A High-Volume Call Center with a Random Arrival Rate Function and Nonexponential Call Durations

Christos Alexopoulos, Dave Goldsman, and Byeong-Yun Chang

Georgia Tech

Highlights

- The call center operates 24 hours/day (except for a 15-minute maintenance period at 4 a.m.).
- It takes inbound calls primarily.
- Calls come from 3 time zones in the continental U.S.
- The call volume is high (about 30,000/day), and is increasing fast.
- The durations of the calls
  - are typically quite short (about 35-40 seconds)
  - depend on the location of the caller (Southern callers take a bit longer to communicate)
  - depend on the time in the day and day of week.
- There in no routing of calls (good).
- At the beginning, we are not worrying about multi-skilled agents.

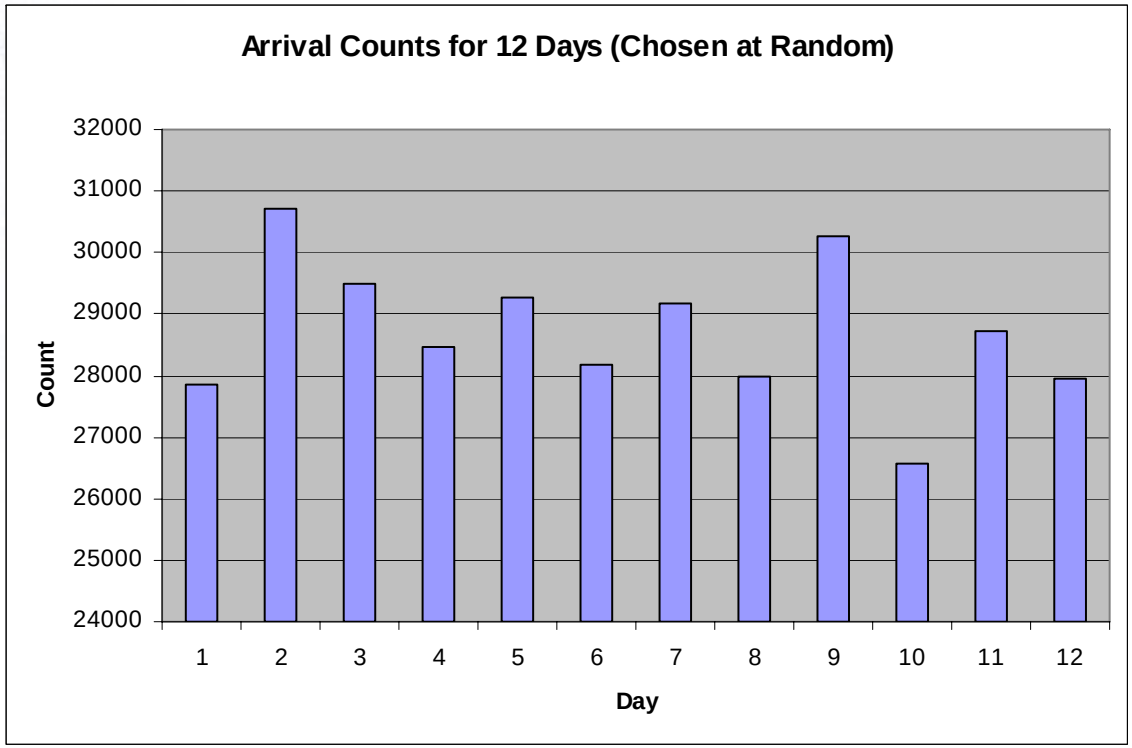
Highlights (continued)

- The arrival process has all the bad properties listed in Avramidis, et al. (2004) and Brown et al. (2005):
  1. The total daily volume has overdispersion relative to the Poisson distribution (the variance is much greater than the mean).
  2. The arrival rate varies considerably with the time of the day.
  3. There is significant correlation between arrival counts in a time partition of the day.
  4. There is correlation between call volumes on successive days.

