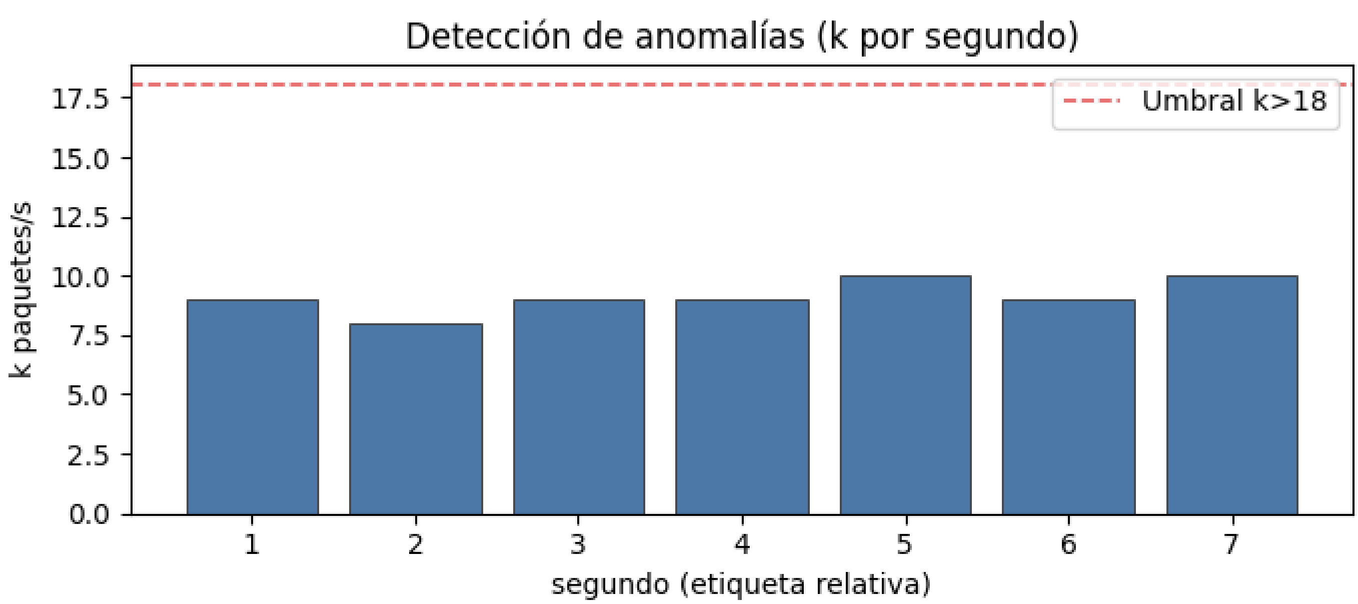
Explicación del resultado actual (ventana 1..7):

- Con $\lambda \approx 9.142857$, el umbral es $k > 18$. En los 7 segundos analizados todos los conteos cumplen $k \leq 18$, por lo que no se marcan anomalías y `anomalies.csv` queda vacío (solo encabezados: `segundo,paquetes_en_segundo`).

Cómo obtener anomalías (escenarios típicos):

- Analizar ventanas que incluyan picos reales (segundos con k muy altos), por ejemplo ampliando el rango temporal.
- Usar un umbral más sensible: reducir z (p. ej., 2 en lugar de 3) en la regla $k > \lambda + z\cdot\sqrt{\lambda}$.
- Cambiar cómo se estima λ: si se incluye todo el rango con segundos en 0, λ baja y también el umbral; así, cualquier segundo con k relativamente alto podría superar el nuevo umbral.
- Alternativa más rigurosa: marcar como anomalía si $P(X\geq k \mid \text{Poisson}(\lambda)) < \alpha$ (p. ej. $\alpha=0.01$); esto es equivalente a usar valores críticos de cola de Poisson.

