Technion - Israel Institute of Technology
The William Davidson Faculty of Industrial Engineering & Management

Center for Service Enterprise Engineering (SEE)

https://seelab.net.technion.ac.il

SEEStat 3.0 Tutorial

Created: June 27, 2010
Last Revised: November 2, 2016

SEEStat 3.0 Tutorial

Introduction

SEEStat is a system for Exploratory Data Analysis (EDA) in real-time. It enables users to easily conduct statistical and performance analyses of massive datasets; in particular, analyzing datasets representing operational histories of large service operations (e.g. call centers, hospitals, internet sites), such as those available through the SEELab server. SEEStat can also automatically create sophisticated reports in Microsoft Excel, which support research and teaching.

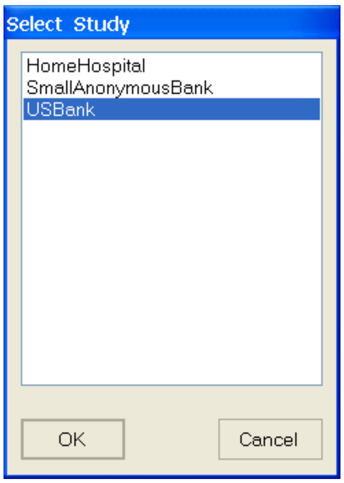
Both SEEStat and the SEELab Server were developed at the Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology. More information on the SEELab can be found at its homepage https://seelab.net.technion.ac.il/.

How to connecting to SEEStat on the Technion SEELab Server can be found here.

SEEStat Tutorial

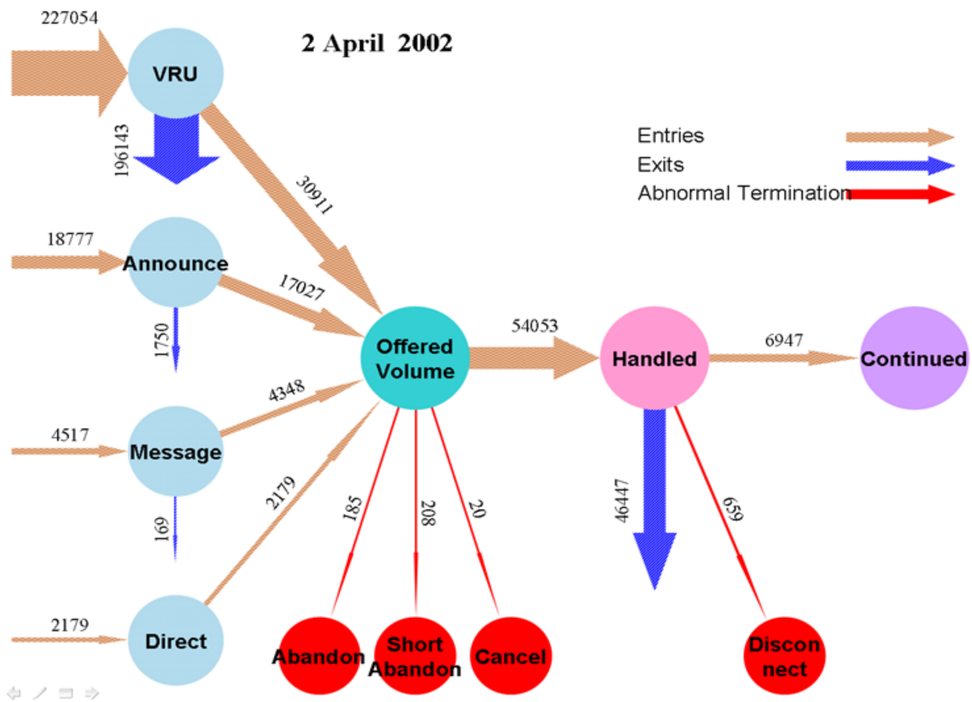
USBank Data

<u>Background</u>: The source of this example database is a large call center of a U.S. Bank. This call center has sites in 4 states, which are integrated to form a single virtual call center: Calls are queued up, when appropriate, in a central queue; they are then served by agents across sites, by fitting service types to agent skills using SBR (Skills-Based Routing) algorithms.
The virtual call center has about 900–1200 agent positions on weekdays, and 200–500 agent positions on weekends. Agents process up to 300,000 calls per day (about 20% reach the agent-queue, and the rest complete their service process within the VRU = Voice Response Unit).



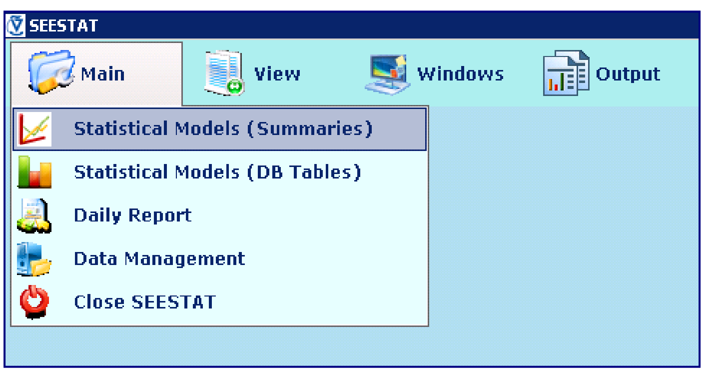
Customer flow chart of the USBank call center

The following flow-chart describes the process-flow of calls in a typical day (Tuesday, April 2, 2002). There are 4 entry points to the system: through the VRU (Voice Response Unit), Announcement, Message, and Direct group (callers that directly connect to an agent). The most commonly used is the VRU. Most of the calls end service in the VRU (196143 calls - about 80% of all calls); while around 20% of the callers entering the system seek service by an agent ('Offered Volume'). Less than 1% of the Offered Volume calls will not reach an agent service – those customers abandon the queue while waiting (a few are disconnected due to technical problems – 'Cancel'). When an appropriate agent becomes available the customer is getting served ('Handled'), after which the customer either exits the system or is moved to a secondary service ('Continued').



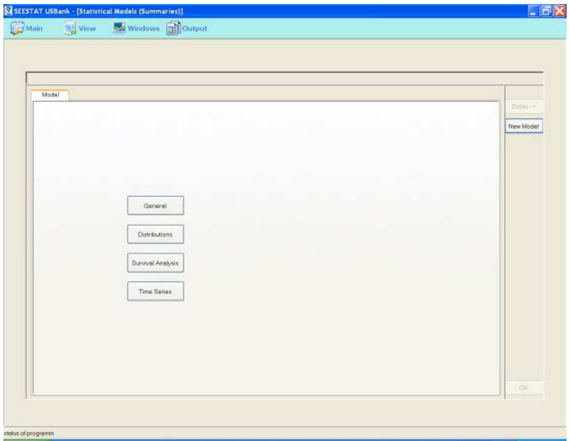
Part 1

After connecting to the server, click the SEEStat 3.0 icon to open the program.
On the top of the screen you see the main menu. Click **"Main"**. We shall work with **"Statistical Models (Summaries)".** Click it.



A list-box with SEEStat *studies* appears (three databases in our case). Select **USBank** (the database we shall be working with), click **"OK"** and wait a few seconds.

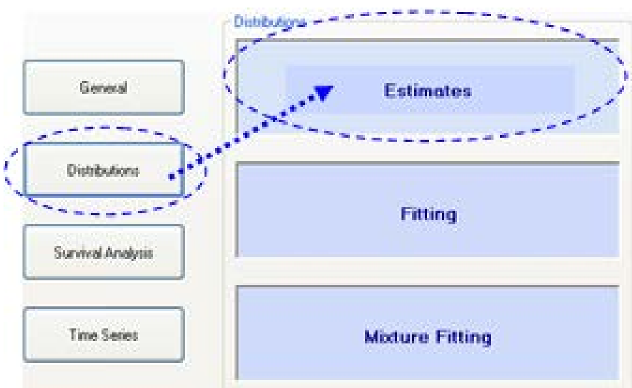
Now you see the **"Model"** panel.



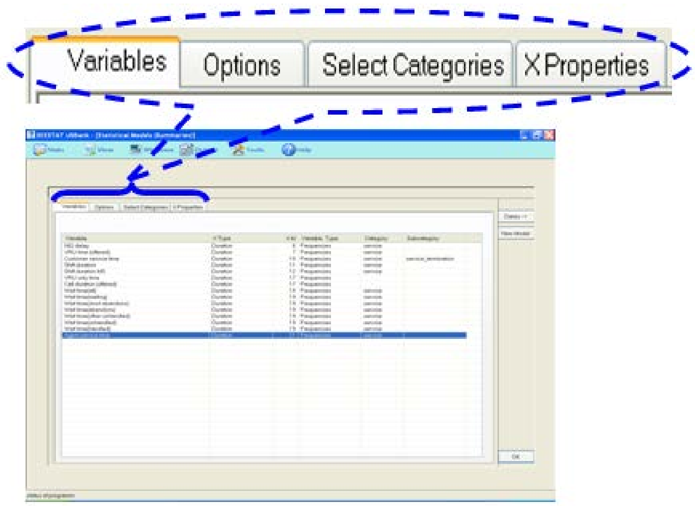
Example 1.1: ***Distributions***

We shall now create a histogram of the ***service time*** *(duration) distribution, a 1-second resolution.*

Click the **"Distributions"** button. Three available distribution models appear.
Select **"Estimates"**.



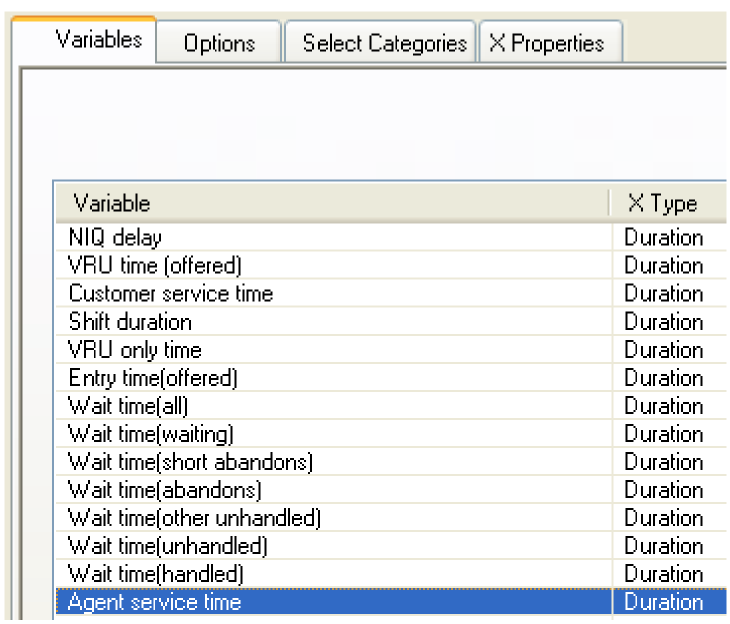
You see the tab control that has 4 tabs:
“Variables”, “Options”, “Select Categories” and “X Properties”.



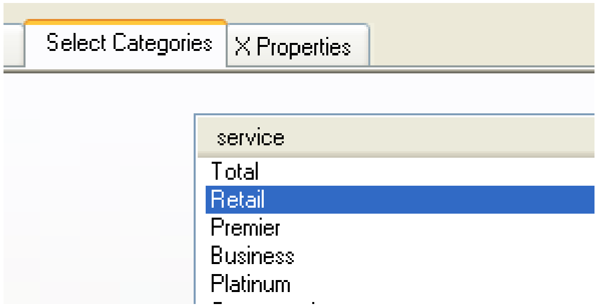
The first one **"Variables"** is active. This tab is *mandatory*, which means you must select variable(s) before moving forward. The three other tabs are *optional*, which means that they already have default values.

NOTE: You can select (click) several variables simultaneously by holding the Ctrl key and clicking on the variables one by one.

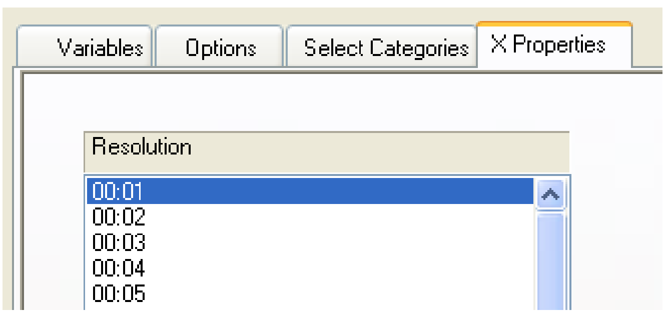
Select "**Agent service time"**(the last entry in the list).



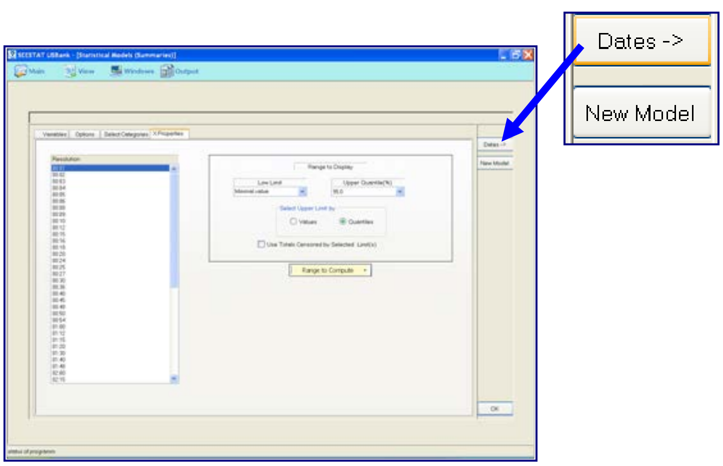
Now move to the **"Select Categories"** tab. You see a list box with all the service types that are offered by USBank. Select **"Retail"**, which is the Bank's main service.



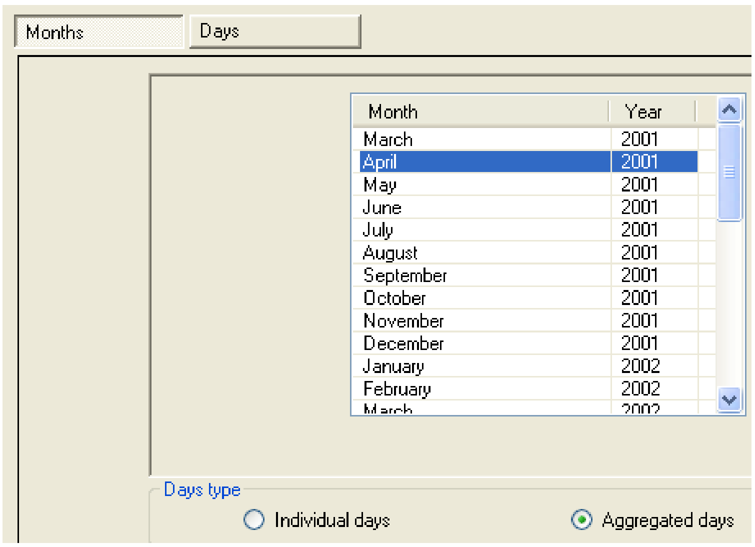
Open the **"X Properties"** tab. It is used to set properties of charts and tables. On the left side you can see the **"Resolution"** list box. The default resolution (bin-size of the histogram) of 5 seconds is marked. Select the minimal resolution **00:01** = 1 second, in order not to miss any details of the histogram.



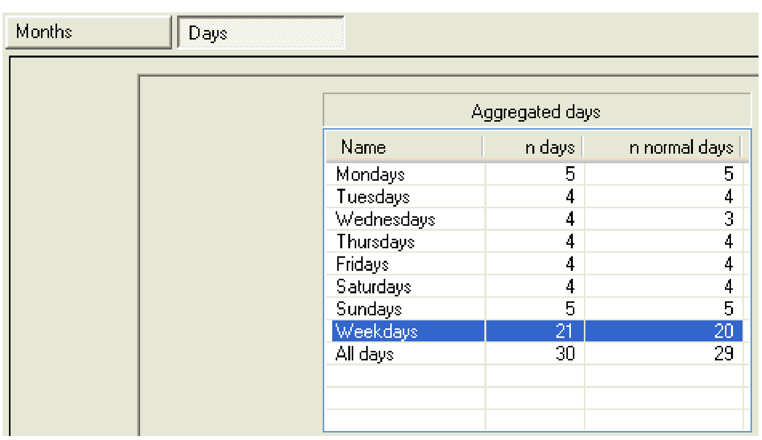
Now you must select the dates we focus on. Click the **"Dates ->"** button on the right side.