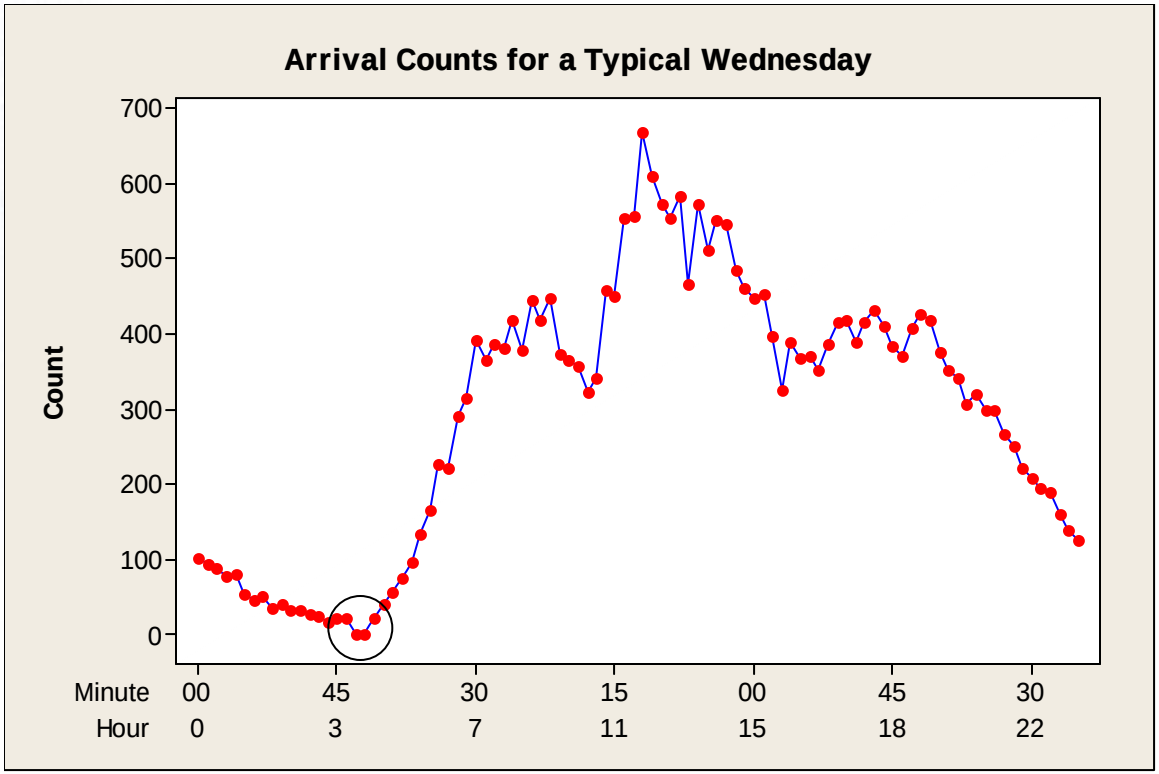
Property 1

Sample Mean = 28,719

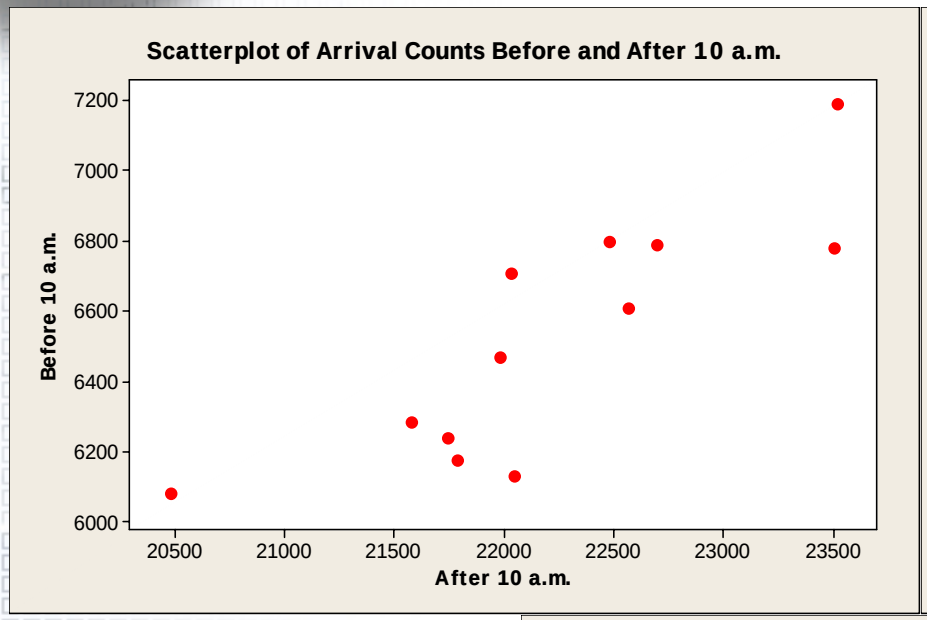
Sample Variance = 1,307,632 (St. Dev. = 1,143)

