Course Review: **Introductory Part.**

- **Introduction to Services and Queues (Service Nets = Queueing Nets)**
  
  Our Service Economy.
  Tele-Services (Telephone, Internet, email, Fax, Chat).
  Queues in service systems are here to stay (at least for a while).
  **Operational Queues:** Perpetual, Predictable, Stochastic.

- **Measurements:** The First Prerequisite
  
  Transaction-based (time-based) measurements.
  Face-to-Face, Telephone, Transportation, Internet, Administrative Services.
  Scenario Analysis (vs. Simulation or Analytical Models): very typical or rare event.

- **Models:** The Second Prerequisite
  
  The Skeptic (Flanders).
  vs. The Believer/Practitioner (Larson, our class).

- **The Fluid View; A Deterministic Service-Station**
  
  Averaging over many (similar enough) scenarios.
  Capacity/Bottleneck analysis (via spreadsheet,LP).
  Utilization Profiles for resources.
  Inventory Buildup Diagrams (via “National Cranberry” HBS case).

- **The Processing Network Paradigm**
  
  TQM (80’s), continued by BPR (90’) = Business Process ReEngineering.
  Dynamic Stochastic Project/Processing Networks (DSP-nets = DS-PERTs).
  Applications: Arrest-to-Arraignment, Israeli Electric Company, Multi-Project Management;
  Operational Q’s: scarce resources; synchronization/coordination gaps, design constraints.
  Q1: Can we do it? via Bottleneck Analysis ↔ the fluid view.
  Q2: How long will it take? typically via stochastic networks.
  Q3: Can we do better? via parametric/sensitivity/what-if analysis.
  Q4: How much better? via optimality/approximation analysis.

- **Towards modelling a Stochastic Service Station:** the main building blocks
  
  **Arrivals’ epochs:** Poisson = the model for completely random arrivals.
  **Service durations:** within the Phase-type framework.
  **Customers’ patience**
Service Engineering

Class 6

Modeling Arrivals to a Service Station: The Poisson Process, and Relatives.

- Empirical Introduction, via DataMOCCA.
  - PASTA = Poisson Arrivals See Time Averages.
  - Biased Sampling.
- Animation: from Bernoulli to Poisson, or The Law of Rare Events.
- Non-homogeneous Poisson Processes.
- Testing: Poisson or not Poisson.
- Modeling Arrivals to a Service Station.
- Forecasting of the Arrival Rate.
- Poisson Alternatives: eg. Internet Applications (Heavy Tails, Long-Range Dependence).
- On Limits Theorems in Probability: SLLN, CLT, Rare Events.
Arrivals to a Call Center (Israel, 1999): Time Scales

Strategic

Yearly

Tactical

Monthly

Operational (Predictable Var.)

Daily

Regulatory (Stochastic)

Hourly
Arrivals to a Call Center (U.S., 1976): Queueing Science

Arrival Process, in 1976

(E. S. Buffa, M. J. Cosgrove, and B. J. Luce, “An Integrated Work Shift Scheduling System”)

Yearly

Monthly

Daily

Hourly

Figure 1: Typical distribution of calls during the busiest hour for each week during a year.

Figure 2: Daily call load for Long Beach, January 1972.

Figure 3: Typical half-hourly call distribution (Bundy D. A.).

Figure 4: Typical intrahour distribution of calls, 10:00-11:00 A.M.
## Monthly Arrivals to Service

### U.S. Bank: Daily Arrival-Rates, over a Month, 2002

<table>
<thead>
<tr>
<th>Month</th>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td><img src="image1.png" alt="Graph" /></td>
<td><img src="image2.png" alt="Graph" /></td>
<td><img src="image3.png" alt="Graph" /></td>
<td><img src="image4.png" alt="Graph" /></td>
<td><img src="image5.png" alt="Graph" /></td>
<td><img src="image6.png" alt="Graph" /></td>
<td><img src="image7.png" alt="Graph" /></td>
<td><img src="image8.png" alt="Graph" /></td>
<td><img src="image9.png" alt="Graph" /></td>
<td><img src="image10.png" alt="Graph" /></td>
<td><img src="image11.png" alt="Graph" /></td>
<td><img src="image12.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

---

Note: The graphs above depict the daily arrival-rates for each month from September 2002 to December 2002.
Intraday Arrival-Rates (per hour) to Call Centers

December 1995 (700 U.S. Helpdesks)

May 1959 (England)

November 1999 (Israel)

Observation: Peak Loads at 10:00 & 15:00
Arrivals to queue
September 2001

Number of cases


Time (Resolution 30 sec.)

Stochastic Variability
USBank Arrivals to queue

Time (Resolution 30 sec.)

Number of cases


Scatter vs. Polygon (in SEE Stat)

Observations: stochastic variability
Smoothing in SSEEStat (Automatic)

USBank Arrivals to queue

Time (Resolution 2 min. 30 sec.)