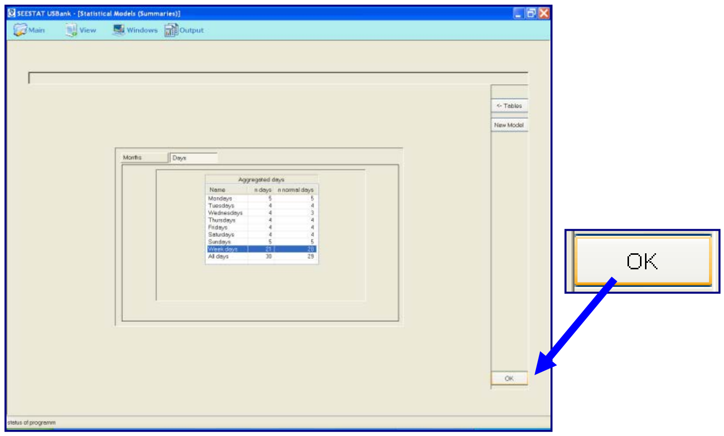
You see the list of months for which the requested data is available.
Select **"April 2001"**.
Below the list of months, you see two options for date-selection (Date type): "Aggregated days" and "Individual days". **"Aggregated days"** is the chosen-default, which we now follow.



Click **"Days"** to make the selection of days, and select **"Weekdays"** – an aggregation of all *5 working days of the week. (Holidays and some special days, such as when there is a system failure, are excluded.)*

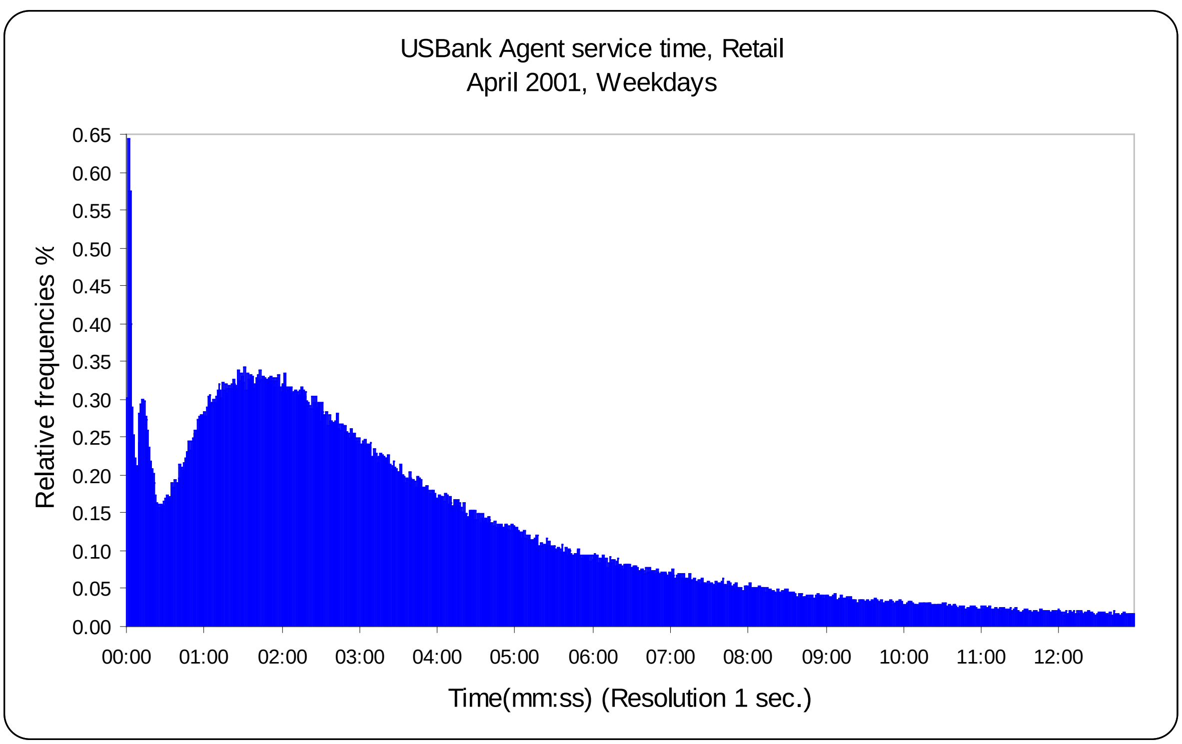


All selections have now been completed: click **"OK"** at the bottom right.



Wait a few seconds *– SEEStat is processing your request: you now see the chart/histogram, produced as "Chart 1" within an Excel spreadsheet.*
NOTE: All the examples in this tutorial, from now on, will be accumulated in this Excel file. DO NOT modify or close this Excel file.

Looking at the chart, you see some irregularities on the left (near the origin). We shall look at these more carefully later.

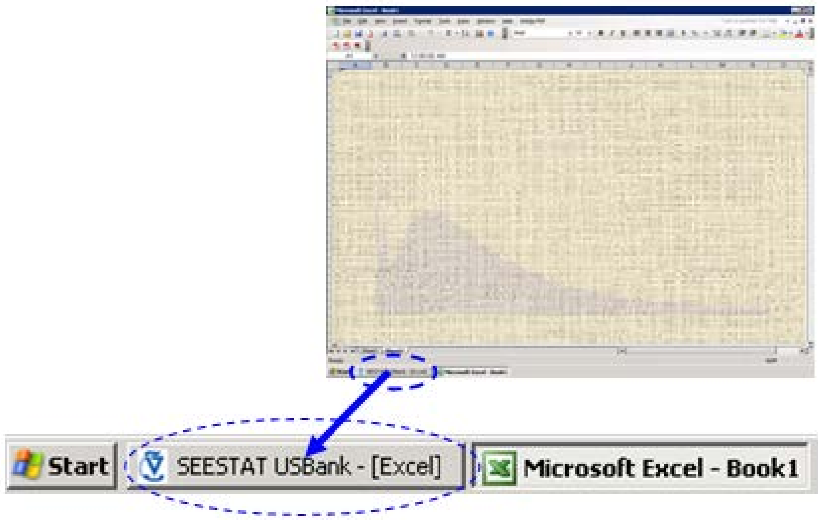


*In fact, two sheets have been created: The first is the chart in "Chart1"; the second is "Table1", which includes Table(s) that are associated with the chart, with the default one being the "**Statistics**" table. Click **"Table1"** to see the contents of this table (N=619,096 is the number of observations); skim through the summary data and then return to **"Chart1"**.*

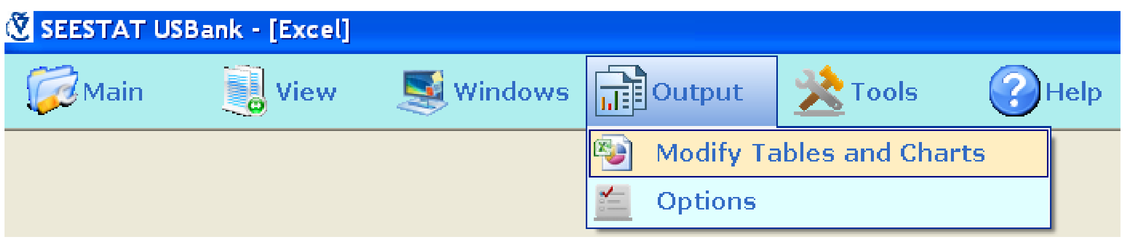
Statistics	
	Agent service time
N	619096
N(average per day)	30954.8
Mean	4 min 19 sec
Standard Deviation	4 min 35 sec
Variance	21 min^2
Median	2 min 57 sec
Minimum	0
Maximum	59 min 53 sec
Skewness	3.062
Kurtosis	15.38
Standard Error Mean	0 sec
Interquartile Range	3 min 53 sec
Mean Absolute Deviation	3 min 2 sec
Median Absolute Deviation(MAD)	1 min 42 sec
Coefficient of Variation (CV) (%)	106.17
L-moment 2 (half of Gini's Mean Difference)	2 min 6 sec
L-Skewness	0.383
L-Kurtosis	0.245
Coefficient of L-variation (L-CV) (%) (Gini's Coefficient)	48.57

You can easily make modifications to charts and tables, as long as they do not require the loading of new data from the database. You will now go through an example of such a modification.

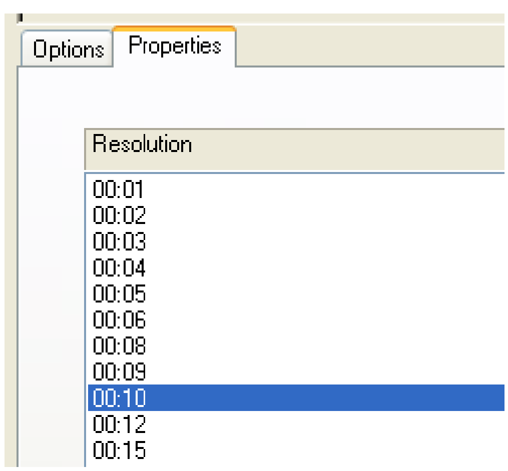
First, return to the SEEStat main menu by clicking the SEEStat USBank button, on the task bar on the lower-left side of the screen – you will repeat this action each time you wish to transfer from Excel to SEEStat.



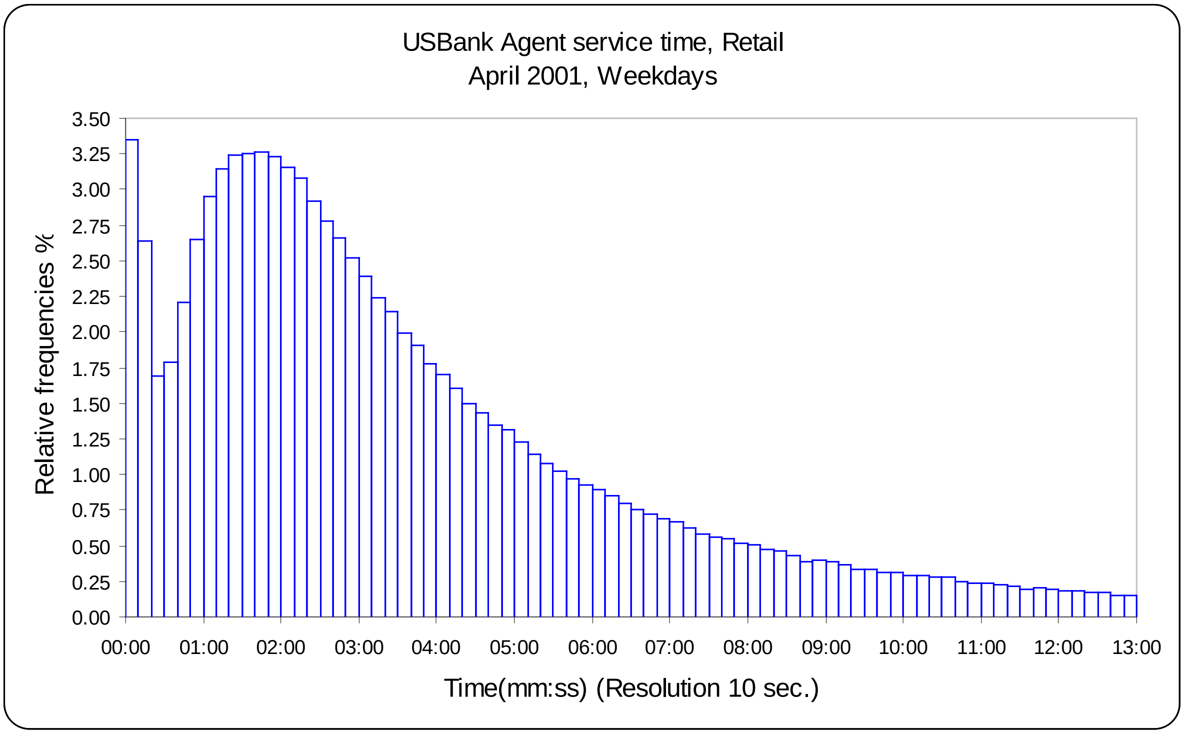
Click **"Output"** on the right side of the top main menu; after this click **"Modify Tables and Charts"**



Two tabs are available: "Options" and "Properties". Open **"Properties"** and change the resolution to **00:10** = 10 seconds.



Click **"OK"**.

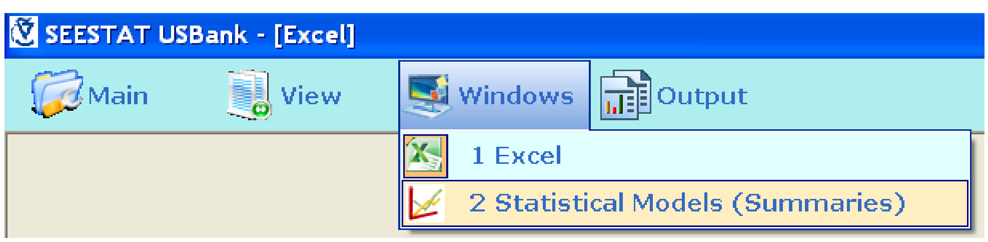


The chart is becoming smoother, but at the cost of losing some details on the left, near the origin.

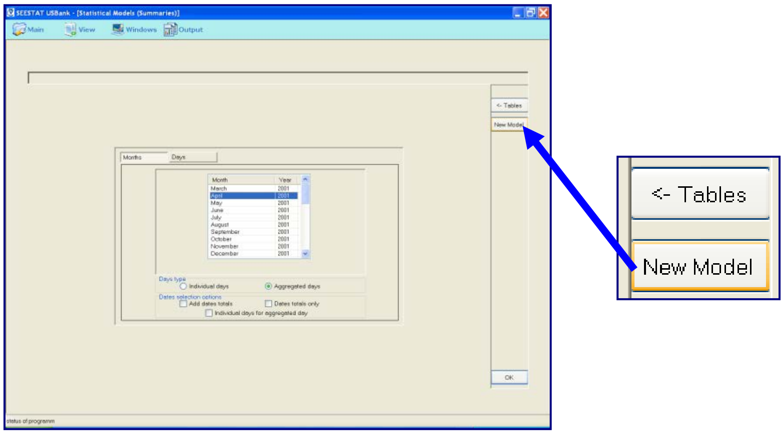
<u>Example 1.2:</u> Intraday time series

We now create a chart of arrival-counts to the call center(s) of USBank, during several days in a September.

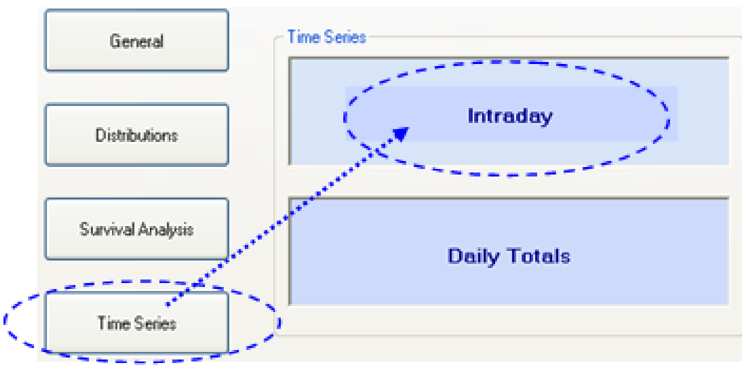
First you must return to the **"Statistical Models (Summaries)"** window. Click the **SEEStat** button on the task bar (left-bottom), next click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**



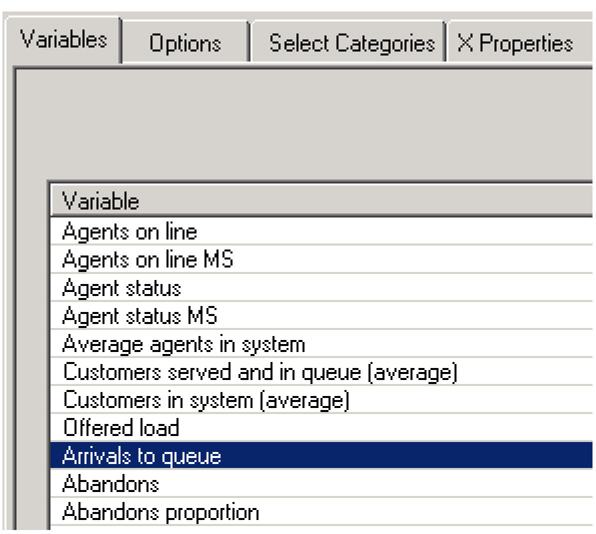
We are now changing models. To this end, select the **"New Model"** button (right side).