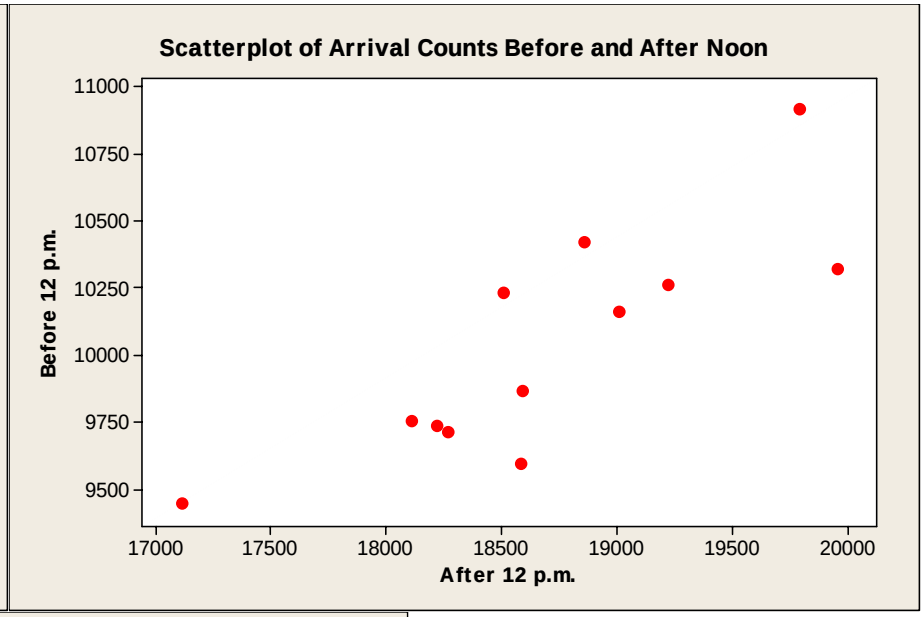


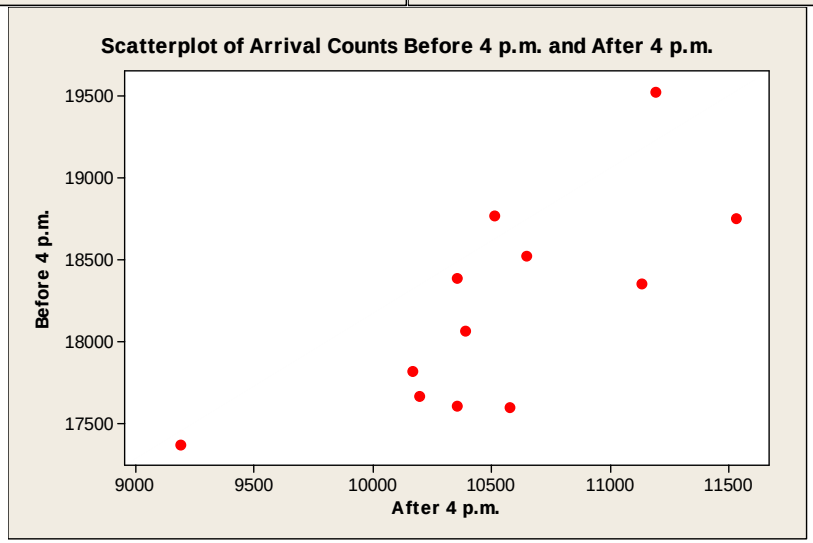
Property 2


Arrival Counts for a Typical Wednesday
Count
700
600
500
400
300
200
100
0
Minute 00 45 30 15 00 45 30
Hour 0 3 7 11 15 18 22

Property 3


Scatterplot of Arrival Counts Before and After 10 a.m.
Before 10 a.m.
7200
7000
6800
6600
6400
6200
6000
20500
21000
21500
22000
22500
23000
23500
After 10 a.m.