Number of cases

Smoothing: sophisticated averaging
Predictable vs. Stochastic Variability

\[ \lambda(t) = \Lambda \times \lambda_0(t) \Rightarrow \Lambda: \text{Stochastic var.} \quad \lambda_0(t): \text{Predictable (shape)} \]
30 sec → 1 hour: averaging

Arrivals to queue
September 2001

Predictable var. (Heavy Tuesday)
Stochastic var.

Time (Resolution 60 min.)
Number of cases

Arrivals to queue
September 2001

Shape of a Tuesday (Predictable)

\[
N(t) = \frac{d(t)}{\int_{0}^{T} d(u) \, du} = \frac{1}{T} \% \text{ of mean} \quad (t = \frac{d(t)}{\sum d(u)})
\]
Second peak at 19:00 (vs. 15:00 in call centers).

How much stochastic variability?
Arrivals to ED: Environment Dependence

Large Israeli ED, 2005-6

HomeHospital Patients Arrivals to ED Department
Week days

Time (Resolution 60 min.)

Average number of cases

Graph showing the number of HomeHospital patients arriving to the ED department on weekdays, with data from April 2005 to September 2006.
Arrivals to ED: Environment Dependence

Number of Arrivals

Home Hospital Patients Arrivals to ED Department
All days

Average number of cases

Percent to Mean

Home Hospital Patients Arrivals to ED Department
All days
Predicting Emergency Department Status

Houyuan Jiang†, Lam Phuong Lam†, Bowie Owens†, David Sier† and Mark Westcott†

† CSIRO Mathematical and Information Sciences, Private Bag 10, South Clayton MDC, Victoria 3169, Australia
‡ The Judge Institute of Management, University of Cambridge, Trumpington Street, Cambridge CB2 1AG, UK

Abstract

Many acute hospitals in Australia experience frequent episodes of ambulance bypass. An important part of managing bypass is the ability to determine the likelihood of it occurring in the near future.

We describe the implementation of a computer program designed to forecast the likelihood of bypass. The forecasting system is designed to be used in an Emergency Department. In such an operational environment, the focus of the clinicians is on treating patients, there is no time carry out any analysis of the historical data to be used for forecasting, or to determine and apply an appropriate smoothing method.

The method is designed to automate the short term prediction of patient arrivals. It uses a multi-stage data aggregation scheme to deal with problems that may arise from limited arrival observations, an analysis phase to determine the existence of trends and seasonality, and an optimisation phase to determine the most appropriate smoothing method and the optimal parameters for this method.

The arrival forecasts for future time periods are used in conjunction with a simple demand modelling method based on a revised stationary independent period by period approximation queueing algorithm to determine the staff levels needed to service the likely arrivals and then determines a probability of bypass based on a comparison of required and available resources.

1 Introduction

This paper describes a system designed to be part of the process for managing Emergency Department (ED) bypass. An ED is on bypass when it has to turn away ambulances, typically because all cubicles are full and there is no opportunity to move patients to other beds in the hospital, or because the clinicians on duty are fully occupied dealing with critical patients who require individual care.

Bypass management is part of the more general bed management and admission–discharge procedures in a hospital. However, a very important part of determining the likelihood of bypass occurring in the near future, typically the next 1, 4 or 8 hours, is the ability to predict the probable patient arrivals, and then, given the current workload and staff levels, the probability that there will be sufficient resources to deal with these arrivals.

Here, we consider the implementation of a multi-stage forecasting method [1] to predict patient arrivals, and a demand management queueing method [2], to assess the likelihood of ED bypass.

The prototype computer program implementing the method has been designed to run on a hospital intranet and to extract patient arrival data from hospital patient admission and ED databases. The program incorporates a range of exponential smoothing procedures. A user can specify the particular smoothing procedure for a data set or to configure the program to automatically determine the best procedure from those available and then use that method.

For the results presented here, we configured the program to automatically find the best smoothing method since this is the way it is likely to be used in an ED where the staff are more concerned with treating patients than configuring forecast smoothing parameters.
For the optimisation we assume no a priori knowledge of the patient arrival patterns. The process involves simply fitting each of the nine different methods listed in Table 1 to the data, using the mean square fitting error, calculated using (3), as the objective function. The smoothing parameters for each method are all in (0, 1) and the parameter solution space is defined by a set of values obtained from an appropriately fine uniform discretization of this interval. The optimal values for each method are then obtained from a search of all possible combinations of the parameter values.
From the data aggregated at a daily level, repeat the procedure to extract data for each hour of the day to form 24 time series (12am–1am, 1am–2am, . . . , 11pm–12am). Apply the selected smoothing method, or the optimisation algorithm, to each time series and generate forecasting data for those future times of day within the requested forecast horizon. The forecast data generated for each time of day are scaled uniformly in each day in order to match the forecasts generated from the previously scaled daily data.

**Output:** Display the historical and forecasted data for each of the sets of aggregated observations constructed during the initialisation phase.