Select now **"Time Series"** and then select **"Intraday".**

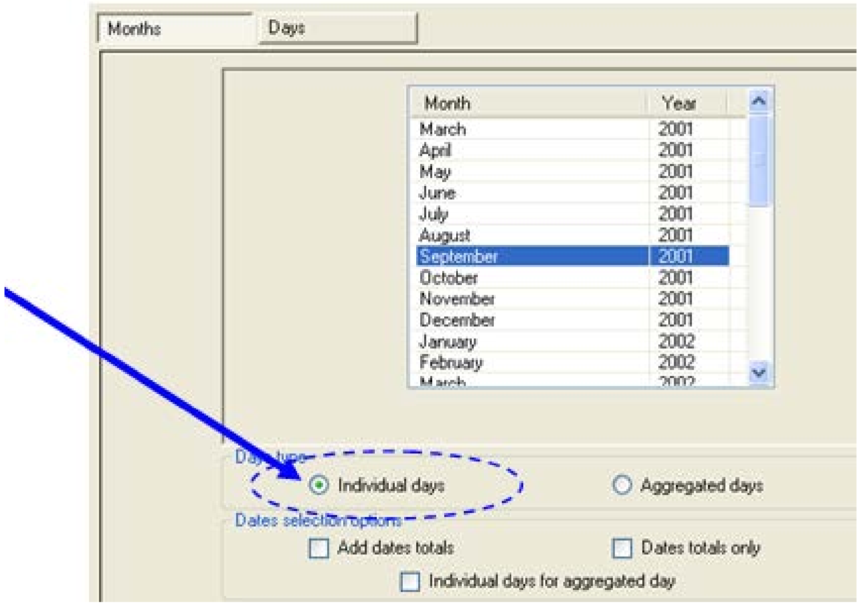


As in [Example 1.1](), four tabs appear. In the **"Variable"** tab, select **"Arrivals to queue"**. In the **"Select Categories"** tab, select **"Total"**.



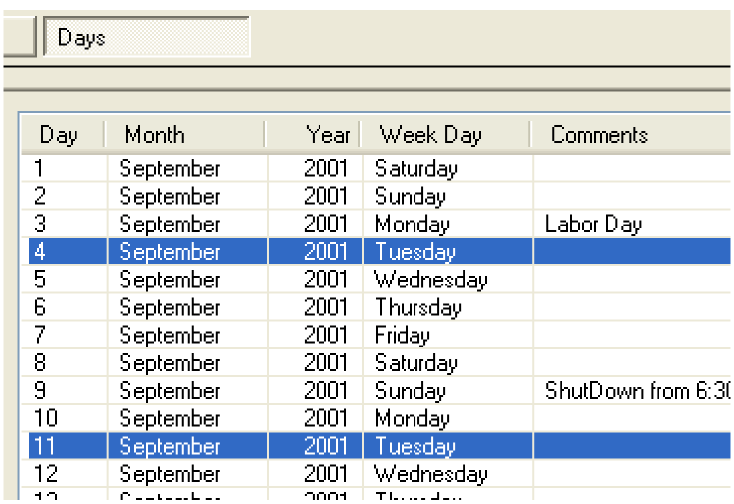
Now select dates: Click **"Dates ->"; ** Select **September 2001** from the **"Months"** list; Mark **"Individual days"**, and click the **"Days"** button.

.



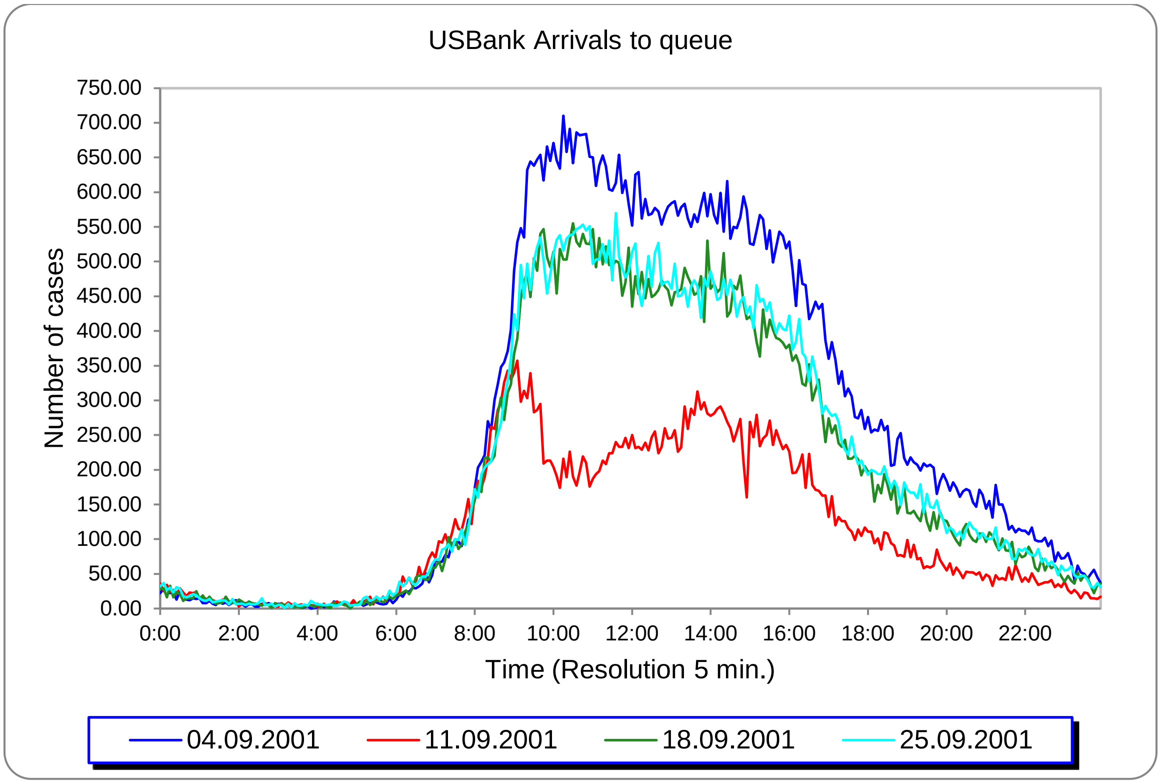
The list of days contains the date, the day of the week and comments if any. For example, Monday, September 3rd, was Labor Day.
It is expected that the Tuesday following a holiday will be a busy day. We thus compare all Tuesdays of the month: September 4, September 11, September 18 and September 25.

Hold down the **"Ctrl"** key, and in parallel click, one by one, the **four Tuesdays** of September 2001.



Day	Month	Year	Week Day	Comments
1	September	2001	Saturday	
2	September	2001	Sunday	
3	September	2001	Monday	Labor Day
4	September	2001	Tuesday	
5	September	2001	Wednesday	
6	September	2001	Thursday	
7	September	2001	Friday	
8	September	2001	Saturday	
9	September	2001	Sunday	ShutDown from 6:3(
10	September	2001	Monday	
11	September	2001	Tuesday	
12	September	2001	Wednesday	

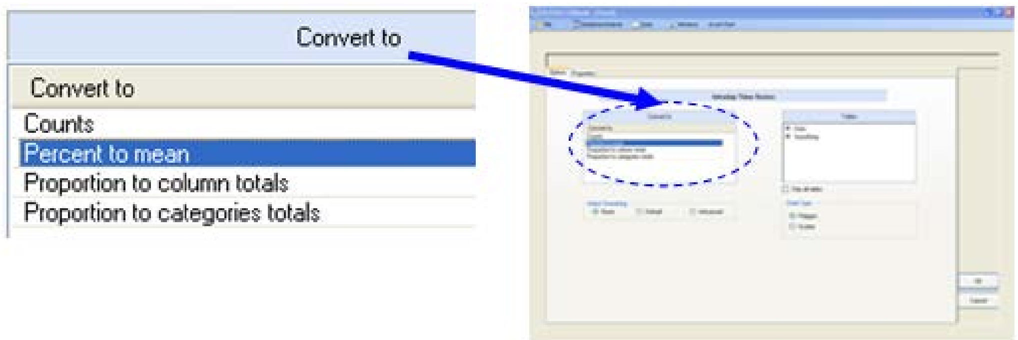
Then click **"OK"** (bottom right).



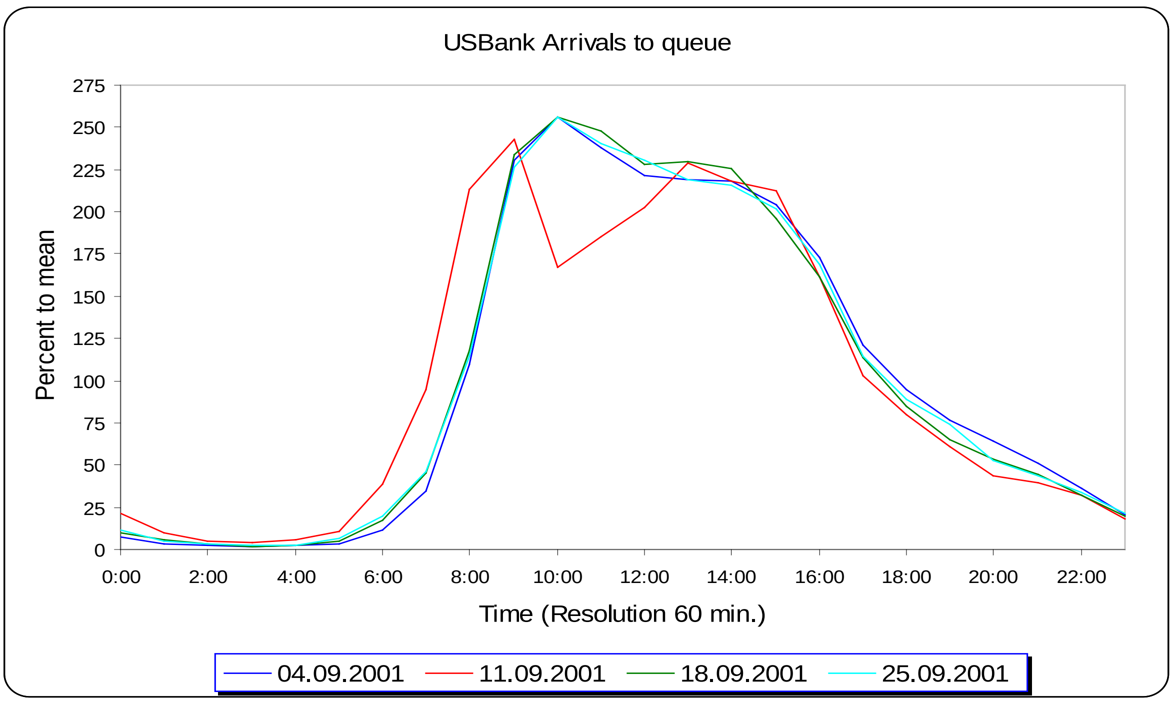
Note: The graphs appear in "Chart2" of Excel. As before,"Table2" contains the corresponding numerical data.

You see a sharp drop in the number of calls around 09:00 a.m. on ***September 11, 2001*** *– this is of course not surprising, given that one of the call centers of US Bank was in NYC, and the others located on the U.S. East Coast.*
You also see that the Tuesday after Labor Day (4.9) is indeed a heavily-loaded Tuesday, as anticipated.
The chart is noisy, due to its 5-minute resolution. We shall momentarily increase the resolution to 1 hour (60 minutes). We also note the following:
On the two Tuesdays after September 11, the number of calls is low, relative to the Tuesday after Labor Day. A natural question now arises: Is there a "shape of a Tuesday"? To seek a common pattern for (the shape of) a Tuesday, if there is any, we change the graphs from absolute counts to "percent to mean" (mean = average number of calls per resolution period).

Go back to the main menu via the **SEEStat** tab (bottom-left). In the main menu select **“Output”** then **“Modify Tables and Charts”**. In the **“Options”** tab, under the **“Convert to”** table on the left, select **“Percent to mean”**, and in the **“Properties”** tab set resolution to **60:00** = 1 hour,



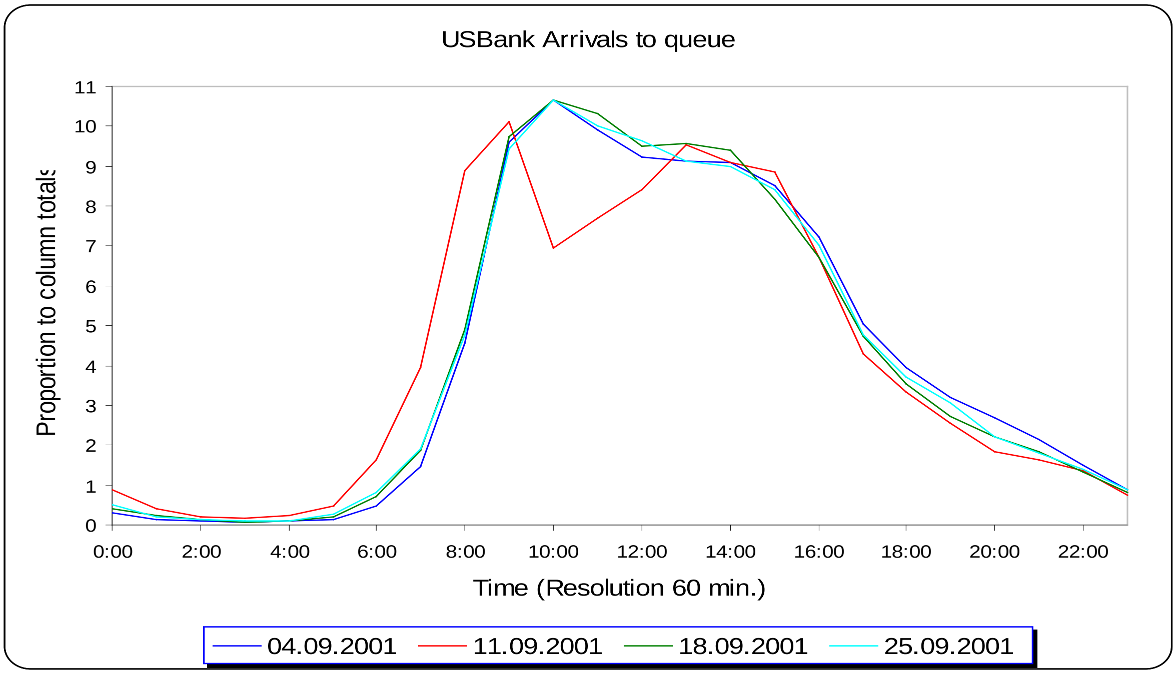
Click **"OK"**.



The "Shape of a Tuesday" is clearly manifested: the distribution of calls over the day is almost the same for the three Tuesdays, both normal and heavily-loaded. (Surprisingly, September 11 also catches up from around 13:00 or so.) For example, the arrival rate during the peak hour—from 10:00–11:00—is about 2.5 times that of an average hour.

*Instead of "Percent to mean", you can plot according to **“Proportion to column totals”** which, in simple words, means the "hourly fraction of load":*

Going via the **“SEEStat”** tab, **“Output”**, **”Modify Tables and Charts”**, **“Proportions to column totals”**, and then **“OK”.**

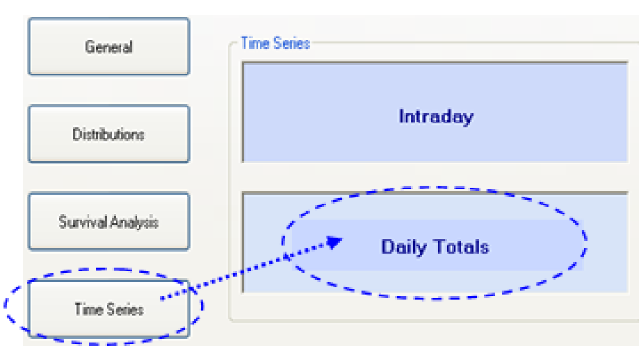


You see that the arrivals during the peak hour 10:00–11:00 constitute 10% of the daily total. (Such observations make load-predictions much easier: indeed, only the daily total must be predicted. Once the daily total is determined, the number of arrivals per hour is allocated according to the shape of the day; e.g. 10% allocated to 10:00–11:00.)

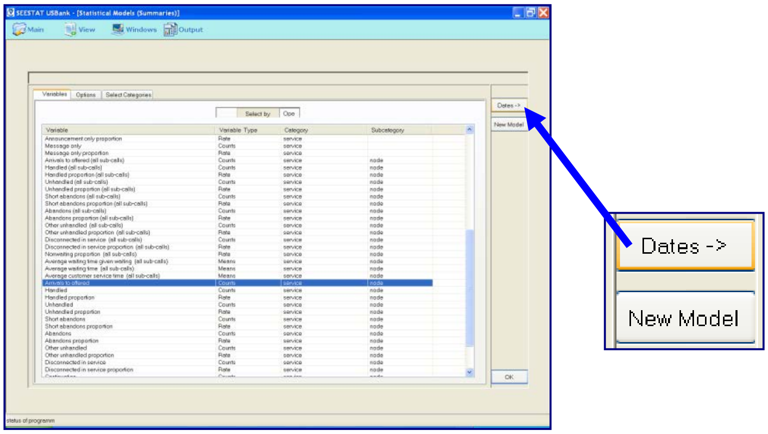
<u>Example 1.3:</u> Time series (Daily totals)

There are two types of daily-total time-series: individual days during a specific month and aggregated days by months. We now demonstrate these concepts.