Scatterplot of Arrival Counts Before and After Noon
Before 12 p.m.
11000
10750
10500
10250
10000
9750
9500
17000
17500
18000
18500
19000
19500
20000
After 12 p.m.


Scatterplot of Arrival Counts Before 4 p.m. and After 4 p.m.
Before 4 p.m.
19500
19000
18500
18000
17500
9000
9500
10000
10500
11000
11500
After 4 p.m.

Highlights (continued)

- In addition:

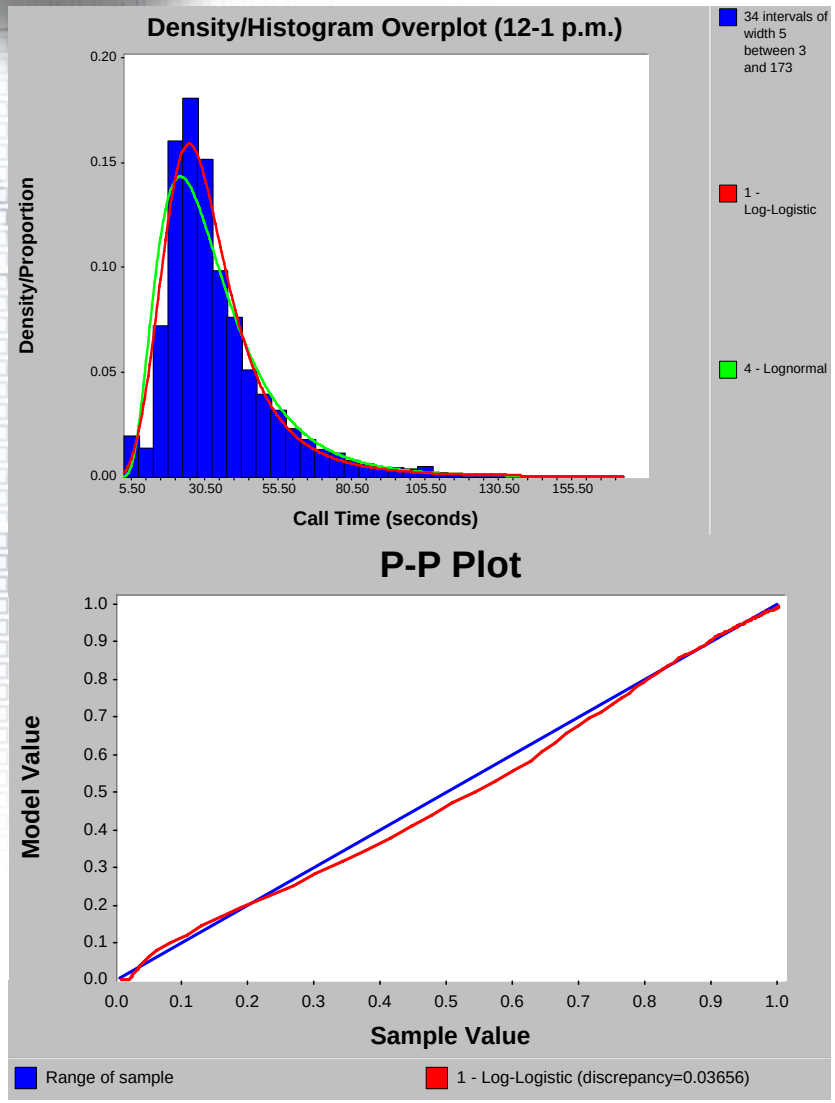
1. The calling population changes due to the addition and subtraction of call points.
2. We observe callers after a potential queueing delay at each point of origin.

Highlights (continued)

Service requirements:

- $\Pr\{\text{Wait} \leq T_j\} \geq 0.98$.
- T_j depends on the time interval j, and varies from 3 to 12 seconds.
- *Failovers* over 15 secs are unacceptable.

Call Times (Log-logistic)



Anderson-Darling Test with Log-Logistic Model

Sample size	250
Test statistic	0.30248
Note:	The following critical values are exact.

Critical Values for Level of Significance (alpha)

Sample Size	0.250	0.100	0.050	0.025
	0.010	0.005		
250	0.426	0.562	0.659	0.768
	0.905	1.009		

Reject? No

Kolmogorov-Smirnov Test with Log-Logistic Model

Sample size	250
Normal test statistic	0.03878
Modified test statistic	0.61311
Note:	The following critical values are exact.