The generalisation of these stages is straightforward. For example, if the data was aggregated to a four-weekly (monthly) level, then the first scaling step would be to extract the observations from the weekly data to form four time series, corresponding to the first, second, third and fourth week of each month. Historical data at timescales of less than one day are scaled to the daily forecasts. For example, observations at a half-hourly timescale are used to form 48 time series for scaling to the day forecasts.

### 4.3 Output from the multi-stage method

Figures 2 and 3 show some of the results obtained from using the multi-stage forecasting method to predict ED arrivals using the 60 weeks of patient arrival data described in Section 3. The forecasted data were generated from an optimisation that used the multi-stage forecasting method to minimise the residuals of (3) across all the smoothing methods in Table 1.

![Figure 2: Hourly historical and forecasted data 25/7/2002–31/7/2002](image_url)

![Figure 3: Four-hourly historical and forecasted data 25/7/2002–31/7/2002](image_url)
Arrival Patterns, Israeli Telecom, 2005

Arrivals to queue

Time (Resolution 30 min.)

Average number of cases

Percent to mean

Averaging (30 min) : over time
Aggregation (All Sundays) : over days

% to Mean
Mondays (Busiest) and Thursdays (Lightest), 2005

Arrivals to queue

Time (Resolution 30 min.)

January 2005 Mondays
January 2005 Thursdays
February 2005 Mondays
February 2005 Thursdays
March 2005 Mondays
March 2005 Thursdays
April 2005 Mondays
April 2005 Thursdays

Arrivals to queue

Time (Resolution 30 min.)

January 2005 Mondays
January 2005 Thursdays
February 2005 Mondays
February 2005 Thursdays
March 2005 Mondays
March 2005 Thursdays
April 2005 Mondays
April 2005 Thursdays
Mondays, 2004-5 (Averages)

Arrivals to queue
Mondays

Average number of cases

Time (Resolution 30 min.)

Percent to mean

Nov-04  Dec-04  Jan-05  Feb-05  Mar-05  Apr-05  May-05  Jun-05  Jul-05  Aug-05
Sep-05  Oct-05  Nov-05  Dec-05
Mondays, 2005 (Individual Days)

Arrivals to queue

Number of cases

Time (Resolution 30 min.)

Percent to mean

Why Model Arrivals? (model, building-block)

Predict

Forecast Performance: Example

US Bank: Forecast “Performance”
(Weinberg, Brown, Stroud, 2005)

Wider confidence intervals for number of calls.

Narrower confidence intervals for arrival rate
(Poisson parameter).

Note: staffing models require an arrival rate as input.
Within-Day Updating

Comparison between Day-Ahead and Within-Day Predictions
(Weinberg, Brown, Stroud, 2005)

![Predicted Volumes for Tuesday 09/02/03](image)

**Conclusion:** Morning information is important but no significant difference between 10am and 12am.
THE BEST LINEAR UNBIASED ESTIMATOR FOR CONTINUATION OF A FUNCTION

BY YAIR GOLDBERG*, YA'ACOV RITOV* AND AVISHAI MANDELBALM†

The Hebrew University* and Technion-Israel Institute of Technology†

We show how to construct the best linear unbiased predictor (BLUP) for the continuation of a curve in a spline-function model. We assume that the entire curve is drawn from some smooth random process and that the curve is given up to some cut point. We demonstrate how to compute the BLUP efficiently. Confidence bands for the BLUP are discussed. Finally, we apply the proposed BLUP to real-world call center data. Specifically, we forecast the continuation of both the call arrival counts and the workload process at the call center of a commercial bank.

1. Introduction. Many data sets consist of a finite number of multidimensional observations, where each of these observations is sampled from some underlying smooth curve. In such cases it can be advantageous to address the observations as functional data rather than as multiple series of data points. This approach was found useful, for example, in noise reduction, missing data handling, and in producing robust estimations (see the books Ramsay and Silverman, 2002, 2005, for a comprehensive treatment of functional data analysis). In this work we consider the problem of forecasting the continuation of a curve using functional data techniques.

The problem we consider here is relevant to longitudinal data sets, in which each observation consists of a series of measurements over time that describe an underlying curve. Examples of such curves are growth curves of different individuals and arrival rates of calls to a call center or of patients to an emergency room during different days. We assume that such curves, or measurement series that approximate these curves, were collected previously. We would like to estimate the continuation of a new curve given its beginning, using the behavior of the previously collected curves.

Although each observation consists of a finite number of points, the observation can be thought of as a smooth function. This dual representation leads to two different approaches when attempting to solve the prediction problem. In the discrete approach, each observation is a longitudinal vector of length $p + q$. We are interested in the prediction of the last $q$-length part

Keywords and phrases: functional data analysis, best linear unbiased predictor, call center data, B-splines
Fig 1. **Arrival count in five-minutes resolution** for six successive weeks, grouped according to weekday (Friday was omitted due to space constraints). There is a clear difference between workdays, Saturdays, and Sundays. For the working days, it seems that there is some common pattern. Between 7 AM and 10 AM the call count rises sharply to its peak. Then it decreases gradually until 4 PM. From 4 PM to 5 PM there is a rapid decrease followed by a more gradual decrease from 5 PM until 12 AM. The call counts are smaller for Saturday and much smaller for Sunday. Note also that the main activity hours for weekends are 8 AM to 5 PM, as expected.

compare our results to the mean of the preceding days, from 12 PM on.