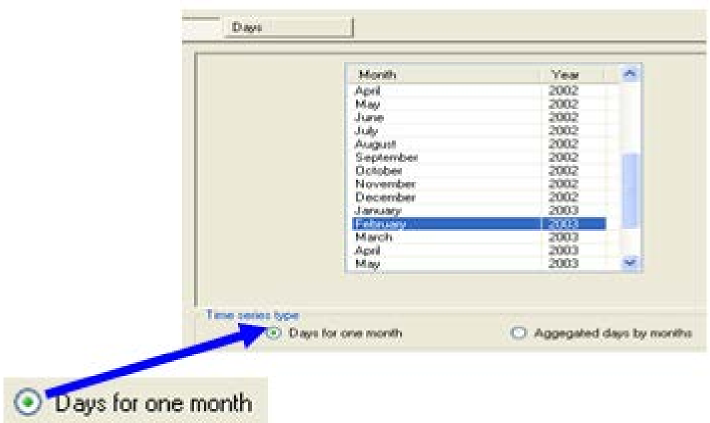
Click the **SEEStat** button on the task bar (left-bottom); next click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**. Click the **"New Model"** button. Select **"Time Series"**, then **"Daily totals"**.



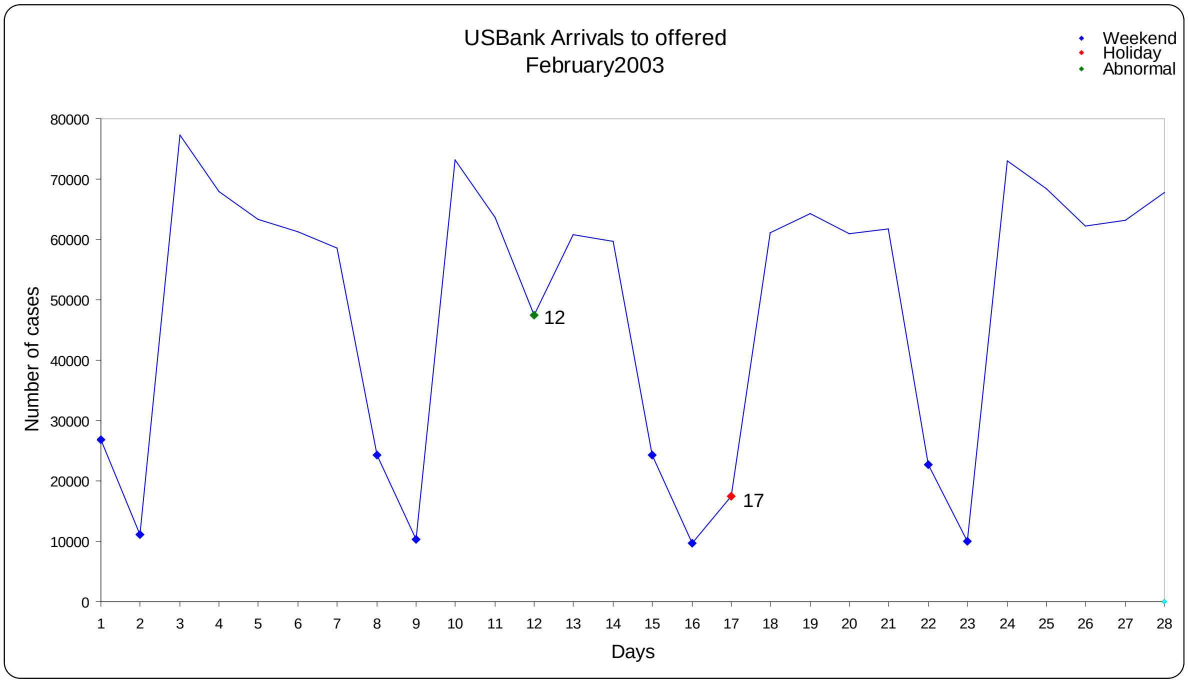
From the variables list select **"Arrivals to offered"** (around the middle of the list – it counts incoming calls that reached the queue for an agent service).
Click the **"Dates->"** button.



Mark **"Days for one month"** and select (after scrolling down) **February 2003.**

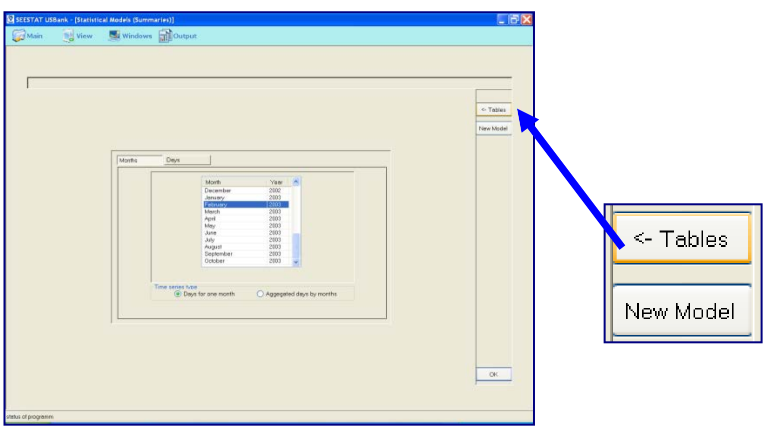


Open the **"Days"** tab (there is no need for you to select anything, but note the Comments).
Click **"OK"**.



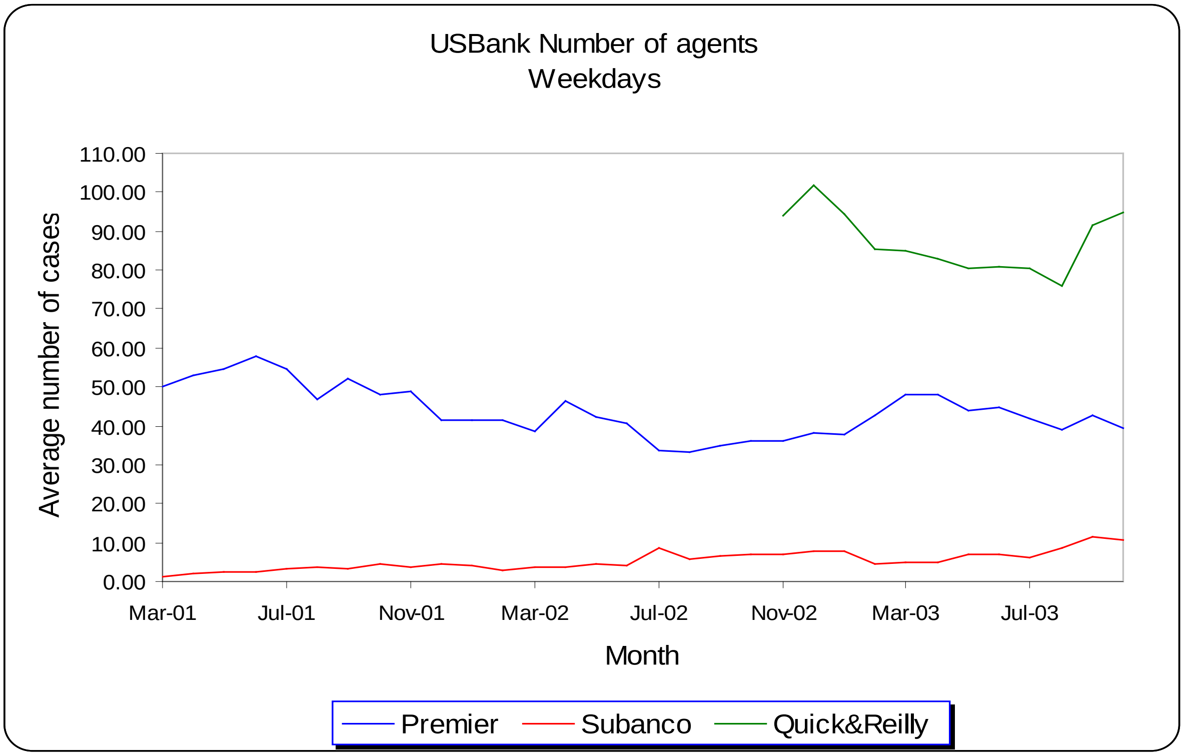
We first observe the weekly pattern in which weekdays have much higher arrivals than weekends (weekends are marked blue).
Note also that on February 12, the system stopped working at 4:00 PM, and February 17 was a holiday—Washington's birthday. This is manifested on the chart, where these special days are marked as Abnormal (green) and Holiday (red).

Return to the **"Statistical Models (Summaries)"** window via the **SEEStat** tab.
(A reminder: Click the **SEEStat** button on the task bar (left-bottom), click **"Windows"** on the main menu (at the top) and select **"Statistical Models (Summaries)"**).
Click the **"<-Tables"** button (top right).



From the variables list select **"Number of agents"** (the first option).
Open the **"Select Categories"** tab. Select the following **three** services: **"Premier"** (priority Retail service) **"Subanco"** (Spanish language) and **"Quick&Reilly"** (brokerage). (In order to do so, hold down the **Ctrl** key and click the three options, one by one.)

Now click the **"Dates->"** button. Mark **"Aggregated days by months"** and click **"Select all"**.
Open the **"Days"** tab and select **"Weekdays"**.
Click **"OK"**.



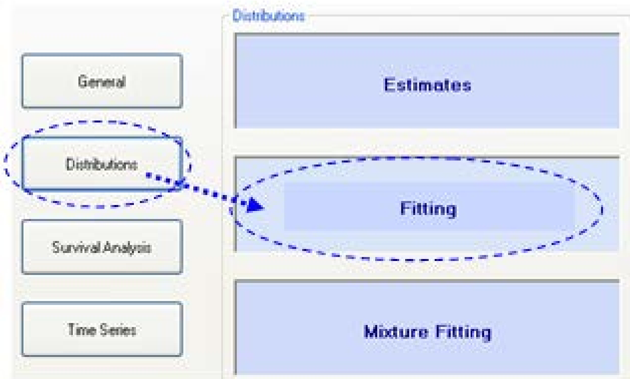
You see that one of the selected services (Quick&Reilly) was integrated into the Call Center of USBank only in November 2002.

Part 2

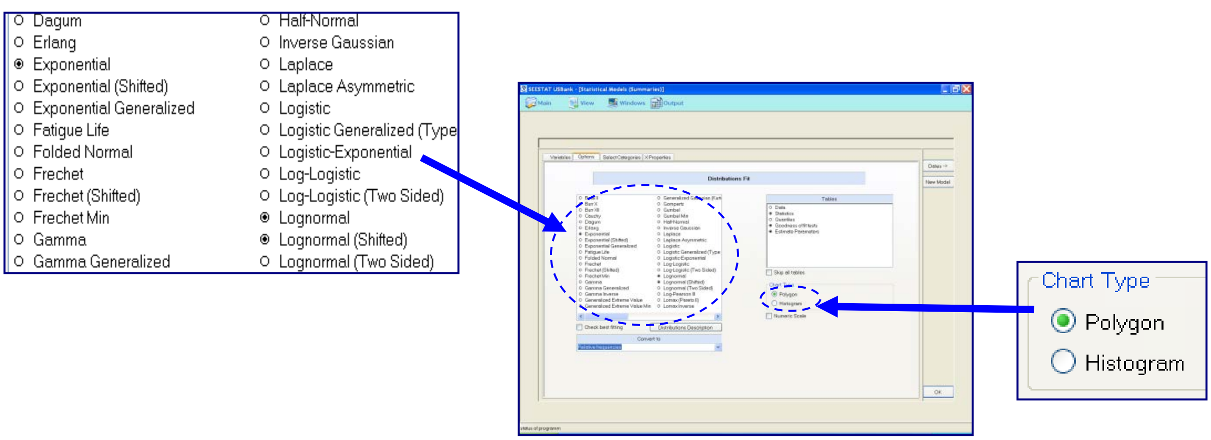
Example 2.1: Distribution fitting

We now fit a parametric service-time distribution to the service-time data from [Example 1.1]

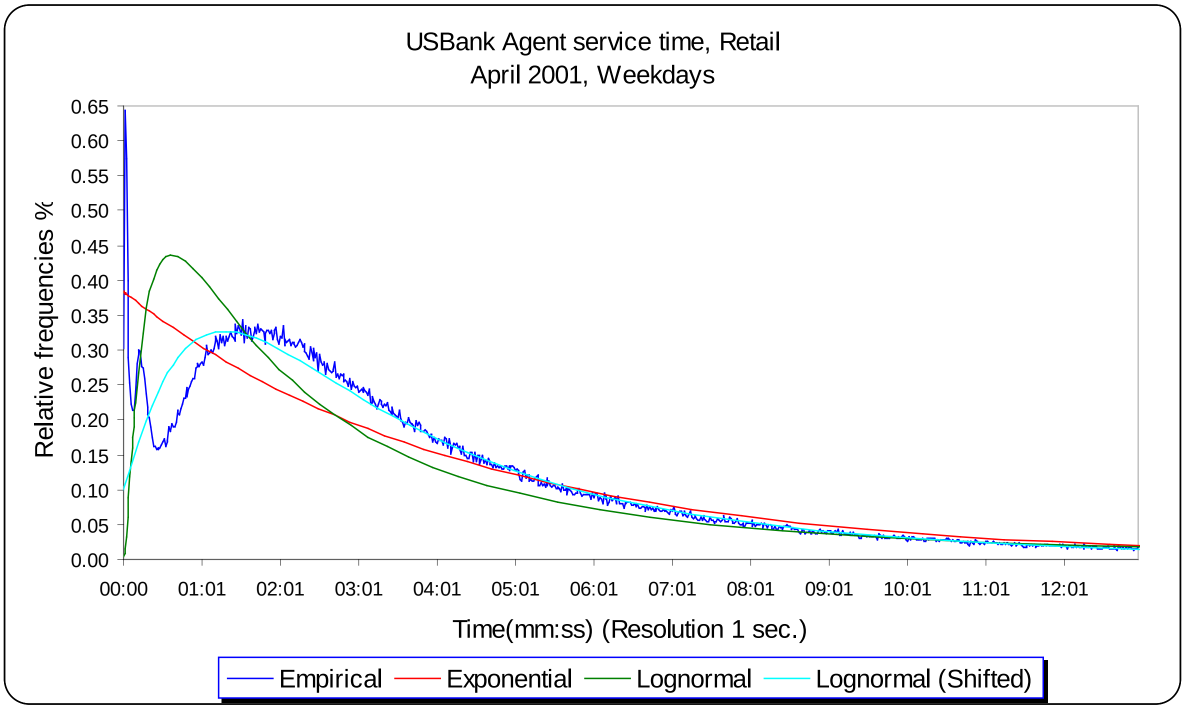
Open window **"Statistical Models (Summaries)"**. Click **"New Model"** and select **"Distributions"** and **"Fitting"**.



From the variables list select **"Agent service time"**.
Open the **"Options"** tab. You see the list of distributions available for fitting.
Mark simultaneously **3** of them: **Exponential, Lognormal,** and **Lognormal (Shifted)**.
Set chart type to **"Polygon"**.



Open the **“X Properties”** tab and set resolution to **00:01** = 1 second.
Click the **“Dates->”** button. Select **April 2001** and **“Aggregated days”**, open the **“Days”** tab and select **“Weekdays”**.
Click the **"<-Tables"** button.
On the **“Select Categories”** tab select **Retail.**
Click **“OK”**.



Observe again the irregularities near the origin. It looks as though there are at least three distributions involved: very short calls, abnormally short calls and, after around 30 seconds, the pattern looks rather regular. The best fit is produced by the Lognormal (Shifted) distribution, but clearly, close to the origin from the right, the fit is inadequate.