Critical Values for Level of Significance (alpha)

Sample Size	0.100	0.050	0.025	0.010
50	0.708	0.770	0.817	0.873
infinity	0.715	0.780	0.827	0.886

Reject? No

First (Naïve) Approach

- Assume a NHPP arrival model.
- Optimize *M/M/N* systems over 15-minute periods.
  - Collect data from the central database.
  - Create a forecast of the arrival rate function and the call times for each day of next week separately.
  - Compute an initial guess at staffing level using the square root rule. (Such a rule is usually conservative.)
  - Use a simulation-based search to find the number of agents that satisfies the service requirements. The simulation model was built with Simkit, a Java-based DES package (available from http://diana.gl.nps.navy.mil/Simkit/).

Benefits from First Approach

- Staffing task can be done much more quickly than before.
- Users can run “worst-case” scenarios to test robustness of particular staffing levels.
- Users can see the consequences of new business growth in the future. More stores translate into:
  - Much more efficiency in terms of server utilization.
  - Better performance in terms of contracted metrics.

Estimation of the Arrival Rate Function from Event Count Data

- We collect call arrivals during an interval $(0, S]$ (e.g., $S = 1440$ minutes).
- Suppose that we collect call arrivals over the time interval $(0, S]$ for k days.
- Partition $(0, S]$ into m subintervals

$$(a_0, a_1], (a_1, a_2], \ldots, (a_{m-1}, a_m] \quad (a_0 = 0, a_m = S).$$

- For example, we could use 15-minute intervals.
- o_j = observed number of calls in $(a_{j-1}, a_i]$, $j = 1, \ldots, m$, over all k realizations.
- The estimate of the arrival rate function $\lambda(\cdot)$ is

$$\tilde{\lambda}(t) = \frac{o_j}{k(a_j - a_{j-1})} \quad \text{for } a_{j-1} < t \leq a_j;\ j = 1, \ldots, m.$$

- The following is an alternative estimator for the mean-value function:

$$\tilde{\Lambda}(t) = \left(\sum_{i=1}^{j-1} \frac{o_i}{k} \right) + \frac{o_j(t - a_{j-1})}{k(a_j - a_{j-1})} \quad \text{for } a_{j-1} < t \leq a_j;\ j = 1, \ldots, m.$$

Generation of Arrivals Based on Event Count Data: Notation

The following algorithm uses the next-event approach, which schedules the next arrival when the current arrival is processed.

- $T =$ time of current call arrival.
- $E \sim$ Exponential(1).
- The algorithm returns
  - $\tilde{\Lambda}^{-1}(\tilde{\Lambda}(T) + E)$ as the time of the next call arrival
  - -1 if no further arrivals are generated (we have exceeded time S).
- $f_j = o_j/k$, $j = 1, \ldots, m$.
- $F_j = \sum_{i=1}^{j} f_i = F_j - F_{j-1}$, $j = 1, \ldots, m$. $(F_0 = 0)$
- CumRate.Now $= \tilde{\Lambda}(T)$.
- CumRate.New $= \tilde{\Lambda}(T) + E$.
- CumRate = value of cumulative rate function at the right endpoint of the interval associated with the next arrival time.

Generation of Arrivals Based on Event Count Data

Max $\leftarrow F_m$
$j \leftarrow 1$
while $(T > a_j)$
    $j \leftarrow j + 1$
endwhile
CumRate.Now $\leftarrow F_{j-1} + o_j(T - a_{j-1})/(k(a_j - a_{j-1}))$
CumRate $\leftarrow F_j$
Generate $U \sim$ Uniform$(0, 1)$
CumRate.New $\leftarrow$ CumRate.Now $+ E$
if (CumRate.New $\leq$ Max) **then**
    while (CumRate.New $>$ CumRate)
        $j \leftarrow j + 1$
        CumRate $\leftarrow$ CumRate $+ f_j$
    endwhile
    return $a_j - (\text{CumRate} - \text{CumRate.New})(a_j - a_{j-1})/f_j$
else
    return -1
endif

Facts

- As $k \to \infty$, the estimator $\tilde{\Lambda}(t)$ converges to the actual function *only* at the endpoints of the intervals.
- The estimate $\tilde{\Lambda}(t)$ is zero intervals containing no observations. In this case, no arrivals will be generated in those intervals. This is convenient for modeling breaks.
- An alternative method uses all observed arrival times (Leemis 2001).