For a detailed description of the first example’s data, the reader is referred to Weinberg, Brown and Stroud (2007), Section 2. For an explanation of how the second example’s workload process was computed, the reader is referred to Reich (2010). The data for the third example was extracted using SEEStat, which is a software written at the Technion SEELab\textsuperscript{1}. We refer the reader to Donin et al. (2006) for a detailed description of the U.S. commercial bank call-center data from which the data for all three examples was extracted. The U.S. bank call-center data is publicly downloaded from SEELab server\textsuperscript{1}.

5.3. **Forecast implementation.** The forecast was performed by Matlab implementation of the BLUP algorithm from Section 3, where we enable regularization as in (9). For the implementation we used the functional data

\textsuperscript{1}SEELab: The Technion Laboratory for Service Enterprise Engineering. Webpage: \url{http://ie.technion.ac.il/Labs/Serveng}
Note that the direct workload forecast results are slightly better than the indirect workload forecast in most of the categories. Also note that in almost all categories, there is an improvement in the 10 AM and 12 PM forecasts over the forecast based solely on past days. The RMSE mean decreases by about 11% (9%) for the 10 AM forecast, and by 15% (12%) for the 12 PM forecast for the direct (indirect) forecast. Figure 3 presents a visual comparison between the direct and the indirect forecast methods on a specific day. The two forecasts look roughly the same, which is also true for all other days in this data set.

While in this example there is no significant difference between the direct and indirect workload forecasts, we expect these methods to obtain different forecasts when the arrival rate changes during an average service time. This

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2}
\caption{Forecasting results for the week following Labor Day (Sept. 2-5, 2003) for the call arrival process of the first example. Labor Day itself (Monday) does not appear since holiday data is not included in the data set. The black dots represent the true call counts in five-minutes resolution. The forecasts based on previous days, 10 AM data, and 12 PM data are represented by the blue, red, and green lines, respectively.}
\end{figure}
Fig 3. **Workload forecasting** for Friday, September 5, 2003, using both the direct and the indirect methods. The black curve represents the workload process estimated after observing the data gathered throughout the day. The blue and red curves represent the workload forecasts for the indirect and direct forecasts, respectively, given data up to 12 PM.

variance does not change drastically (see Figure 1).

5.7. **Confidence bands.** Following Weinberg, Brown and Stroud (2007), we introduce the 95% confidence band coverage (COVER) and the average 95% confidence band width (WIDTH). Specifically, for each day \( j \), let

\[
COVER_j = \frac{1}{K} \sum_{k=1}^{K} I(F_{L,jk} < N_{jk} < F_{U,jk}) ; \quad WIDTH_j = \frac{1}{K} \sum_{k=1}^{K} (F_{U,jk} - F_{L,jk}) ,
\]

<table>
<thead>
<tr>
<th>Example 3</th>
<th>RMSE</th>
<th>APE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day ahead</td>
<td>10 AM</td>
</tr>
<tr>
<td>Minimum</td>
<td>3.66</td>
<td>3.62</td>
</tr>
<tr>
<td>Q1</td>
<td>5.37</td>
<td>5.63</td>
</tr>
<tr>
<td>Median</td>
<td>6.80</td>
<td>7.01</td>
</tr>
<tr>
<td>Mean</td>
<td>7.64</td>
<td>7.19</td>
</tr>
<tr>
<td>Q3</td>
<td>9.01</td>
<td>8.97</td>
</tr>
<tr>
<td>Maximum</td>
<td>16.12</td>
<td>11.84</td>
</tr>
</tbody>
</table>

Table 3. Summary of statistics (minimum, lower quartile (Q1), median, mean, upper quartile (Q3), maximum) of RMSE and APE for the forecast based on the mean of the previous days and the BLUP, using 10 AM and 12 PM cuts for the weekends data set.
where \((F_{L,jk}, F_{U,jk})\) is the confidence band of day \(j\), evaluated at the beginning of the \(k\)-th interval (see (16)). The mean coverage and mean width, for all three examples, are presented in Table 4. First, note that for all three examples, the width of the confidence band narrows down as more information is revealed. In other words, the width of the confidence band for the 12 PM forecast is narrower than the width for the 10 AM forecast which, in turn, is narrower than the width for the previous days’ mean. We also see that the mean coverage becomes more accurate as more information is revealed. Figure 4 depicts the confidence bands for the arrival process on a particular Sunday. Note that the confidence bands for the previous days’ forecast and the 10 AM forecast almost coincide. However, at 12 PM, when enough information on this particular day becomes available, the confidence band narrows down and does capture the underlying behavior.