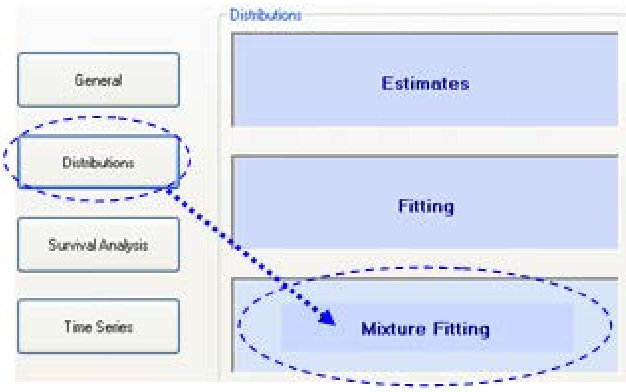
*You could use the tables on the previous sheet (the one accompanying the graph-sheet) to statistically validate the fit: **scroll down** until reaching the "**Parameter-Estimates**" and "**Goodness-of-Fit tests**" tables.*

Distribution	*Goodness-of-Fit Tests*				
	Residuals Std	*Kolmogorov-Smirnov*		*Cramer-von Mises*	
		Statistic	*p Value*	*Statistic*	*p Value*
Exponential	0.0333583	0.0648110	<.0001	688.91	<.0001
Lognormal	0.0504281	0.0878340	<.0001	1574.35	<.0001
Lognormal (Shifted)	0.0070425	0.0211673	<.0001	30.71	<.0001

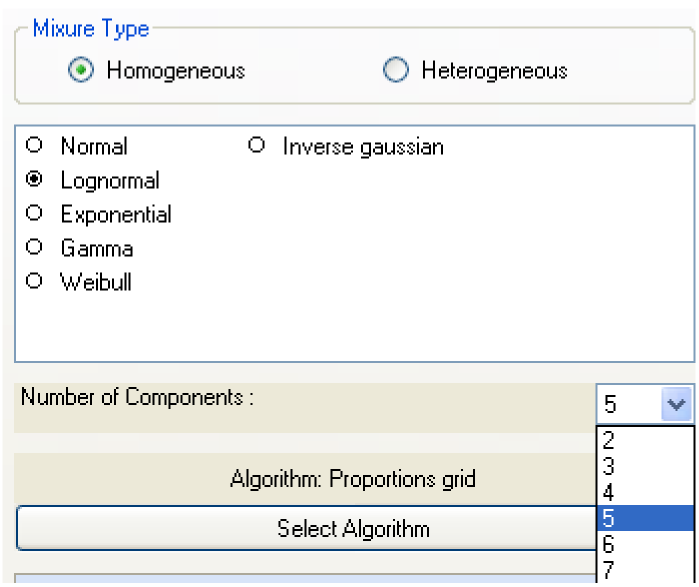
***Example 2.2*: Distribution mixture fitting**

We now try to accommodate the behavior to the right of the origin in the "Agent service time" histogram by a ***mixture*** *of distributions.*

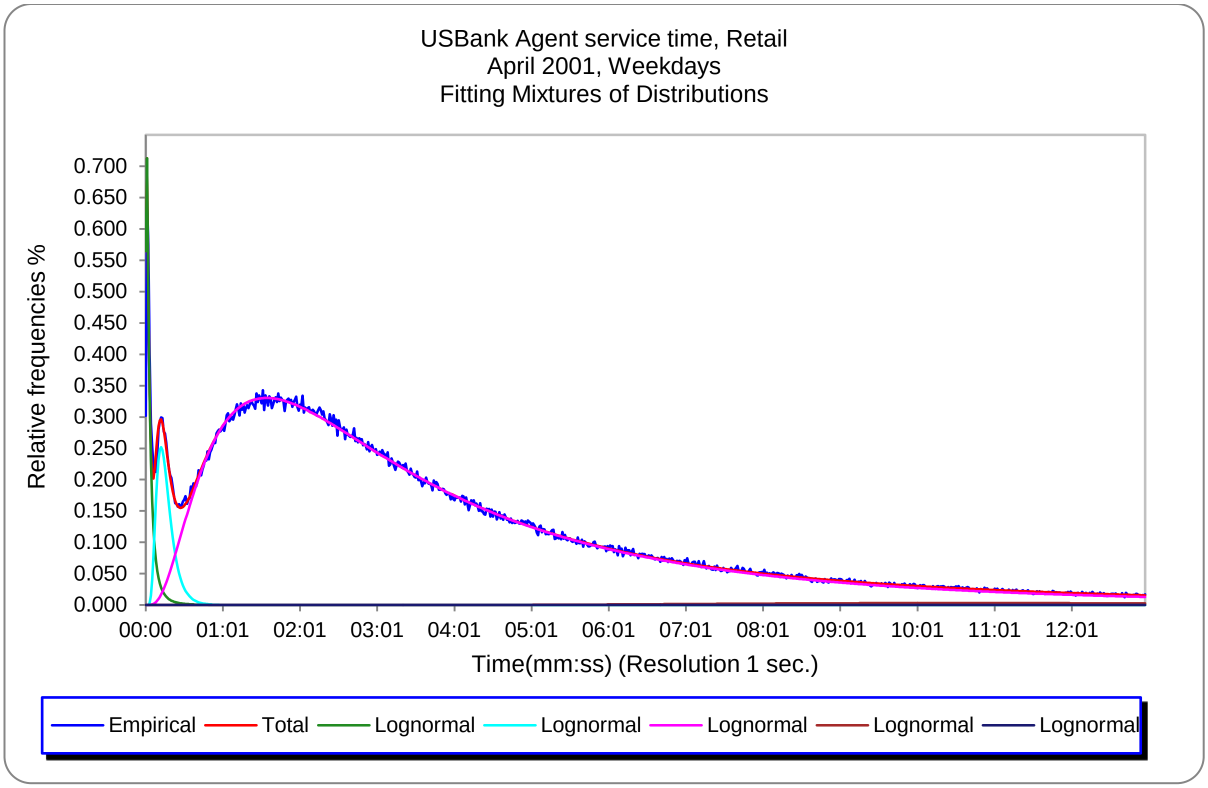
Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"New Model"**, select **"Distributions"** and **"Mixture fitting"**.



Open the **"Options"** tab. You can select a homogeneous or heterogeneous (mixture of various distributions) option. The former is the default. Select **"Lognormal"**. Set the number of mixture components to **5**, select chart type **Polygon**.



Click **"OK".**



You observe an excellent fit (Red line). In particular, on the left side (near the origin), there are two components, accommodating the very short and short calls.

Going to the previous Excel sheet, to view the corresponding Tables (by scrolling down), one notes that the main component has a weight of 91% in the mixture—its role in the chart is to fit the part beyond 30 seconds, which it does very well.

Parameter Estimates	
Components	*Mixing Proportions (%)*
1. Lognormal	3.19
2. Lognormal	3.55
3. Lognormal	91.09
4. Lognormal	1.83
5. Lognormal	0.34

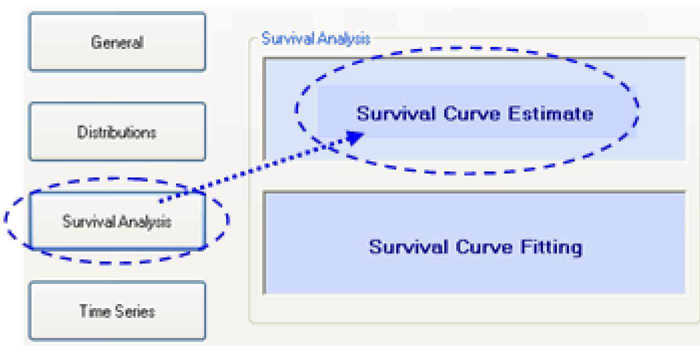
See Appendix A for a short explanation on Mixture fitting and the algorithms used for distribution fitting.

Example 2.3: ***Survival analysis with smoothing of hazard rates***

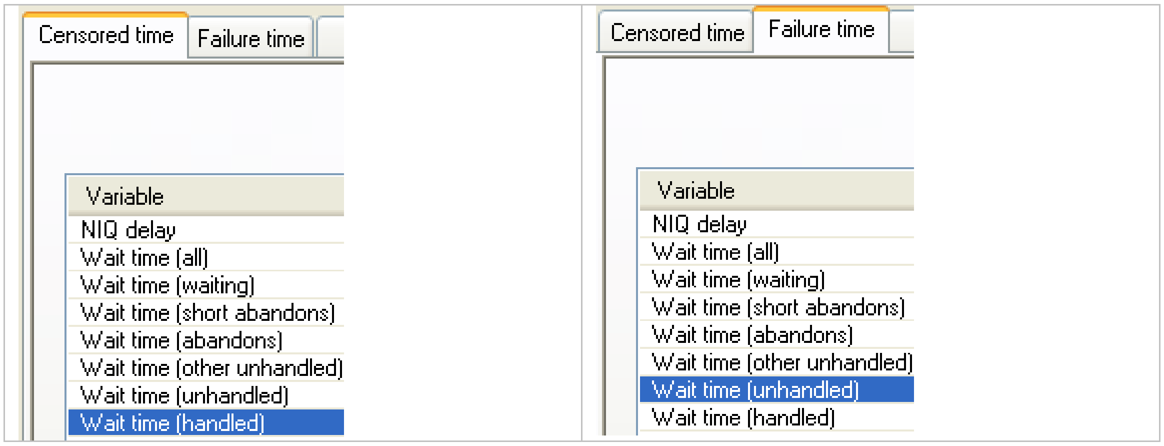
*SEEStat supports several survival models. These are required, for example, in order to achieve insight into customers' **(im)patience**, namely the time customers are willing to wait prior to hanging up. Indeed, for those customers who got served, their waiting time provides only a lower bound on how long they are willing to wait—their (im)patience constitutes **censored** observations. One must thus "uncensor" the data to produce adequate estimates. To this end, we now use simple tools from survival analysis. These will produce hazard-rate functions, which provide natural statistical summaries of (im)patience.*

Return, via **SEEStat**, to the **"Statistical Models (Summaries)"** window, click **"New Model"**.

Select **"Survival analysis"** and **"Survival Curve Estimate"**.

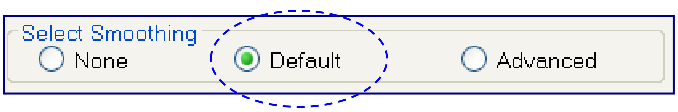


There are two variable tabs. The first tab **"Censored time"** is open. Select **"Wait time (handled)"**: this corresponds to the waiting time of the customers who received service. Open the **"Failure time"** tab and select **"Wait time (unhandled)":** this corresponds to the waiting of customers who joined the queue but did not receive service (mainly due to abandonment, though there are sometimes other reasons such as system malfunction).

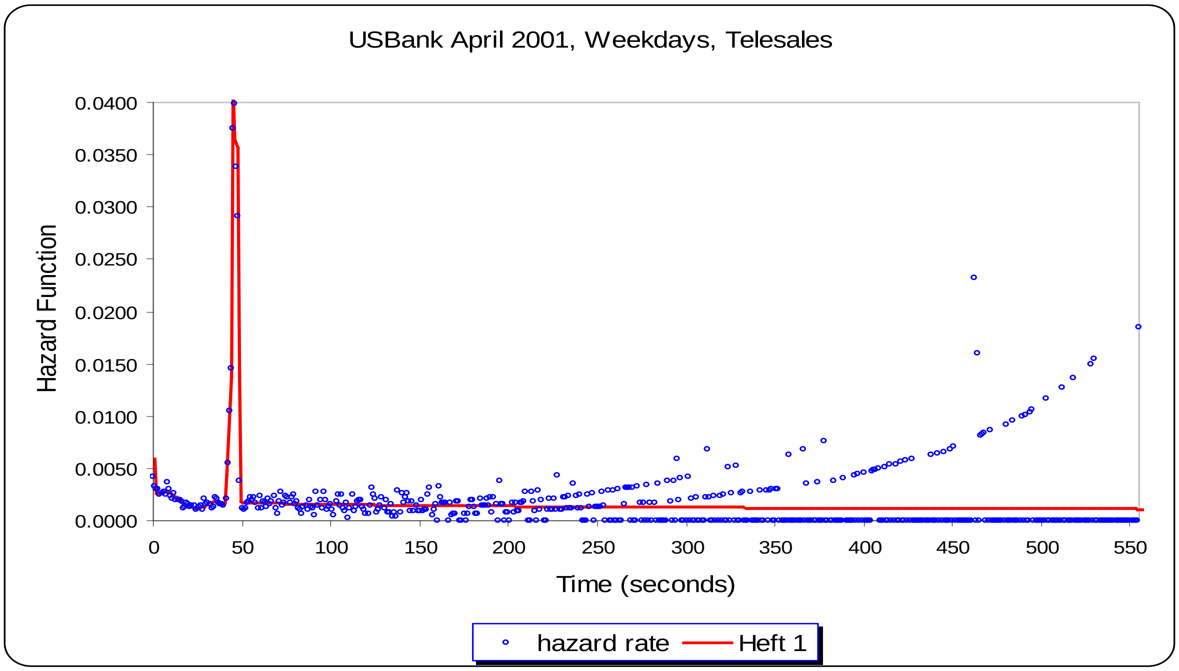


Open the **"Options"** tab. SEEStat supports several methods of smoothing, which are applicable to hazard rates and beyond. Select **"Default"** smoothing (which, this time, happens to be the method of HEFT).

See Appendix B for a reference list on smoothing methods.



From the tab **"Select categories"** select **"Telesales"**.
Click **"Dates"**. Select **"April 2001"** and on the tab **"Days"** select "**Weekdays**".
Click **"OK".**

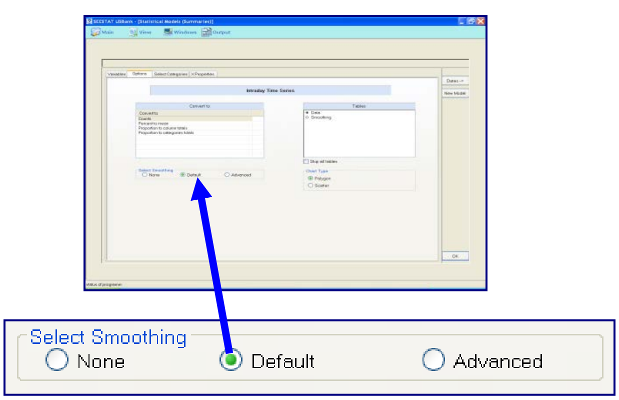


A noticeable peak in the hazard rate indicates that there is a trigger for customers to abandon after about 50 seconds of waiting (which, based on our experience, could be the result of a voice-announcement at that time: such announcements, regardless of their content, “reminds” customers of their wait and thus increase their likelihood of abandonment).

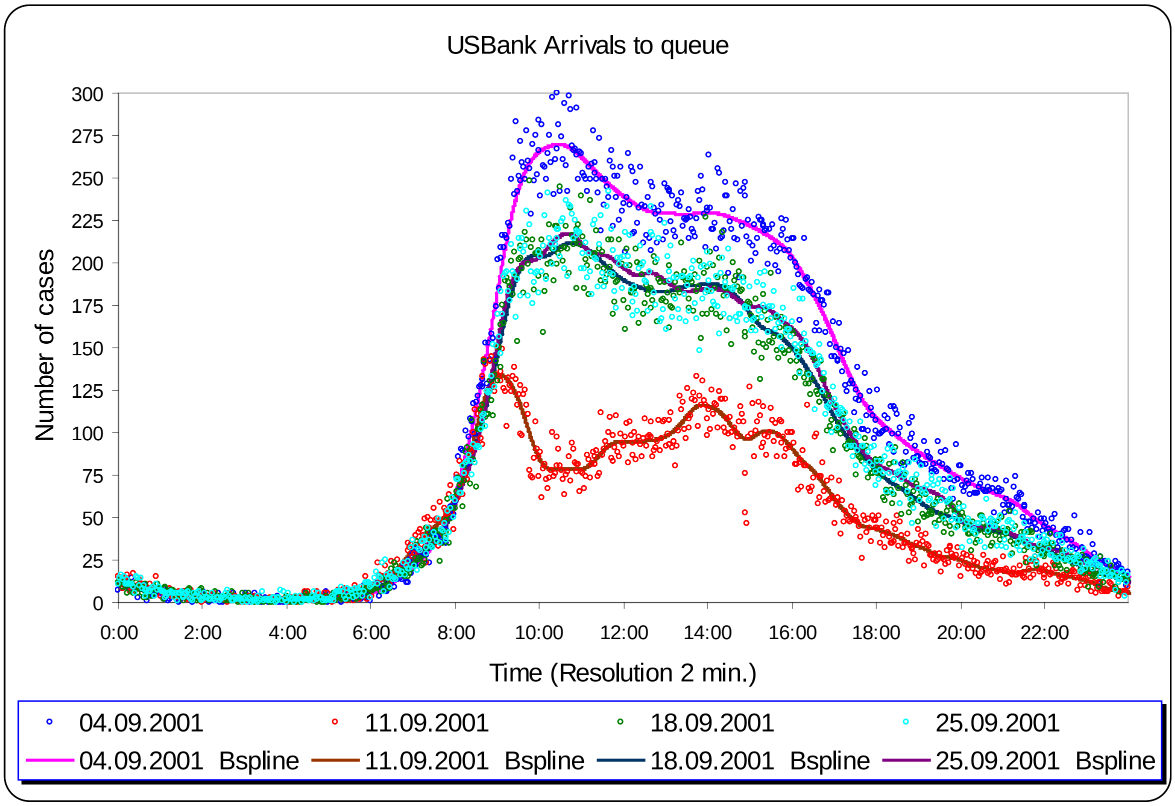
<u>Example 2.4:</u> Smoothing of intraday time series

Smoothing algorithms are available for several statistical models. We now demonstrate the application of smoothing on the data used in <u>Example 1.2</u>.

