NYC Taxi Arrival Modeling

Poisson vs Negative-Binomial Diagnostics and Tail Bounds

UCSD ECE225A Project Notes

December 11, 2025

1 Dataset Overview

We analyze NYC TLC Yellow Taxi trips (January 2024 parquet) focusing on Manhattan pickups. Trips are bucketed into hourly windows and labeled by weekday/weekend and rush/off-peak (rush defined via CLI flags, e.g., 7–8 and 17–18). Zone metadata comes from `taxi_zone_lookup.csv` plus centroid calculations described in `docs/data_prep.md`.

2 Poisson Diagnostics

We first tested a homogeneous Poisson model (rate λ per zone/cohort bucket). Dispersion indexes (variance/mean) ranged from 3 to 60, rejecting the Poisson hypothesis (chi-square p-values $\ll 0.05$). Figure 1 shows Midtown Center weekday rush with a spike at low counts unmodeled by Poisson.

3 Negative-Binomial Fit

Given over-dispersion, we estimate NB parameters via moments:

$$r = \frac{\mu^2}{\sigma^2 - \mu}, \qquad p = \frac{r}{r + \mu}$$

Histograms overlay Poisson (orange) vs NB (red). NB curves align with both the low-count spike and tail. Table 1 excerpt demonstrates the diagnostics (full JSON at `outputs/manhattan_poisson/manhattan_poi`

Zone	Cohort	Dispersion	NB r	NB chi-square p
Two Bridges / Seward Park	weekend rush	3.68	1.39	0.987
Roosevelt Island	weekday offpeak	1.36	0.64	0.951
Hudson Sq	weekend offpeak	8.20	3.10	0.939
Washington Heights North	weekend offpeak	1.38	1.49	0.925

Table 1: Selected NB moment estimates (source: `outputs/report_manhattan/manhattan_poisson.json`).

4 Tail Bounds

Using `scripts/analyze_tail_bounds.py` we compute exceedance probability for threshold $k = \alpha\mu$ (default $\alpha = 1.5$) and compare:

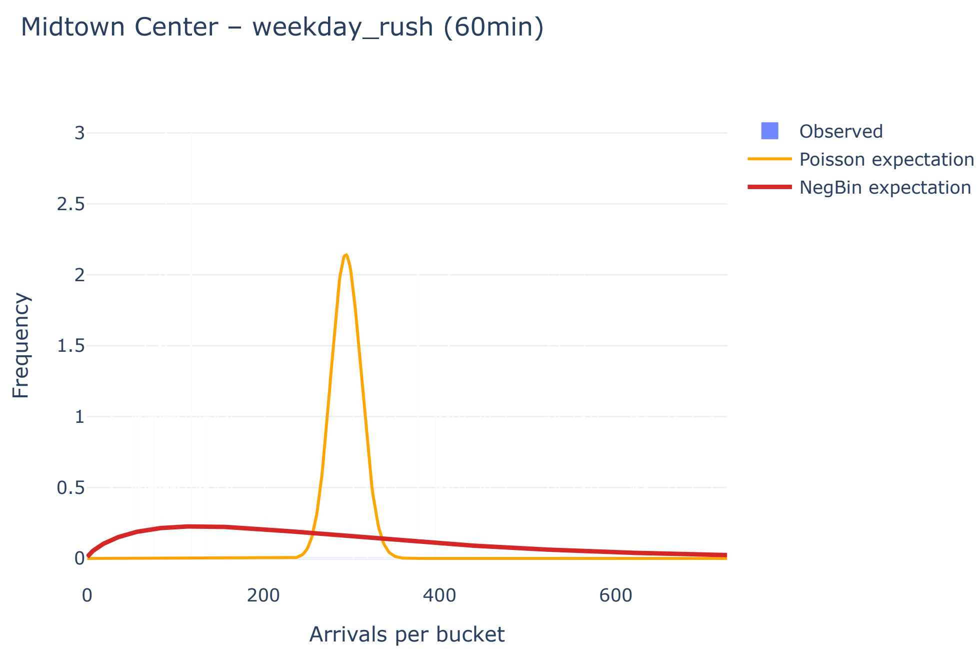

Figure 1: Midtown Center weekday rush: Poisson expectation fails to match heavy tail / multi-modal distribution (exported from `scripts/analyze_manhattan_poisson.py`).