**Table 4.** The mean confidence band coverage and the mean width for the forecasts based on the previous days’ mean, the 10 AM cut and the 12 PM cut for the arrival process on the working days data set (Example 1), the workload process on the working days data set (Example 2) and the arrival process on the weekends data set (Example 3).

<table>
<thead>
<tr>
<th></th>
<th>Coverage</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Example 1</td>
<td>Example 2</td>
</tr>
<tr>
<td>Mean</td>
<td>93.19%</td>
<td>91.69%</td>
</tr>
<tr>
<td>10 AM</td>
<td>94.14%</td>
<td>92.27%</td>
</tr>
<tr>
<td>12 PM</td>
<td>94.86%</td>
<td>93.08%</td>
</tr>
</tbody>
</table>

**Figure 4.** Confidence bands for Sunday, August 10, 2003. The black dots represent the true call counts in fifteen-minutes resolution. The confidence bands based on previous days, 10 AM data, and 12 PM data are represented by the blue, red, and green lines, respectively.
Recall: **4 Constructions of the Poisson Process**

**Interarrival times:** Exponential iid; for Simulations.

**Probability-of-arrival during small intervals:** Counting & Levy (stationary independent increments), with the properties:

\[
P\{ A(t + dt) - A(t) = 1 \} = \lambda dt + o(dt),
\]

\[
\{ = 0 \} = 1 - \lambda dt + o(dt),
\]

\[
\{ \geq 2 \} = o(dt).
\]

**Axiomatic:** Counting & Levy suffices!
(But up to \( \lambda \)).

**Intuitive:** from Bernoulli to Poisson
(The Law of Rare Events).
Intuitive Construction (Animation): from Bernoulli to Poisson

Model for “completely random” arrivals, over the time interval $[0, T]$, at rate $\lambda$:

- **Large** number of customers $n$, each one calling during $[0, T]$, with a **small** probability $p_n \approx \frac{\lambda T}{n}$ (rate $\lambda$).
- Times of calls **uniformly** distributed over $[0, T]$.
- Then: number of calls $A(T) \overset{d}{=} \text{Bin}(n, p_n)$.
- Note: $np_n \to \lambda T$, as $n \to \infty$.
- By **Law of Rare Events**: $A(T) \Rightarrow \text{Poiss}(\lambda T)$.

**Simulation Examples (Mathlab)**

$n = 10000$, $p_n = 0.01$  \hspace{1cm} $n = 100000$, $p_n = 0.001$
§3.1 Definition 3.2 requires too much. As discussed, Levy + counting ⇒
\[ \exists \lambda > 0 \ni N(t) - N(s) \sim \text{Poisson}[\lambda(t-s)]. \]
In particular,
\[
P\{ N(t+dt) - N(t) = 1 \} = \lambda dt + o(t)
\]
\[
P\{ N(t+dt) - N(t) = 0 \} = 1 - \lambda dt + o(t).
\]
\[
P\{ N(t+dt) - N(t) > 1 \} = o(t)
\]

§3.2 Derivation of the Poisson distribution from Bernoulli.

§3.3 Properties of the Poisson Process.

1. Poisson marginals: number of events in any interval is Poisson;
\[
EN_t = \lambda t, \text{Var } N_t = \lambda t
\]
\[
\Rightarrow C = \frac{\sigma}{E} = \frac{\sqrt{\lambda t}}{\lambda t} = \frac{1}{\sqrt{\lambda t}} \text{ small for } t \text{ large.}
\]

2. Interarrival times which are iid exp (\(\lambda\)).

Beginning of proof: \( P(T_1 \geq t) = P(N_t = 0) = e^{-\lambda t}, \ t \geq 0. \)
This is a characterizing property that is practical for simulation.
Extensions to \(T_2, T_3, \ldots,\) and their independence, if rigorous, requires more than the “it should be apparent” in Hall, pg. 58.

3. Memoryless property: time till next event does not depend on the elapsed time since the last event.

4. \(S_n = T_1 + \cdots + T_n \sim \text{Gamma}(n, \lambda) = \text{Erlang}.\)

5. Order-statistics property: Given \(N(t) = n,\) the unordered event times are distributed as \(n\) iid r.v., uniformly distributed on \([0, t].\)
\[
\Rightarrow \text{simulation over } [0, t]: N(t) \sim \text{Poisson}(\lambda t); \quad U_1, U_2, \ldots, U_{N(t)} \text{ iid } U[0, t].
\]

§3.4 Goodness of Fit

How well does a Poisson model fit our arrival process?