Return as usual to **"Statistical Models (Summaries)"**, click **"New Model"**, select **"Time Series"** and **"Intraday"**. Select **"Arrivals to queue"**. In **"Options"** tab select **“Default”** smoothing (this time, the default is the method of Cubic Splines).

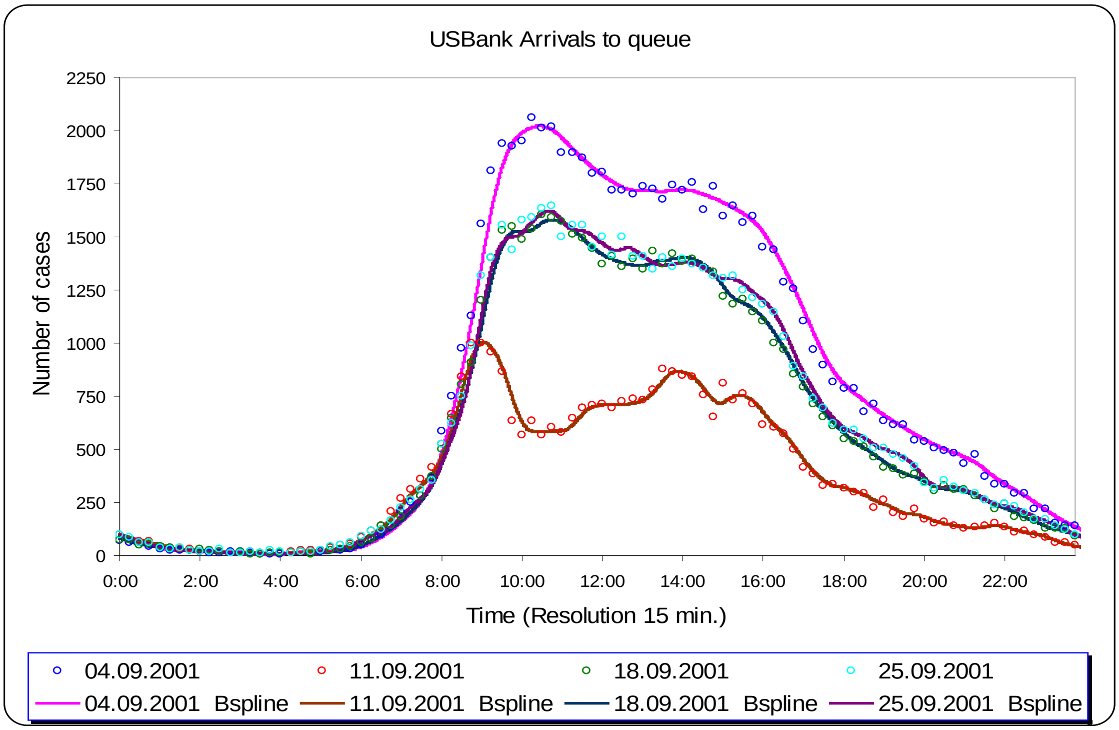


Select **"Scatter"** as chart type.
In the **"X Properties"** tab, set resolution to **02:00** = 2 minutes.
Click **"Dates"**, mark **"Individual days"** and select **"September 2001"**.
In the **"Days"** tab, if not already selected, select (with **"Ctrl"** and click) all four **Tuesdays** of September.
Click "**OK**"



For this small resolution of 2 minutes, there is plenty of noise, but the smoothed data clearly identifies the regular pattern that was discovered before. (Note that the smoothed curves are computed with the minimal resolution for this variable, which is 30 seconds; the 2-minute resolution is only for display.)

Click **"Output"** on the main menu, then click **"Modify Tables and Charts"**.
Open the **"Properties"** tab, set resolution to **15 min** and click **"OK"**.



The Averaged Data (over 15 minutes) is now much closer to the smoothed curves, as expected.

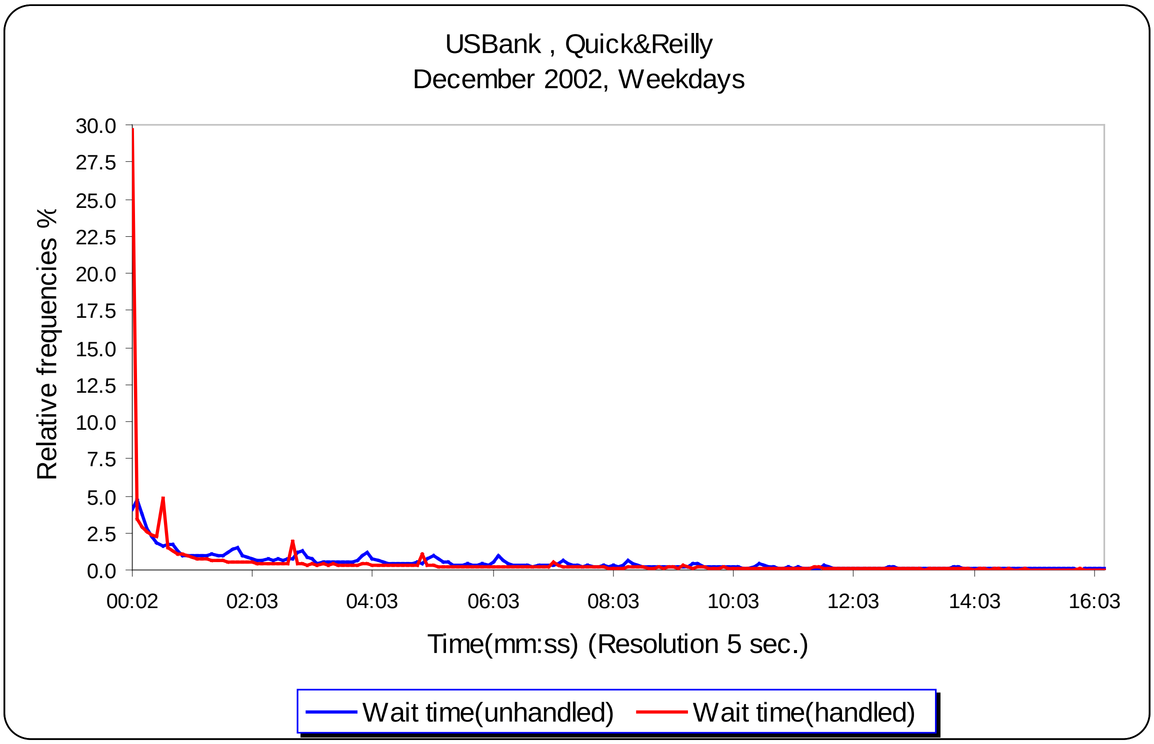
Part 3

Some additional interesting examples.

<u>Example 3.1:</u> Queue regulated by a protocol & announcements

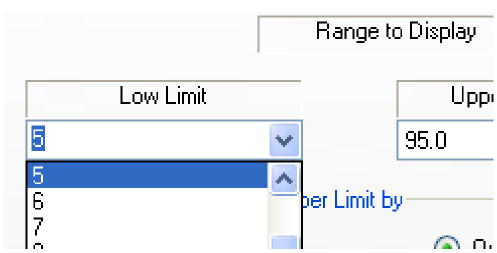
Via **SEEStat** return to the **"Statistical Models (Summaries)"** window, click **"New Model"**, then click the **"Distributions**" button. Three available distribution models appear. Select **"Estimates"**. In the **"Variables"** tab select (using **Ctrl**) both **"Wait time (unhandled)"** and **"Wait time (handled)".**

In the **"Options"** tab select chart type **Polygon**. Click "**Dates->"**, select **December 2002**, make sure the **"Aggregated days"** option is selected, and in **"Days"** select **Weekdays**. Click **"<-Tables". In "Select Categories"** select **"Quick&Reilly".** Click "**OK**".

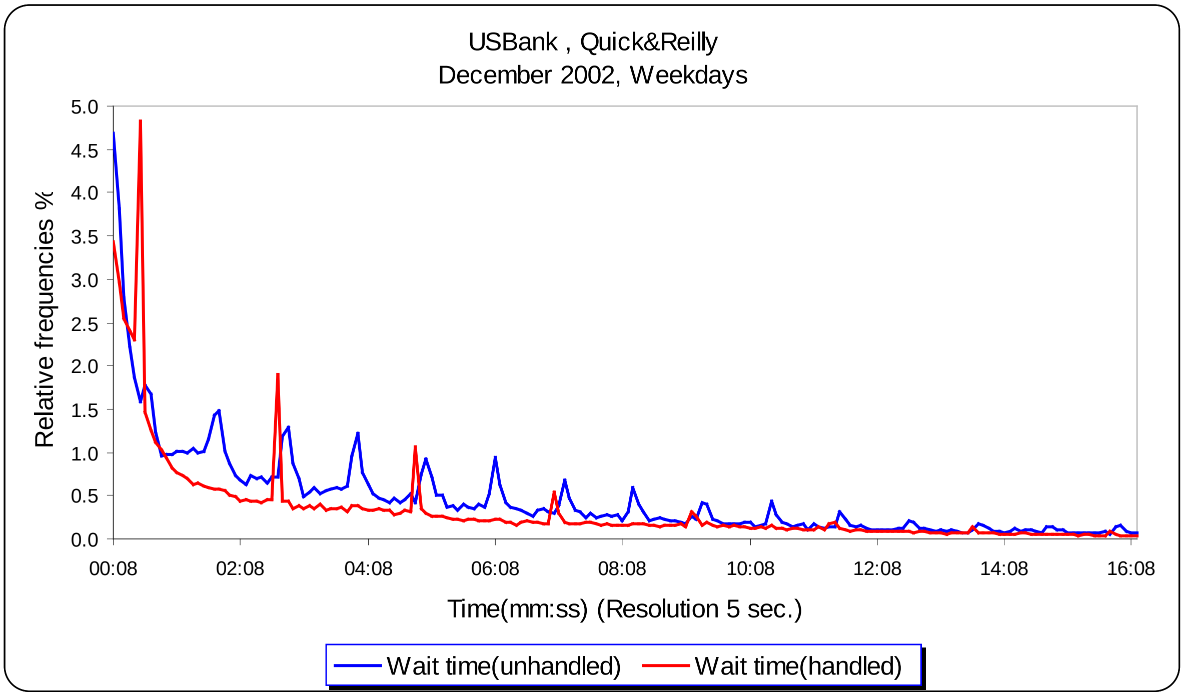


Both lines are periodical. To get a better focus, you will cut the chart on the left side.

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
Open **"Properties"**, set the low limit to **5 seconds**.



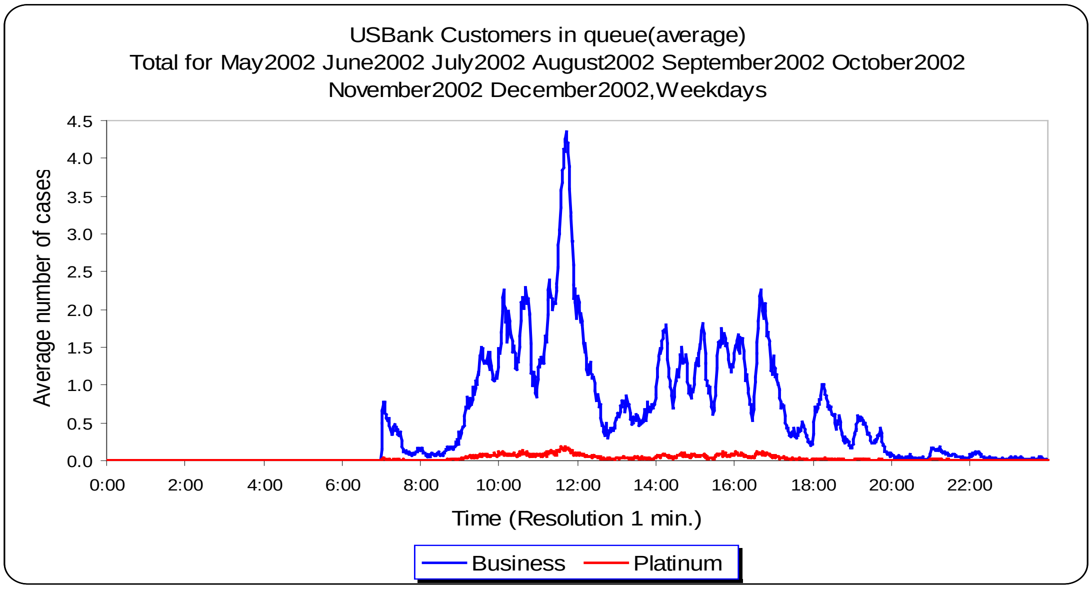
Click **"OK"**.



As you see, the Wait time (unhandled), in blue, peaks every 65 sec. The Wait time (handled), in red, peaks every 130 seconds. These interesting observations are yet to find their explanations. (Our experience suggests that peaks in the "Wait-time (unhandled)" are "psychological", for example a reaction of a customer to an announcement; and peaks in the "Wait-time (handled)" are "protocol-driven", for example a result of a priority upgrade.)

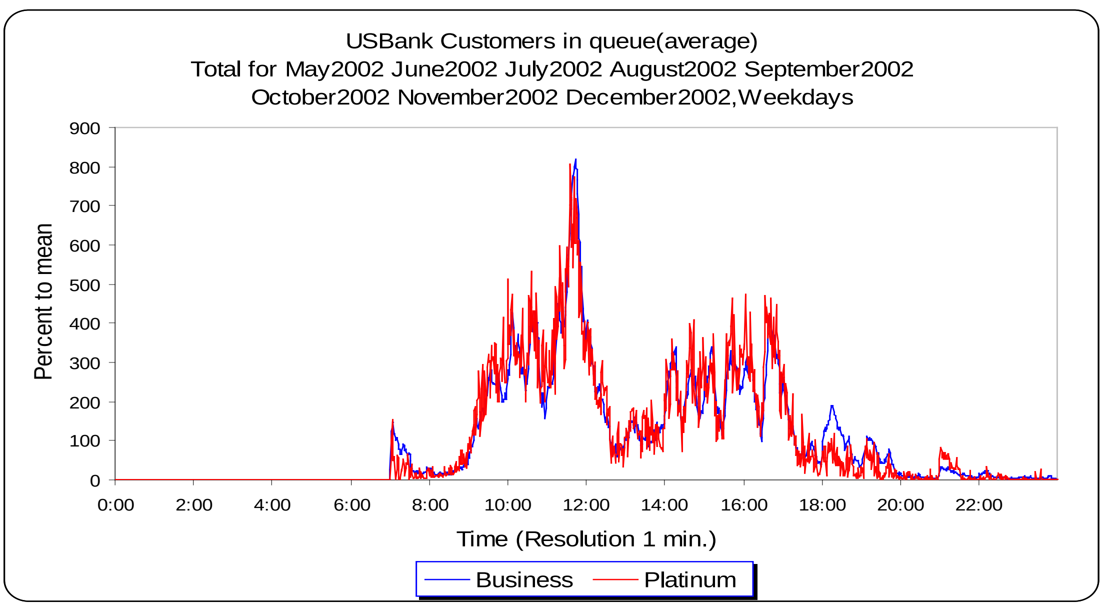
***Example 3.2*: Queue length & state-space collapse**

Via **SEEStat** return to the **"Statistical Models (Summaries)"** window,
click **"New Model"**. Click the **"Time Series"** button and select **"Intraday"**.
In the **"Variables"** tab select **"Customers in queue (average)"**.
In the **"Options"** tab select smoothing **"None"** and chart type **"Polygon".**
In the **"X Properties"** tab select resolution **1 minute.**
In **"Select Categories"** tab select (with **Ctrl** and click) **Business** and **Platinum**.
Click **"Dates->"**, select **"Dates totals only",** select the 8 months from **May 2002** to **December 2002** and select **Weekdays** in the **"Days"** tab. Click **"OK"**.



Platinum is a small-scale service. You will now normalize the chart in order to identify patterns.

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.
Open the **"Options"** tab and select **Percent to mean**. Click **"OK"**.

