Towards Predictive Analytics in Transit:
Thinking Out Loud About Probability and the Bx24
TheOpenBus.com
December 14, 2015

1 Set Up

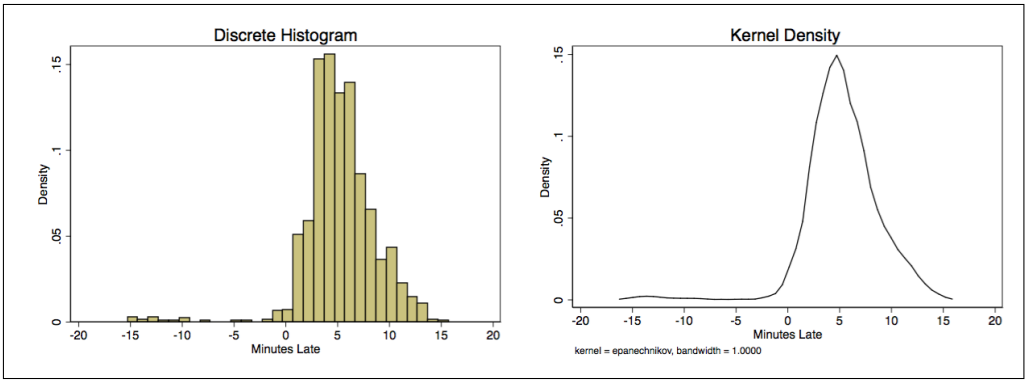
The Bx24 is a local bus route which serves the east side of The Bronx in New York City. It is scheduled to arrive every 30 minutes during day time hours. This report will use arrival time data for a particular stop to investigate the probability characteristics that determine how a hypothetical bus rider interacts with the bus. From a rider's perspective, waiting for a bus constitutes a random process in that the exact arrival of the bus is unknown. This report will attempt to illuminate some non trivial probability questions that would be of interest to a rider.

2 Basic Data Properties

Data was scraped from the Metropolitan Transportation Authority's (MTA) online GPS bus location service. A single stop on the Bx24 route is used for analysis (eastbound at Westchester Ave and Roberts Ave). Data covers every instance of a bus passing this stop from September 1, 2015 - October 31, 2015, excluding weekends, forming a data set of 1,380 observations. Each observation is time stamped to the nearest second.

Official schedule data for the Bx24 is taken from the MTA public records and digitized. Schedule data is combined with scraped data such that every scraped observation is matched to the scheduled arrival time that is closest to its observed time. The difference between the scheduled time and the observed time can then be interpreted to indicate how late or early the bus was relative to the schedule. Figure 1 shows the distribution of early/late times across the 1,380 observations. Figure 1 displays both a discrete histogram, and a smoothed kernel approximation of the density function. Interestingly, the Bx24 is late 98% of the time, suggesting the official schedule is not a best estimate of actual arrival time. On average the bus arrives 5.30 minutes late, and has a standard deviation of *lateness* of 3.39 minutes.

Figure 1: Bx24 Bus - Early/Late Distribution



3 Odds of Catching a Bus

Let us formally consider the central problem facing the rider:

If the number of minutes late the bus arrives is a random variable (X), and assuming the rider also arrives some number of minutes late relative to the schedule (Y), what is probability the rider will meet (catch) the bus? And what will be the wait time the rider should expect to face?

$X \to \mathbb{R}, Y \to \mathbb{R}$. Arriving early is simply a realization of X or Y which is less than 0.

The bus will not wait for a rider to show up, but a rider will be willing to wait some amount of time after arriving at the stop. Let us consider the probability that a rider waits less than 5 minutes for the bus.

First consider a simple, stylized case as a baseline in which the density with which the rider shows up ($\mathrm{g}(y)$) is distributed uniformly ranging from 5 minutes early to 5 minutes late, and centred at the scheduled time. The density of the bus arrival ($\mathrm{f}(x)$) is distributed uniformly between 5 minutes early and 10 minutes late. The event of interest is then every instance in which $X \geq Y$ and $X \leq Y + 5$. This event is represented by the shaded area in Figure 2.