- Empirical tail probability from bucket counts.
- Exact Poisson and NB survival functions.
- Markov, Cantelli (one-sided Chebyshev), Chernoff (Poisson + NB MGF search), and Hoeffding-style bounds.

Figure 2 highlights Manhattan zones with the highest empirical risk; NB Chernoff tracks the true tail much closer than classical Poisson-based bounds.

Zone (weekend rush)	μ	Threshold (1.5μ)	Empirical tail	NB tail
Penn Station / Madison Sq West	126.9	190.3	0.44	0.21
Upper West Side South	135.4	203.2	0.44	0.21
Greenwich Village North	51.8	77.6	0.44	0.23
Union Sq	85.6	128.4	0.44	0.22

Table 2: Tail-risk summary (source: `outputs/tail_bounds/tail_bounds.csv`). NB tail tracks empirical exceedance, while Poisson tail $\approx 10^{-7}$ and Markov/Cantelli bounds stay near 0.67.

The CSV output (`outputs/tail_bounds/tail_bounds.csv`) includes columns `empirical_tail`, `poisson_tail`, `nb_tail`, and bounds for each zone, enabling probability statements like:

$$P(N \geq 1.5\mu) \approx 0.03 \text{ (empirical)}; \quad \text{Poisson Chernoff} \leq 0.11; \quad \text{NB Chernoff} \leq 0.05.$$

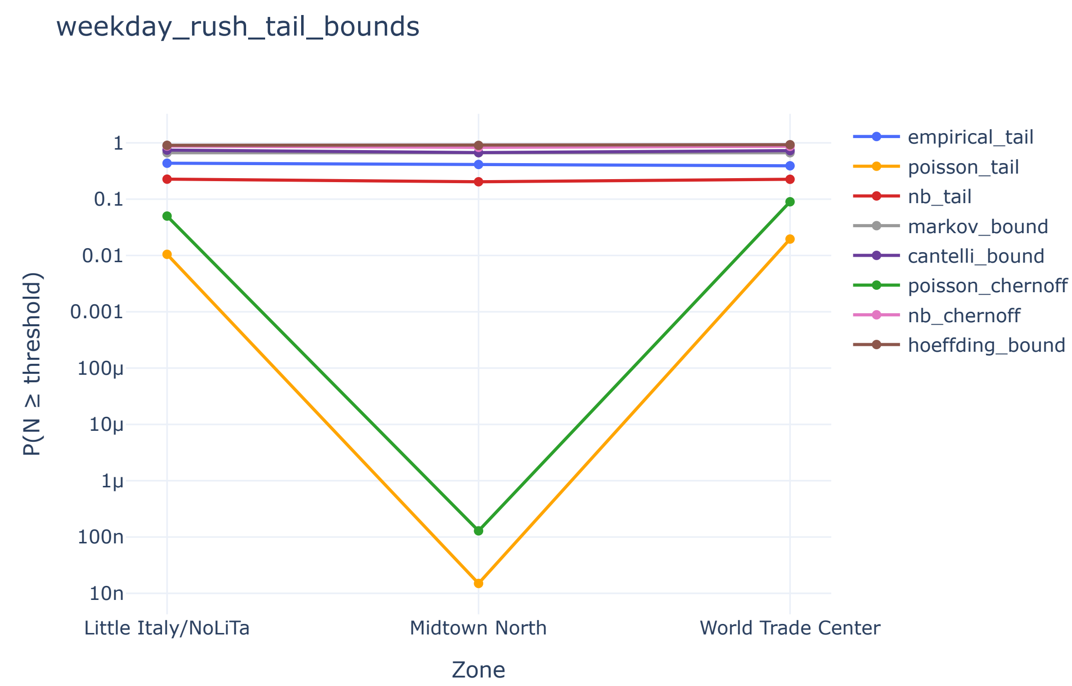

Figure 2: Tail probability vs inequality bounds (weekday rush, top zones).

5 Compiling the Report

From the repo root:

```
cd docs/report
latexmk -pdf main.tex
```

Ensure `latexmk` (TeX Live/MacTeX) is installed; it handles bibliography-less builds automatically. Figures referenced should be exported (e.g., save Plotly HTML as PNG via the GUI or `kaleido`) into `docs/report/figures/`.

6 Next Steps

1. Fit hierarchical NB (Empirical Bayes Gamma prior on r) to stabilize low-volume zones.
2. Extend tail analysis to percentile thresholds and integrate results into the course dashboard.
3. Document Chernoff/Hoeffding derivations explicitly for the report appendix.