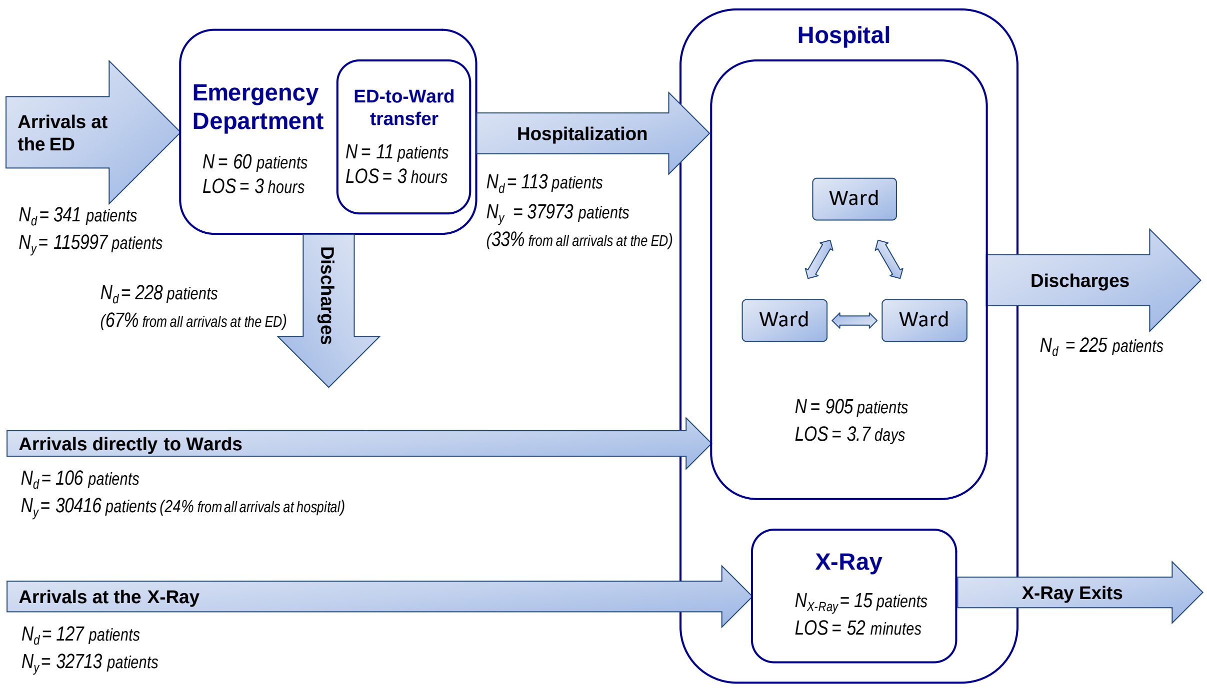


Note the essentially overlapping patterns of the queue lengths of the two customer types. (This phenomenon is predicted by asymptotic analysis of queues in heavy traffic, where it is referred to as State-Space-Collapse.)

Exit SEEStat, either via the "X" on the top-right corner, or by clicking "Close SEEStat" in the Main menu. (Don't exit the terminal.)

HomeHospital Data

<u>Background</u>: *The data we rely on was collected at a large Israeli hospital. This hospital consists of about 1000 beds and 45 medical units. The data includes detailed information on patient flow throughout the hospital, over a period of several years (January 2004–October 2007). In particular, the data allows one to follow the paths of individual patients throughout their stay at the hospital, including admission, discharge, and transfers between hospital units. The data does not acknowledge resolutions within the ED or within wards.*



N_d - average number of patients that arrived per weekday, for period January 1, 2004–October, 31, 2007

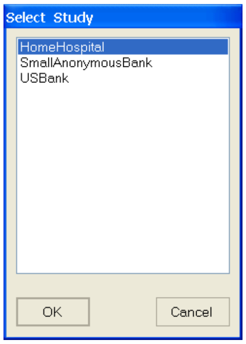
N_y - average number of patients that arrived per year, for years 2004, 2005, 2006, all days (for year 2007 data not fully completed; missing two months—November and December).

N - average number of patients in ED/ED-to-Ward transfer/Wards, recorded at 12:00 per weekday, for period January 1, 2004–October, 31, 2007

$N_{X\text{-}Ray}$ - average number of patients in X-Ray at 10:00 per weekday, for period January 1, 2004–October, 31, 2007

LOS - length of stay in ED/ED-to-Ward transfer/Wards/X-Ray per weekday, for period January 1, 2004–October, 31, 2007

Reopen SEEStat 3.0 and select **HomeHospital** study.



Part 4: Hospital

Example 4.1: ***Arrivals - Average per one weekday over entire month***

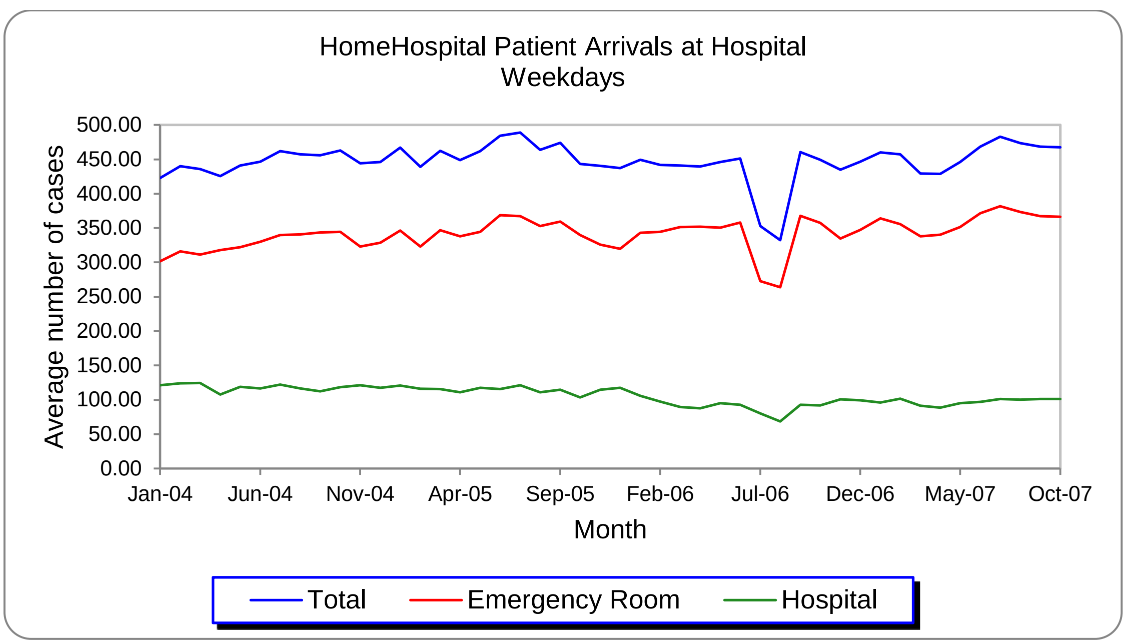
Click **"Main"** and **"Statistical Models (Summaries)".** Select **"Time Series"**, then **"Daily totals"**.

From the variables list select **"Patient Arrivals at Hospital".**

In the **"Select Categories"** tab, select all categories.

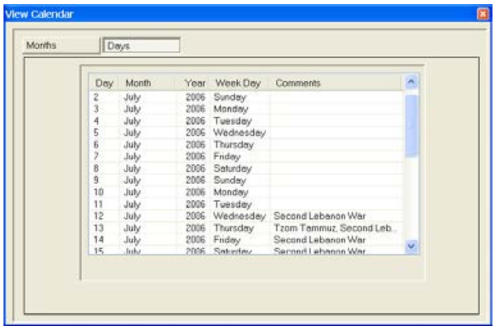
Click the **"Dates->"** button. Click the **"Select all"** button, open tab **"Days"** and select **"Weekdays"**.

Click **"OK"**.



We see a drop in the number of arrivals during July and August 2006. The reason for this phenomenon is the "second Lebanon war" which took place during that period and affected mostly the northern part of Israel in which the hospital is located. This could be verified via ***Calendar****, in which special days (such as holidays) and special events are noted.*

Click **View**-> **Calendar**. Mark **"Individual days"** and select **July 2006.** Open tab **"Days"**.

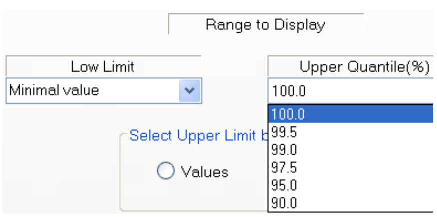


Click **"Months"** tab and select **August 2006.** Open tab **"Days"**.

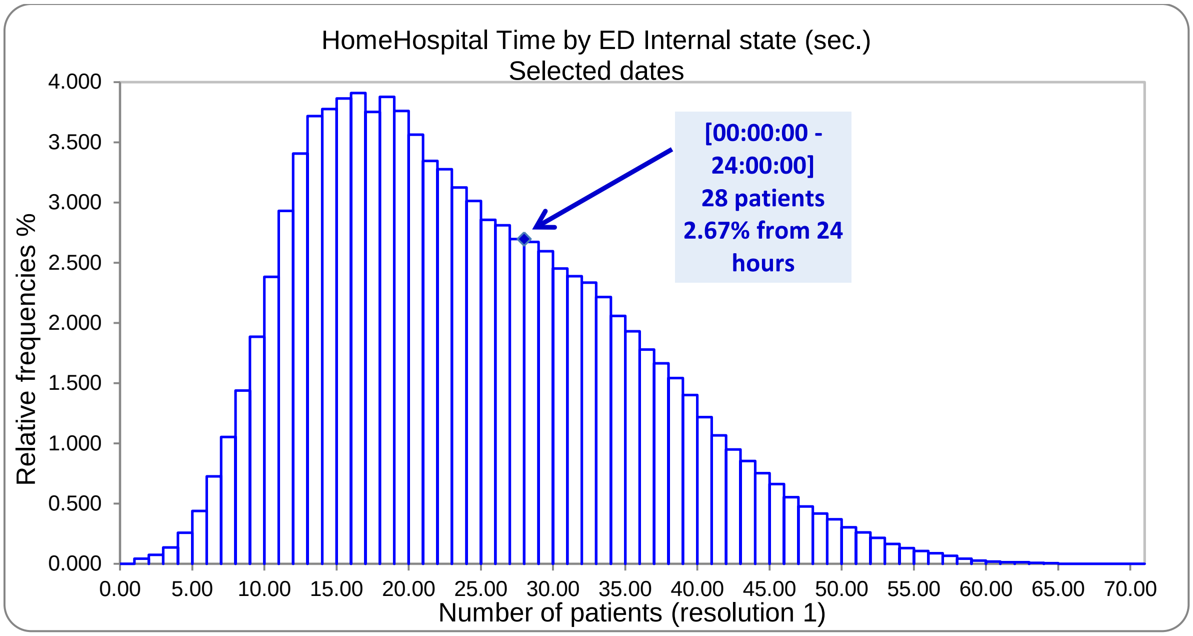
Part 5: Emergency Department

***Example 5.1**: Time by ED Internal state (sec.), or equivalently ED census - Distribution during all 24 hours of the day*

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Estimates"**.
In the **"Variable"** tab, select **"Time by ED Internal state (sec.)"**.
In the **"Select Categories"** tab, select **"Total"**.
In the **"X Properties"** tab change **upper quantile** limit to **100.**



Click the **"Dates->"** button. Select **"Dates totals only"** and all months from **January 2004 to October 2007,** open the **"Days"** tab and select **"All days"**.
Click **"OK"**.



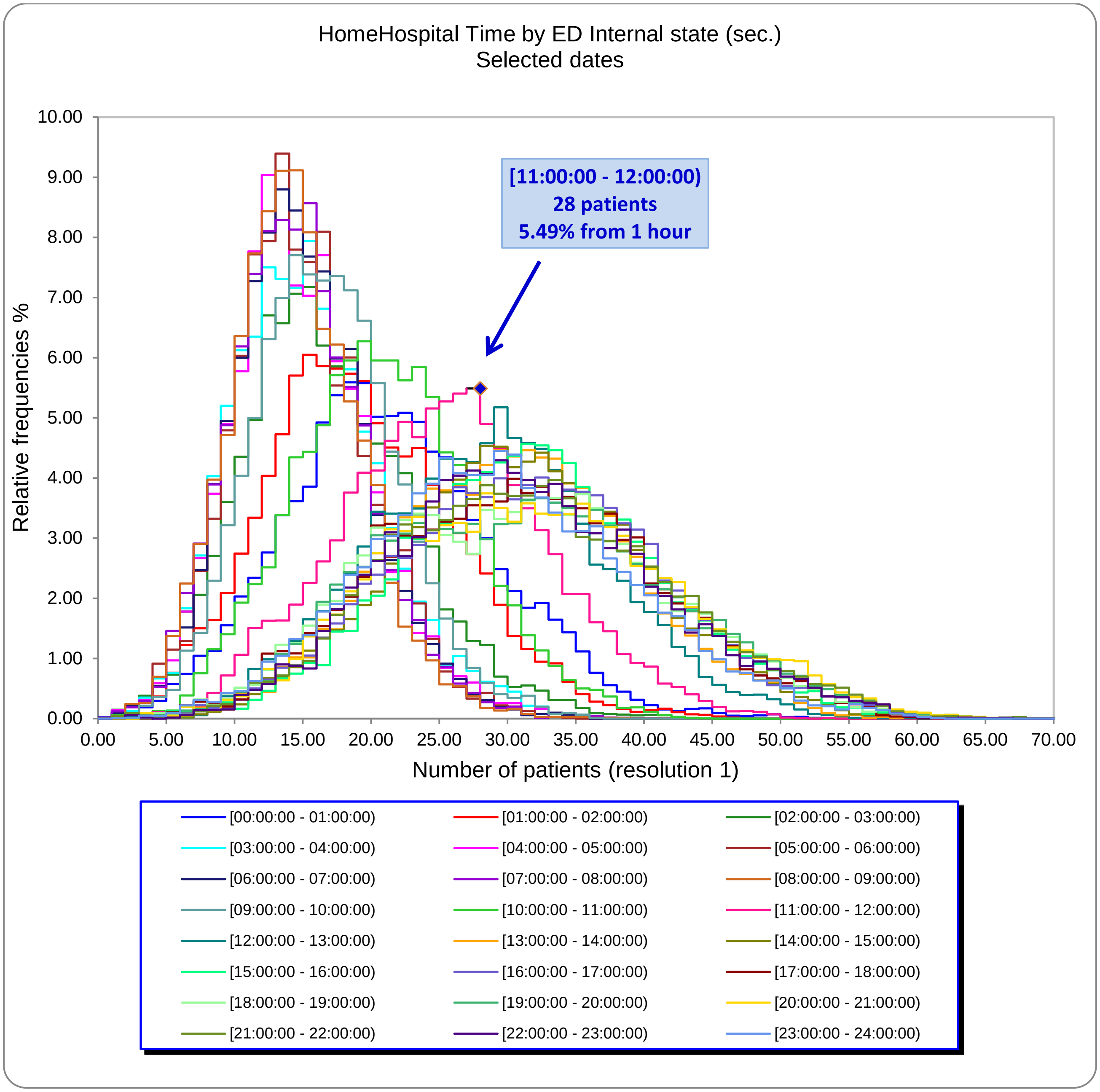
*For example: 28 patients were in Internal ED during 38 minutes and 30 seconds (2.674% from 24 hours: 0.02674*86400 sec=2310 sec) between 00:00 and 24:00.*

Statistics	
	Time by ED Internal state (sec.)
N	120960000
N(average per day)	86400
Mean	23.63
Standard Deviation	10.7
Variance	114.4
Median	22
Minimum	0
Maximum	70

We observe that the distribution has an unusual shape; it is skewed to the left, has a light right tail and has two local peaks. We shall now further investigate it via Example 5.2.

<u>Example 5.2</u>: Time by ED Internal state (sec.), or equivalently ED census - Distribution during each of the 24 hours per day

Return to the **"Statistical Models (Summaries)"** window.
Click the **"<-Tables"** button.
In the **"Select Categories"** tab, select (with shift key) all categories except **"Total"**.
Click **"OK"**.