Qualitative assessments:
Airplanes landing times at a single runway, during an hour: no
Airplanes landing times at a large airport, during an hour: plausible
Job candidates that arrive at their appointments during an hour: no
Visits to a zoo, most of which arrive in groups, during an hour: no
Arrival times at a bank ATM = Automatic Teller Machine, during an hour: plausible
§3.5 **Quantitative Tests**

Graphical Tests:

- cumulative arrivals vs. a straight line (Fig. 3.2)
- paired successive interarrivals (Fig. 3.4)
- exponential interarrivals
  (How do you identify exp (·) when you see one? Use Histograms!)

§3.6 **Parameter Estimation**

Estimate $\lambda =$ arrival rate.

MLE (Max. Likelihood Estimator), given $A(t)$, $t \leq T$: $\hat{\lambda} = \frac{A(T)}{T}$.

Confidence intervals for $\frac{1}{\lambda}: \frac{T}{A(t)} \pm z_{\alpha} \frac{T}{A(T)^{3/2}}$ \hspace{1cm} (3.34)

**Sample-size**: for $(1 - \alpha)$-confidence interval of width $w$, $N \geq \left[ \frac{2\pi}{w\lambda} \right]^2$.

Thus, for $w = \epsilon \cdot \frac{1}{\lambda}$, we need $N \geq \left[ \frac{2\pi}{\epsilon} \right]^2$.

(Eg.: 95%-confidence interval of width = 10% of mean, requires $N \geq \left[ \frac{2 \times 1.96}{0.1} \right]^2 \approx 1500$!)
**PASTA = Poisson Arrivals See Time Averages**

A rigorous proof in the following general setting was first presented by R. Wolff.

**Arrivals (Observations):**

\[ A = \{ A(t), \ t \geq 0 \}, \]

\[ A(t) = \text{number of arrivals in } [0, t]. \]

**System:**

\[ X = \{ X(t), \ t \geq 0 \}, \]

\[ X(t) = \text{state at time } t. \]

\[ \bar{t} \triangleq \lim_{T \to \infty} \frac{1}{T} \int_0^T X(t) dt \]

\[ \bar{\xi} \triangleq \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^N X(S_n -) \]

where \( S_n = n-\text{th arrival time.} \)

**PASTA assumptions:**

(i) \( A \) is Poisson, and

(ii) **Lack of Anticipation.** For every \( t \geq 0 \), \( \{ A(t + u) - A(t) : u \geq 0 \} \) is independent of \( \{ X(s) : 0 \leq s \leq t \} \).

Then \( \bar{t} = \bar{\xi} \), in the following precise sense: If one limit exists, then the other exists as well, in which case they are equal.

**Second version:** Let \( B \) be an arbitrary subset of a process \( X \). Define by

\[ \bar{t}_B \triangleq \lim_{T \to \infty} \frac{1}{T} \int_0^T I\{ X(t) \in B \} dt \]

the fraction of time that the system spends in \( B \), and let

\[ \bar{\xi}_B \triangleq \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^N I\{ X(S_n -) \in B \} \]

be the fraction of arrivals that find ("see") the system in \( B \).

Then \( \bar{t}_B = \bar{\xi}_B \) in the same sense as above.
Application of PASTA: Biased Sampling

A renewal process is a counting process with iid interarrivals.

Descriptions: \( R = \{ R(t), \ t \geq 0 \} \) or \( \{ T_1, T_2, \ldots \} \) iid, or \( \{ S_1, S_2, \ldots \} \)

Example: Poisson exponential Erlang

**Story:** Buses arrive to a bus stop according to a renewal process \( R_b = \{ R_b(t), \ t \geq 0 \} \).

- \( T_i^b \) — times between arrivals of the buses.
- Passengers arrive to the bus stop in a completely random fashion (Poisson).
- \( S_i^p \) — arrival times of the passengers.

**Question:** How long, on average, do they wait? Plan service-level.

\[ A = \{ A(t), \ t \geq 0 \} = \text{Poisson arrivals of passengers.} \]

\[ X = \{ X(t), \ t \geq 0 \} = \text{state = Virtual waiting time.} \]

PASTA: 
\[
\lim_{T \to \infty} \frac{1}{T} \int_0^T X(t) \, dt = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} X(S^p_{n\to}) = \bar{\tau}
\]

\[ \Rightarrow \bar{\tau} = \frac{1}{T} \cdot \text{(area under } X, \text{ over } [0, T]) \]

\[ \approx \frac{1}{T} \cdot \left( \frac{1}{2} (T_1^b)^2 + \frac{1}{2} (T_2^b)^2 + \cdots + \frac{1}{2} (T_{R_b(T)}^b)^2 \right) \]

\[ = \frac{R_b(T)}{T} \cdot \frac{1}{2} \cdot \frac{T_1^b + \cdots + T_{R_b(T)}^b}{R_b(T)} \xrightarrow{\tau \to \infty} \frac{1}{E(T_1^b)} \cdot \frac{1}{2} \cdot E(T_1^b)^2, \text{ by SLLN} \]

\[ = \frac{1}{2} E(T_1^b) \left[ 1 + \frac{\sigma^2}{E} \right], \quad c = \frac{\sigma}{E} \text{ coefficient of variation.} \]