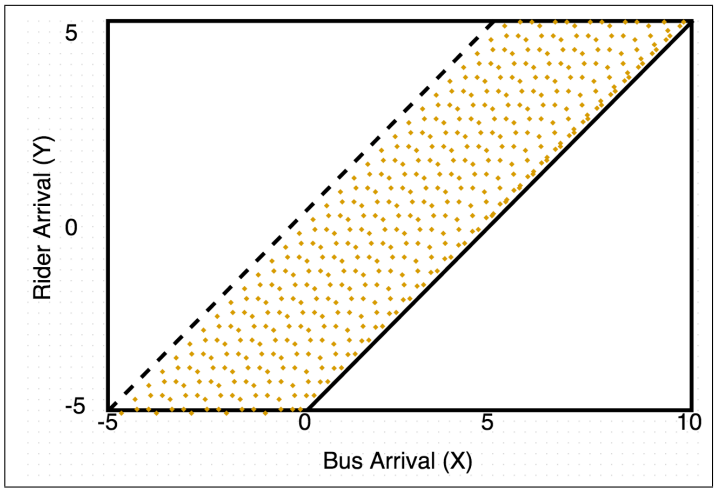
Given that both distributions are assumed uniform, the probability of any two equally sized areas in the box of Figure 2 have equal probability.

$\Rightarrow \mathbb{P}[\text{wait} < 5\text{mins}] = [(10 * 15) - \frac{10*10}{2} - \frac{10*10}{2}]/(10 * 15) = \frac{1}{3}$

So under uniform distribution assumptions the probability of catching the bus in under 5 minutes is 33.3%.

Formally the probability is calculated by taking a double integral over a random vector: $\int_{-5}^{5}\int_{-5}^{10} \mathrm{h}(x,y)dxdy = 0.333$

Figure 2: Bus/Rider Meeting Problem



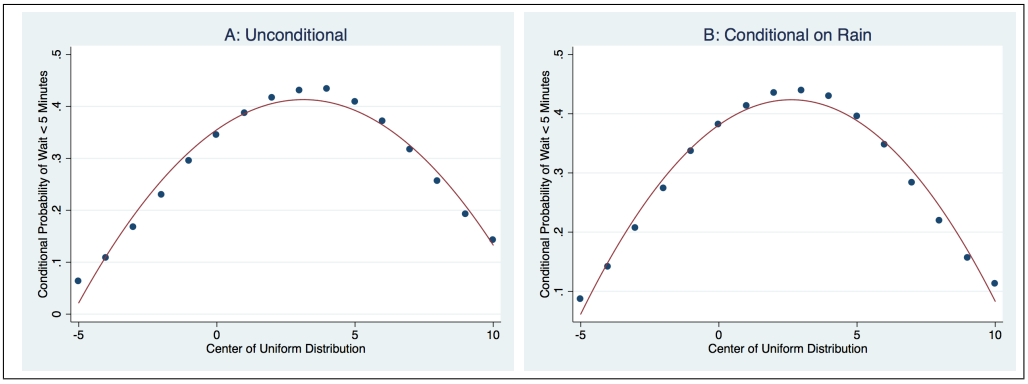
However the assumption that the bus arrival distribution (f(x)) is uniform and bounded between -5 and 10 is unnecessary; we have an estimate of this density (Figure 1). By integrating over the observed distribution of X rather than a uniform distribution, the likelihood of waiting less than 5 minutes given Y∼Unif[-5,5] is 34.4%. However, unlike under the uniform (f(x)), the probability of catching the bus under the estimated (f(x)) depends continuously on the realization of Y.

If we now assume the rider is aware of the distribution of early/late times, but maintain the assumption that Y is restricted to be uniform with a range of 10, it is notable that 'aiming' to arrive at the stop for the scheduled time is not optimal. Figure 3A shows the probability of waiting less than 5 minutes for a bus, given Y∼Unif with a range of 10, around different centre points. The optimal strategy is to aim to be at the bus stop 4 minutes after the scheduled time, when there is a 43.3% chance of waiting less than 5 minutes.

4 Conditioning

For any prospective trip the rider faces the random variable X, the realization of X being the number of minutes late the bus arrives at the stop. This section

Figure 3: Optimal Strategy / Optimal Center of Uniform Y



will explore an observed variable that the rider could potentially condition on, when forming his expectation of X.

As mentioned above the $\mathbb{E}[X] = 5.30$

Now suppose, before the rider leaves for the bus stop they observe the weather (ω), which is the input of an indicator variable R :

$$R(\omega) = \begin{cases} 1 \text{ if raining} \\ 0 \text{ otherwise} \end{cases}$$

The rider can then adjust their $\mathbb{E}[X]$ given this new information.