Figura generada automáticamente (Python):



Conclusiones

Discusión de resultados:

- Discreta (Poisson): En la ventana 1..7 se obtuvo $\lambda \approx 9.142857$ paquetes/seg y un índice Var/Media ≈ 0.052, muy por debajo de 1. Esto sugiere subdispersión, coherente con una ventana corta y conteos relativamente uniformes. Visualmente, la PMF Poisson(λ) superpuesta al histograma muestra un ajuste razonable para el rango observado, aunque la evidencia es limitada por el tamaño muestral.
- Continua (Exponencial): A partir de interarribos, $\lambda \approx 10.5$ 1/s. El KS resulta con p-valor muy bajo, lo que formalmente rechaza la exponencialidad; sin embargo, con pocos datos y resolución a segundos es habitual obtener p-valores pequeños por empates ($\Delta=0$) y discretización. En la práctica, la curva Exp(λ) sirve como primera aproximación y referencia visual.
- Conjuntas (Protocolo×Tamaño): Se estimó $P(TCP)\approx 0.75$ y $P(Grande)=0$. En consecuencia, $P(Grande|TCP)=0$ y la independencia resulta "trivial" por ausencia de paquetes grandes. Este hallazgo está más dominado por la muestra que por una relación estructural entre variables.

Sensibilidad y discrepancias al modificar supuestos:

- Ventana discreta: Si ampliamos la ventana más allá de 1..7 o usamos ventanas móviles, λ puede cambiar (al agregar segundos con 0 llegadas o picos). El índice Var/Media tenderá a acercarse a 1 con más datos si la hipótesis de Poisson es plausible; con ventanas muy cortas puede alejarse más (sub/sobredispersión aparente).
- Interarribos con $\Delta=0$: Ignorar $\Delta=0$ suele aumentar la media de los tiempos y, por ende, disminuir λ estimado. Alternativamente, añadir un jitter pequeño aleatorio para romper empates puede elevar el p-valor del KS, acercando la evidencia hacia la exponencialidad si el proceso latente lo es.
- Umbral de "Grande": Cambiar el umbral (p. ej., 600 o 400 bytes) alterará P(Grande). Si aparecen valores >0, la evaluación de independencia deja de ser trivial; puede observarse dependencia si ciertos protocolos concentran más paquetes grandes.
- Bins y escalas en histogramas: Modificar número de bins o el rango visible puede afectar la percepción visual del ajuste (sobre todo con n pequeño). Para contrastes formales conviene fijar reglas (p. ej., Freedman–Diaconis) y reportar sensibilidad.
- Métricas de anomalía: El umbral $k > \lambda + 3\sqrt{\lambda}$ depende de λ. Cambios en ventana/λ desplazan el umbral y pueden producir más/menos segundos marcados.

En síntesis, los resultados son consistentes con una primera aproximación Poisson–Exponencial, pero están fuertemente condicionados por el tamaño muestral, la resolución temporal y la definición de categorías. Para conclusiones robustas, recomendamos: ampliar muestra, evaluar estabilidad de λ con ventanas móviles, considerar ajustes por discretización en interarribos y complementar con pruebas χ^2/KS con mayor n. Extensiones útiles incluyen modelos con sobredispersión (p. ej., Poisson-Gamma/Negativa Binomial) o alternativas para tiempos (Gamma/Lognormal) si la evidencia empírica lo justifica.

Evidencia numérica (resumen ejecutado):

- λ por conteos (1..7): 9.142857; índice Var/Media: 0.052083.
- Umbral $3\cdot\sqrt{\lambda}$: $k > 18$; anomalías encontradas: 0 (anomalies.csv vacío).
- λ por interarribos: 10.500000; KS exponencial: estadístico 0.905, $p\approx 9.3e-65$ (rechazo con cautela por discretización y n pequeño).