"Deterministic" answer Bias, due to variability

Check Poisson bus arrivals to derive Paradox:

\[ \frac{1}{2} \text{("stochastic" answer)} = \frac{1}{2} \text{("deterministic" answer).} \]
Counting process with independent increments:

\[
P\{ A(t + dt) - A(t) = 1 \} = \lambda(t)dt + o(dt),
\]
\[
\{ = 0 \} = 1 - \lambda(t)dt + o(dt),
\]
\[
\{ > 1 \} = o(dt).
\]

**Main Property:**

Poisson number of arrivals over intervals:

\[
A(T_2) - A(T_1) \overset{d}{=} \text{Poiss} \left( \int_{T_1}^{T_2} \lambda(s)ds \right).
\]

**Construction from time-homogeneous:**

(Time-Change in Stochastic Processes; Thinning here)

Data: Arrival rate \( \lambda(t) \), \( 0 \leq t \leq T \).

Let \( \lambda_{\max} = \max_{t \in [0,T]} \lambda(t) \).

1. Simulate a homogeneous Poisson(\( \lambda_{\max} \)) process.

2. **Thinning.** For each arrival \( S_i \), generate \( U_i \overset{iid}{=} U(0,1) \).

Let \( p_i = \lambda(S_i)/\lambda_{\max} \).

\( U_i \leq p_i \), accept arrival;
\( U_i > p_i \), reject arrival.
Arrivals to a Call Center: How to Model?

Arrivals over the day are not time-homogeneous.

Arrivals over small intervals (15, 30, 60 min) are close to time-homogeneous Poisson.

Arrivals over the day are non-homogeneous Poisson.

Practically: Test (Brown), then model, as a Poisson process with piecewise-constant arrival rates.

How to predict/forecast arrival rates?
Arrivals to a Call Center: 
Variability of the Arrival Rates

Number of Calls at a U.S. bank. 

25 Mondays overall.
- **13:00-13:30**: 25 observations, range: 2,500-3,2000; Sample Mean=$2,842$, BUT Sample Variance=$24,539$!
- **17:00-17:30**: Mean=1,705, Variance=10,356.

Conclude: **Number of calls during “similar” intervals not i.i.d Poisson: over-dispersion.**
A Test for Inhomogeneous Poisson Process

1. Break up the interval of a day into short blocks of time, say $I$ (equal-length) blocks of length $L$.

2. Let $T_{i0} = 0$ and $T_{ij}$: the $j$-th ordered arrival time in the $i$-th block, $i = 1, \ldots, I$ and $j = 1, \ldots, J(i)$, then define

$$R_{ij} = (J(i) + 1 - j) \left( -\log \left( \frac{L - T_{ij}}{L - T_{i,j-1}} \right) \right).$$

3. Under the null hypothesis that the arrival rate is constant within each given time interval, the $\{R_{ij}\}$ will be independent standard exponential variables.

4. Use any customary test for the exponential distribution; for example, Kolmogorov-Smirnov test.
Figure 3: Exponential ($\lambda=1$) Quantile plot for $\{R_{ij}\}$ from Regular calls (11:12am – 11:18am)

$L = 6 \text{ min}, n = 420$, Kolmogorov-Smirnov statistic $K = 0.0316$ and the P-value is 0.2.
Forecasting Problem: Setup

Days are divided into time intervals, with an assumed constant arrival rate over an interval.

Practice: 15 min, 30 min, 1 hour.

\( N_{jk} \) = # of arrivals, during time-interval \( k \), on day \( j \).
Assume \( J \) days overall, with \( K \) intervals per day.

- **One-day-ahead** prediction:
  \( N_1, \ldots, N_{j-1}, \) known. Predict \( N_j, \ldots, N_{jK} \).

- **Several days (weeks) ahead** prediction.

- **Within-day** prediction.
  \( N_1, \ldots, N_{j-1}, N_j, \ldots, N_{j,k-1} \) known.
  Predict \( N_{jk}, \ldots, N_{jK} \).

Practice: Do all the above, via nested rolling horizon (Weekly, Daily, Hourly).
Forecasting: Simple Methods

Most recent observation.
\[ F_{jk} = \text{most recent "similar" call volume.} \]

Example: \[ F_{jk} = N_{j-7,k} \] (previous week).

Moving average.
Average of several (not too many) recent "similar" call volumes.

Most Recent, plus Yesterday’s Correction.
Example: Factor accounting for a “busy yesterday”.

What about sophisticated forecasting methods?
Active research.

Here, we shall compare the performance of simple methods against (given results of) sophisticated methods.
Forecasting: Goodness-of-Fit

$N_{jk}$ – number of calls (day $j$, interval $k$);

$F_{jk}$ – forecast.