Daily weather data for New York City was taken from the Weather Underground online service allowing the variable R to be generated for all days in the data set. A simple univariate regression of the form: $X = \beta_0 + \beta_1 R + \epsilon$, estimates a marginally significant $\beta_1 = -0.61$, suggesting the average bus arrived 0.61 minutes earlier on a day it was raining. Possibly the rain reduces the number of people taking the bus, which makes it easier for a driver to keep on schedule.

The conditional expectation can be computed:

$\mathbb{E}[X|R = 1] = \int x f_{X|R}(x|r)dx = 4.79$

Considering the riders problem, if Y is again assumed $\sim$Unif[-5,5], the conditional probability of waiting fewer that 5 minutes is:

$\Rightarrow \mathbb{P}[\text{wait} < 5\text{mins}|R = 1] = 38.1\%$

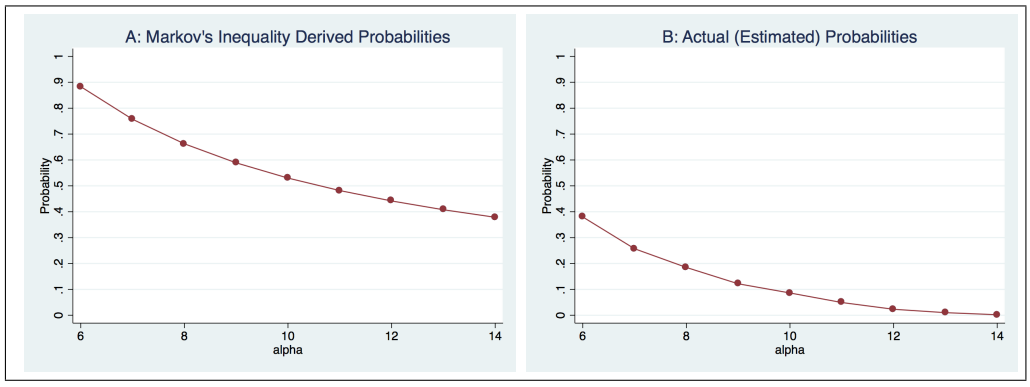
Recall without conditioning $\mathbb{P}[\text{wait} < 5\text{mins}] = 34.4\%$

The presence of rain therefore improves the prospects of a rider with the naive strategy of 'aiming' to arrive at the officially scheduled time. Such a

rider will have a 3.7 percentage point increase in their odds of waiting less than 5 minutes if it is raining.

We could now consider if the optimal strategy (the strategy which maximizes the probability of waiting less than 5 minutes for a bus) is altered by the conditioning variable R. Figure 3B displays the graph of $\mathbb{P}[\text{wait}< 5\text{mins}|R = 1]$ for Y$\sim$Unif, with range 10, but different centers. The highest probability occurs from centring the distribution of Y at 3 minutes. This means the optimal strategy for a rider that observes rain ($R = 1$) is to aim for an arrival at the stop 3 minutes late and would face a 43.9% chance of catching a bus within 5 minutes. Calculated in a similar way: if $R = 0$ the optimal strategy is to aim to arrive 4 minutes late, which results in the rider facing a 44.3% chance of catching a bus within 5 minutes.

Figure 4



5 Missing by Markov

An alternative motivating problem for a rider may be to minimize the probability that they miss a particular bus. Here it is assumed that $\mathbb{E}[X]$ can be communicated to the rider, but the full distribution is too complex to communicate. Assuming a rider needs to catch a specific bus, and is willing to wait an unlimited amount of time, what is the probability they will miss the bus given their arrival time at the stop?

This problem is illuminated by Markov's Inequality (here the 2% of observations for which $X < 0$ are omitted). The rider knows that the bus is

normally 5.30 minutes late, $\mathbb{E}[X] = 5.30$. The probability that the bus is more than α minutes late follows:

$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$

For example if the rider shows up 6 minutes late ($\alpha = 6$):

$\mathbb{P}(X \geq \alpha) \leq \frac{5.30}{6} = 0.883$

Meaning the probability of catching the bus is no more than 88.3%. The $\mathbb{P}(X \geq \alpha) \leq \frac{5.30}{\alpha}$ is plotted for different α values in Figure 4A.

Figure 4B plots the actual proportion of buses that arrived after the same α thresholds. Markov's Inequality is able to provide very little predictive power in this context. For example the estimated probability of catching the bus if arriving 6 minutes late is estimated to be far less than the upper bound of 88.3%; at only 38.0%.