Limitaciones y mejoras:

- Muestra discreta muy corta (7 s): ampliar ventana o usar ventanas móviles para estabilidad.
- Resolución temporal: si hay Δ=0 frecuentes, considerar filtrarlos o añadir jitter controlado y documentarlo.
- Validaciones: complementar con χ^2 (discreto) y KS con mayor n, o ajustar modelos alternativos si hay evidencia de sub/sobredispersión.

Funciones y código utilizado

- `src/data/loaders.py`
  - `read_network_csv`: lee CSV con columnas Timestamp, Packet_Size, Protocol. Soporta formatos con/sin segundos y AM/PM. Devuelve una lista de registros tipados.
- `src/analysis/statistics.py`
  - `group_counts_per_second(timestamps_sec)`: cuenta paquetes por segundo (floor de timestamp).
  - `expand_counts_with_zeros(counts_per_sec)`: construye el vector de conteos por cada segundo del rango, rellenando con 0 los segundos sin llegadas.
  - `estimate_lambda_from_counts(counts_per_sec)`: estima λ como total de llegadas dividido por duración (en segundos).
  - `index_of_dispersion(counts)`: Var/Media (≈1 en Poisson).
  - `interarrival_times(timestamps_sec)`: diferencias consecutivas (s).
  - `estimate_lambda_from_interarrivals(deltas)`: λ = 1/ media(Δ).
  - `poisson_pmf(k, λ)`: e^{-λ} λ^k / k!.
  - `exponential_pdf(x, λ)`: λ e^{-λx} (x≥0).
  - `contingency_protocol_size(protocols, sizes, threshold)`: tabla conjunta Protocolo×Tamaño con P(TCP), P(Grande), P(Grande|TCP) e indicador de independencia.
  - `poisson_anomaly_threshold(λ, z)`: umbral k > λ + z√λ (z=3 por defecto).
- `analysis_cli.py` (script principal de análisis)
  - Parámetros relevantes:
    - `--seconds-range a-b`: analiza solo los segundos relativos [a..b] para la parte discreta.
    - `--excel-table` y `--excel-compact`: exportan tablas listas para Excel.
  - Genera: figuras PNG, CSV de resumen y tablas (counts, histograma Poisson, interarribos, conjuntas, anomalías).
- `generate_report.py`: compone este informe a partir de los CSV generados.

Arquitectura y funcionamiento de la aplicación (GUI):

- `app.py` (Tkinter) ofrece una interfaz para:
  1. Cargar el CSV (Timestamp, Packet_Size, Protocol).
  2. Estimar parámetros: λ (llegadas/seg) y μ (servicios/seg). Si no ingresas media de servicio (ms), μ se estima por heurística (tamaño medio y enlace hipotético).
  3. Simular la cola M/M/1 (`src/sim/queue_mm1.py`) y mostrar métricas: utilización ρ, W/Wq, L/Lq (y sus equivalentes empíricos), según duración y warm-up configurados.
- Flujo típico: Abrir CSV → Estimar parámetros → Ajustar duración/warm-up → Simular → Revisar métricas.

Detalles matemáticos relevantes:

- Poisson: PMF $P(X=k)=e^{-\lambda}\lambda^k/k!$, $E[X]=Var[X]=\lambda$; índice Var/Media≈1.
- Exponencial: $f(t)=\lambda e^{-\lambda t}$, $E[T]=1/\lambda$.
- Independencia TCP/Grande: $P(B|A)=P(B) \Leftrightarrow P(A\cap B)=P(A)P(B)$. En este dataset, $P(\text{Grande})=0 \Rightarrow$ independencia trivial.
- Anomalías: umbral $k > \lambda + 3\sqrt{\lambda}$ ($z=3\sigma$), configurable.

Anexos

- Discreta: counts_per_second.csv, poisson_histogram_table.csv, poisson_histogram_table_compact.csv
- Continua: interarrival_times.csv, exponential_pdf_table.csv
- Conjunta: contingency.csv, contingency_full.csv, contingency_summary.csv