Simple Staffing Assignment

- This approximation was designed for the Markovian $M/M/N$ system.
- λ = arrival rate for calls in a sufficiently long time interval.
- $E[S]$ = mean holding (call) time in this interval.
- $R = \lambda E[S]$ (offered load).
- W = waiting time (delay) of a typical call in steady-state.
- If R and the number of agents grow according to the relationship $N = R + \beta\sqrt{R}$, then the probability that a call will wait is

$$\Pr\{W > 0\} \approx \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1},$$

$$E[W] \approx \Pr\{W > 0\}\frac{E[S]}{\beta\sqrt{R}},$$

$$\Pr\{W > T\} \approx \Pr\{W > 0\}e^{-\beta\sqrt{R}T/E[S]},$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the density and cdf of the standard normal distribution.

A Simple Staffing Assignment (continued)

- Suppose we wish to have

$$\Pr\{W \leq T\} \geq 1 - \epsilon,$$

e.g., $\Pr\{W \leq 12 \text{ sec}\} \geq 0.98$.

- We solve (numerically)

$$\left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1} e^{-\beta\sqrt{R}T/E[S]} = \epsilon$$

for β, and then set

$$N = \left\lceil R + \beta\sqrt{R} \right\rceil.$$

Forecasting Daily Arrivals

We tested two models from Avramidis et al. (2004):

★ **Model 1** (Whitt 1999): $\Lambda(t) = W\lambda(t)$; $W \sim \text{gamma}(\gamma, 1)$.

shape parameter

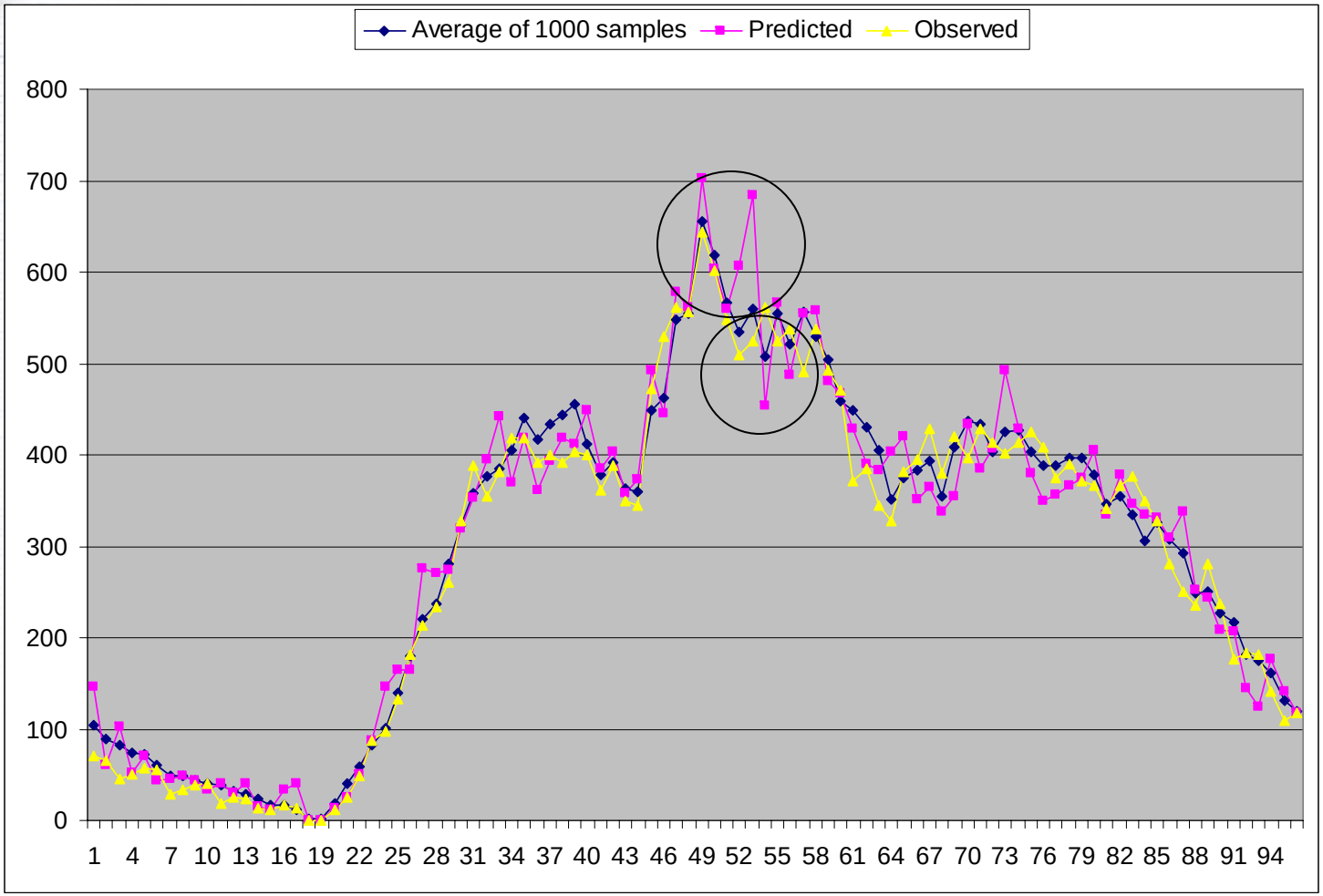
- X_i = arrival count for interval $[t_{i-1}, t_i)$, $i = 1, \ldots, 96$.
- $Y = \sum_{i=1}^{96} X_i$ = total daily count.
- $\lambda_i = \int_{t_{i-1}}^{t_i} \lambda(t)\, dt$.
- $\mathbf{X} = (X_1, \ldots, X_{96})$.
- $\mathbf{X} \sim$ negative multinomial$(\gamma, \lambda_1, \ldots, \lambda_{96})$.
- Facts:
  - $\text{Corr}(X_i, X_j) \geq 0)$.
  - This model yields distributional properties for the remaining demand given the demand observed up to a certain point — good for short term forecasts.