Two ways to quantify forecasting accuracy:

1. **Root Mean-Square Error (RMSE)**

For each day $j$, calculate:

$$RMSE_j = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (N_{jk} - F_{jk})^2}.$$ 

$$RMSE = \sqrt{\frac{1}{J} \sum_{j=1}^{J} RMSE_j}.$$ 

2. **Average Percent Error (APE)**

$$APE_j = \frac{100}{K} \cdot \frac{1}{K} \sum_{k=1}^{K} \frac{|N_{jk} - F_{jk}|}{N_{jk}}.$$ 

$$APE = \sqrt{\frac{1}{J} \sum_{j=1}^{J} APE_j}.$$
Axiomatically, “completely random arrivals” are Poisson.

Arrivals over the day are not time-homogeneous.

Hence, arrivals over the day are non-homogeneous Poisson.

Arrivals over small intervals (15, 30, 60 min) are close to time-homogeneous Poisson.

Practically:
Test (L. Brown), then model, as a Poisson process with piecewise-constant arrival rates.
A (Common) Model for Call Arrivals

Whitt (99’), Brown et. al. (05’), Gans et. al. (09’), and others:

Doubly-stochastic (Cox, Mixed) Poisson with instantaneous rate

\[ \Lambda(t) = \lambda(t) \cdot X , \]

where \( \int_0^T \lambda(t) dt = 1. \)

- \( \lambda(t) = “\text{Shape}” \text{ of weekday} \)  [Predictable variability]
- \( X = \text{Total \# arrivals} \)  [Unpredictable variability]

w/ Maman & Zeltyn (09’):
Above assumes “too-much” stochastic variability!
Over-Dispersion (Relative to Poisson), Maman et al. (’09)

**Israeli-Bank Call-Center**

**Arrival Counts - Coefficient of Variation (CV),** per 30 min.

Sampled CV - solid line, Poisson CV - dashed line

- Poisson CV = $1/\sqrt{\text{mean arrival-rate}}$.
- Sampled CV’s $\gg$ Poisson CV’s $\Rightarrow$ **Over-Dispersion**.
Over-Dispersion: Fitting a Regression Model

\[ \ln(\text{STD}) \text{ vs. } \ln(\text{AVG}) \]

**Tue-Wed, 30 min resolution**

\[ y = 0.8027x - 0.1235 \]
\[ R^2 = 0.9899 \]

\[ y = 0.8752x - 0.8589 \]
\[ R^2 = 0.9882 \]

**Tue-Wed, 5 min resolution**

\[ y = 0.7228x - 0.0025 \]
\[ R^2 = 0.9937 \]

\[ y = 0.7933x - 0.5727 \]
\[ R^2 = 0.9783 \]

Significant linear relations (Aldor & Feigin):

\[ \ln(\text{STD}) = c \cdot \ln(\text{AVG}) + a \]
Over-Dispersion: Random Arrival-Rate Model

The **linear relation** between $\ln(\text{STD})$ and $\ln(\text{AVG})$ motivates the following model:

Arrivals distributed **Poisson with a Random Rate**

$$\Lambda = \lambda + \lambda^c \cdot X, \quad 0 \leq c \leq 1;$$

- $X$ is a random-variable with $E[X] = 0$, capturing the magnitude of **stochastic deviation** from mean arrival-rate.
- $c$ determines **scale-order** of the over-dispersion:
  - $c = 1$, proportional to $\lambda$;
  - $c = 0$, Poisson-level, same as $0 \leq c \leq 1/2$.

In **call centers**, over-dispersion (per 30 min.) is of order $\lambda^c$, $c \approx 0.8 \sim 0.85$. 
Fitting a **Gamma Poisson** mixture model to the data: Assume a (conjugate) prior Gamma distribution for the arrival rate $\Lambda \sim Gamma(a, b)$. Then, $Y \sim Poiss(\Lambda)$ is Negative Binomial.

Very good fit of the Gamma Poisson mixture model, to data of the Israeli Call Center, for the majority of time intervals.

Relation between our $c$-based model and Gamma-Poisson mixture is established.

Distribution of $X$ derived, under the Gamma prior assumption: $X$ is asymptotically normal, as $\lambda \to \infty$. 

*Over-Dispersion: Distribution of $X$?*
Over-Dispersion: The Case of ED’s

**Israeli-Hospital Emergency-Department**

**Arrival Counts - Coefficient of Variation**, per 1-hr. & 3-hr.

- **One-hour resolution**
- **Three-hour resolution**

- Moderate over-dispersion: \( c = 0.5 \) reasonable for hourly resolution.
- **ED beds in conventional QED** (Less var. than call centers ! ?).
Unpredictable Variability: The Multi-Class Case

- Research w/ I. Gurvich & P. Liberman, ongoing.

Unpredictable variability: \( X = (X_1, \ldots, X_I) \)

Pairs: \((X_{\text{Retail}}, X_{\text{Business}})\) and \((X_{\text{Business}}, X_{\text{Platinum}})\)

US Bank: Correlations, 600 weekdays

- Positive correlation (vs. independent in existing research)
- Research: Empirical, then Impact on design and control?