Forecasting Daily Arrivals (continued)

★ **Model 3**:

- $Y = \sum_{i=1}^{96} X_i$ = total daily count $\sim$ 3-parameter gamma.
- $Q_i = X_i/Y$, $i = 1, \ldots, 96$.
- $\mathbf{Q} = (Q_1, \ldots, Q_{96}) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_{96})$, and independent of Y.
- $\tilde{\mathbf{X}} = Y\mathbf{Q}$.
- $\mathbf{X} = [\tilde{\mathbf{X}}]$.
- This model allows negative correlations between X_i and X_j.
- We used counts for the same day in 3 weeks to estimate the parameters of the gamma and Dirichlet distributions.

Arrival Counts for a Wednesday (Model 3)



What's Next?

- We are examining daily arrival process models as more data are collected.
- We are looking at models for call volumes over different days.
- We are looking at flexible staffing assignments.
- We are looking more closely at service time distributions and how they vary from day to day (as well as hour to hour).

My Dream!

It's not just what call centers can do for us…

It's what we can do for call centers!

A Few References

Avramidis, A.N., A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a call center. *Management Science* 50(7):896-908, 2004.

Brown, L., N. Gans, A. Mandelbaum, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: a queueing perspective. *JASA* 100(469):36-50, 2005.

Gans, N., G. Koole, and A. Mendelbaum. Telephone call centers: Tutorial, Review, and research prospects. *Manufacturing and Service Operations Management* 5(2):79-141, 2003.

Halfin, S., and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29:567-587.

Henderson, S.G. Estimation of nonhomogeneous Poisson processes from aggregated data. *Operations Research Letters* 31:375-382, 2003.

Jennings, O.B., A. Mandelbaum, W.A. Massey, and W. Whitt. Server staffing to meet time-varying demand. *Management Science* 42(10)1383–1394, 1996.

Leemis, L.M. Nonparametric estimation of the cumulative intensity function for a nonhomoegenous Poisson process. *Management Science* 37:866–900, 1991.

Leemis, L.M. Nonparametric estimation and variate generation for a nonhomoegenous Poisson process from event count data. *IIE Transactions* 36:1155–1160, 2004.

Testik, M.C., J.K. Cochran, and G.C. Runger. Adaptive server staffing in the presence of time-varying arrivals: a feed-forward control approach. *Journal of the Operational Research Society* 55:233–239, 2004.

Whitt, W. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* 24:205-212.