*For example: 28 patients were in Internal ED during 3 minutes and 17 seconds (5.49% from 1 hour: 0.0549*3600 sec=197 sec) between 11:00 and 12:00*

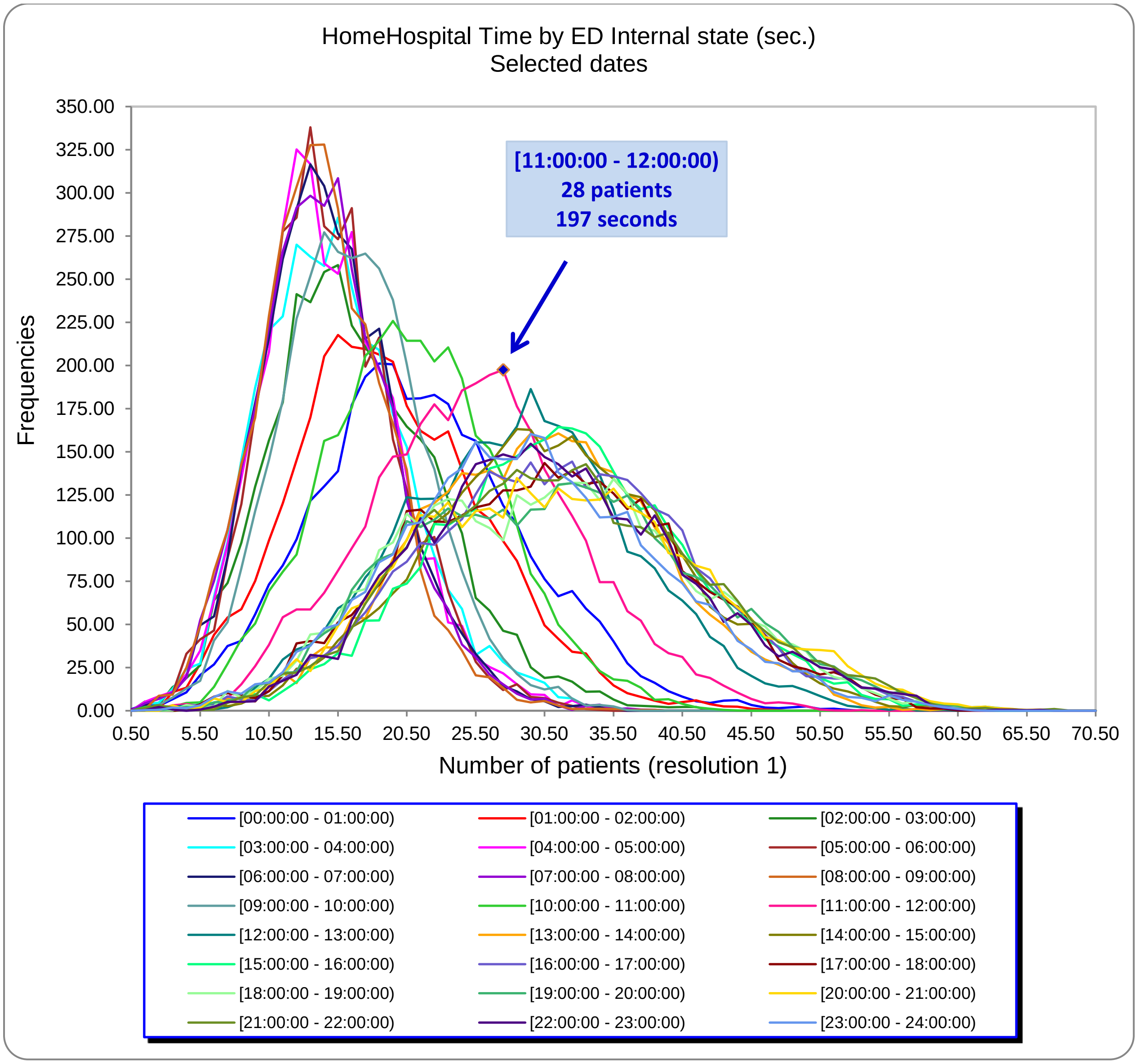
Statistics

	N	N(average per day)	Mean	Standard Deviation	Variance	Median	Minimum	Maximum
[00:00:00-01:00:00)	5040000	3600	20.93	7.6	58.09	20	0	52
[01:00:00 - 02:00:00)	5040000	3600	18.74	6.9	48.6	18	0	48
[02:00:00 - 03:00:00)	5040000	3600	16.35	6.1	37.67	16	1	42
[03:00:00 - 04:00:00)	5040000	3600	14.86	5.3	29.15	14	1	40
[04:00:00 - 05:00:00)	5040000	3600	14.43	5.0	25.26	14	1	35
[05:00:00 - 06:00:00)	5040000	3600	14.45	4.9	24.05	14	1	34
[06:00:00 - 07:00:00)	5040000	3600	14.38	4.8	23.96	14	1	36
[07:00:00 - 08:00:00)	5040000	3600	14.24	4.9	24.55	14	1	38
[08:00:00 - 09:00:00)	5040000	3600	14	4.7	22.62	14	1	39
[09:00:00 - 10:00:00)	5040000	3600	16.17	5.1	26.95	16	1	36
[10:00:00 - 11:00:00)	5040000	3600	20.36	6.4	41.23	20	1	44
[11:00:00 - 12:00:00)	5040000	3600	24.75	7.6	59.01	25	1	53
[12:00:00 - 13:00:00)	5040000	3600	28.02	8.4	71.95	28	2	67
[13:00:00 - 14:00:00)	5040000	3600	29.73	8.8	78.45	30	4	65
[14:00:00 - 15:00:00)	5040000	3600	30.59	8.9	80.57	30	4	62
[15:00:00 - 16:00:00)	5040000	3600	31.2	9.1	83.76	31	4	62
[16:00:00 - 17:00:00)	5040000	3600	31	9.6	93.63	31	4	64
[17:00:00 - 18:00:00)	5040000	3600	30.2	9.8	97.69	30	4	64
[18:00:00 - 19:00:00)	5040000	3600	30.16	10.2	104.93	30	3	64
[19:00:00 - 20:00:00)	5040000	3600	30.62	10.5	110.18	30	2	64
[20:00:00 - 21:00:00)	5040000	3600	31.14	10.5	110.41	31	2	67
[21:00:00 - 22:00:00)	5040000	3600	31.08	10.16	103.26	31	1	70
[22:00:00 - 23:00:00)	5040000	3600	30.58	9.7	94.25	30	1	66
[23:00:00 - 24:00:00)	5040000	3600	29.02	9.5	91.04	29	1	61

Click **"Output"** on the main menu and then **"Modify Tables and Charts"**.

In the **"Options"** tab, under the **"Convert to"** select **"Frequencies"** and select chart type **"Polygon".**

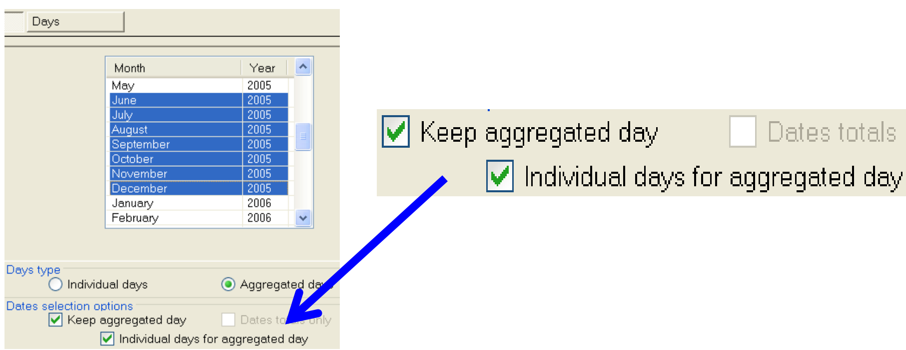
Click **"OK"**.



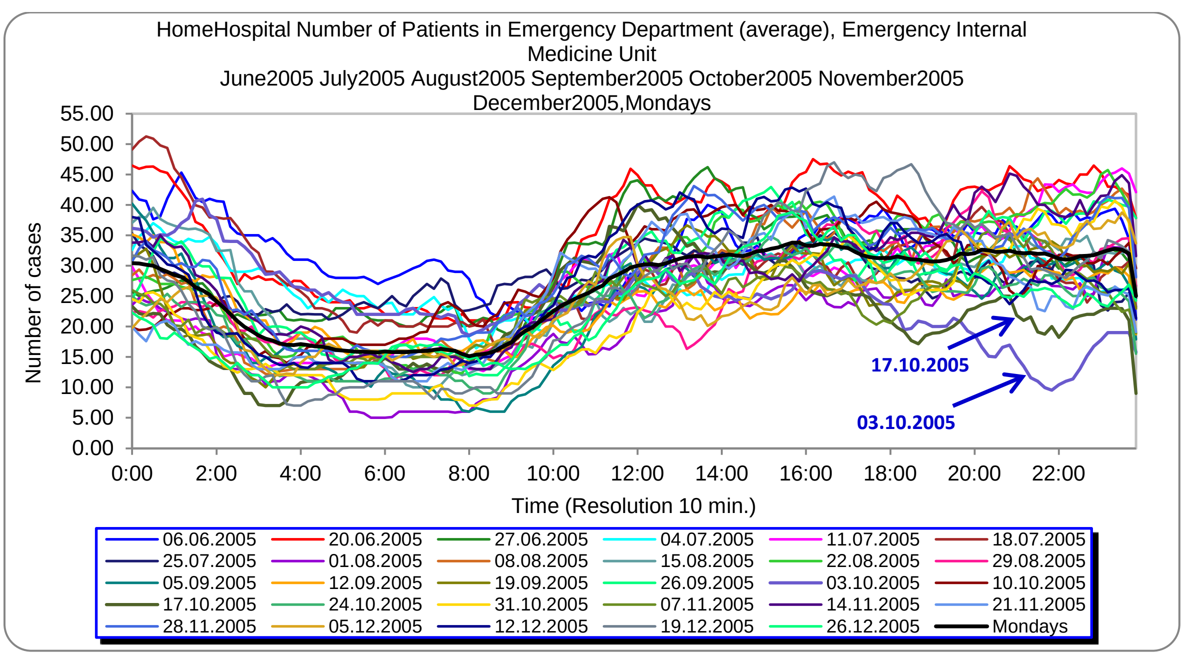
We observe that the distribution of the number of patients in the ED at the time of the day t, (for t = 0, 1, ..., 23) is normal, with mean and variance that vary over time (this was confirmed by the appropriate statistical tests, see Example 5.4). Using these figures, we identify three main patterns that compose the distribution found in Example 5.1: (1) From 02:00 until 09:00, where the average number of patients is around 15; (2) From 12:00 until 22:00, where the average number of patients is about 35; and (3) The rest of the day (09:00–12:00, 22:00–02:00), when the distribution shifts from one group to the other.
This is further observed in Example 5.3.

Example 5.3: Number of patients in Internal ED - Average per 10-minute intervals, only on Mondays during 2005

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Time Series"**, then **"Intraday"**.
In the **"Variable"** tab, select **"Number of Patients in Emergency Department (average)"**.
In the **"Select Categories"** tab, select **"Emergency Internal Medicine Unit"**.
Click the **"Dates->"** button.
Mark **"Individual days for aggregated day".**
Select months from **June 2005** to **December 2005**.



Open the **"Days"** tab and select **"Mondays"**.
Click **"OK"**.

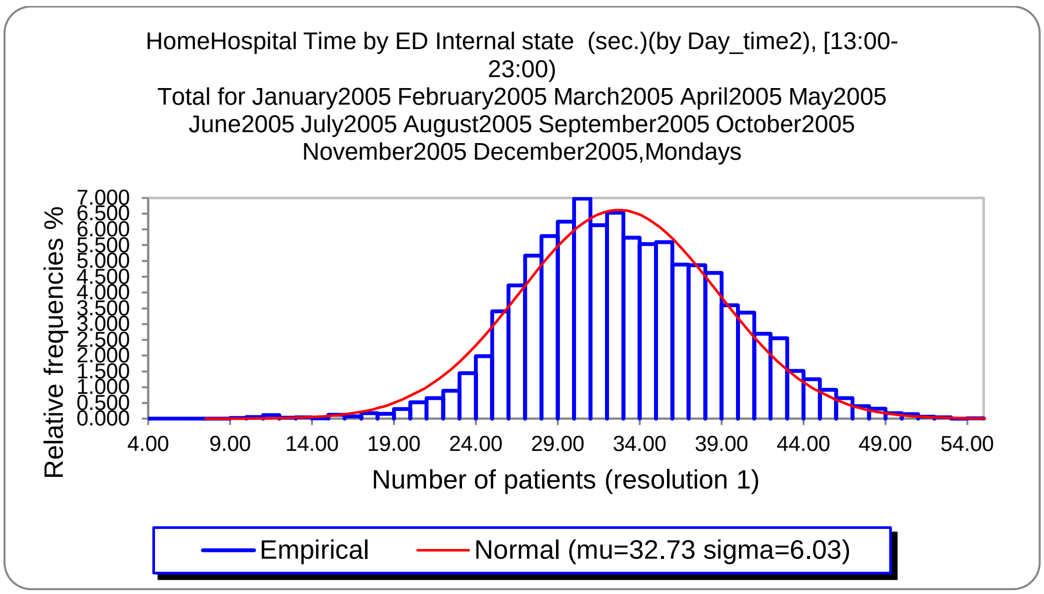


Example 5.4: Time by ED Internal state (sec.) - Fitting distribution during "evening" hours, on Mondays, 2005

Return to the **"Statistical Models (Summaries)"** window.

Click the **"New Model"** button. Select **"Distributions"**, then **"Fitting"**. In the **"Variable"** tab, select **"Time by ED Internal state (sec.) (by Day_time2)"**. In the **"Options"** tab select **Normal** distribution. In the **"Select Categories"** tab, select **"[13:00–23:00)"**. Click the **"Dates->"** button.
Mark **"Dates totals only"**. Select months from **January 2005** to **December 2005**.

Open tab **"Days"** and select **"Mondays"**.
Click **"OK"**.

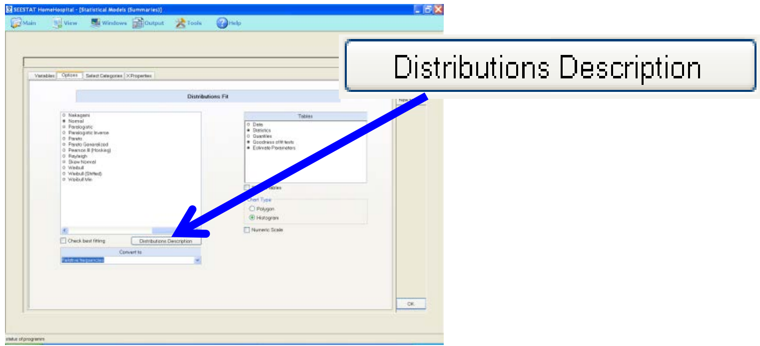


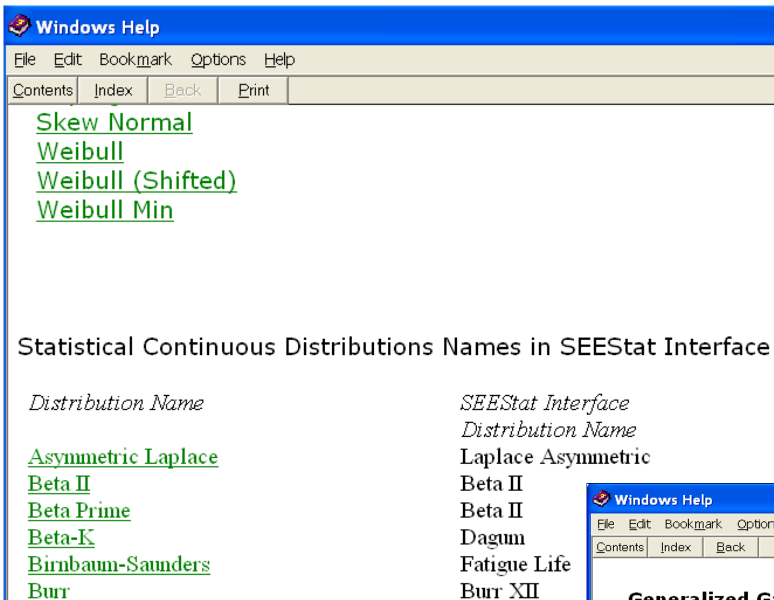
Statistics	
	Time by ED Internal state (sec.)
N	1836000
N(average per day)	36000
Mean	32.73
Standard Deviation	6.029
Variance	36.35
Median	32

Parameters for Normal Distribution	
Parameter	*Estimate*
mu	32.73
sigma	6.03
mean	32.73
std	6.029

Goodness-of-Fit Tests for Normal Distribution			
Test	*Statistic*	*DF*	*p Value*
Residuals Std	0.0209		
Kolmogorov-Smirnov	0.0587		<.0001
Cramer-von Mises	801.9579		<.0001
Anderson-Darling	4532.0570		<.0001
Chi-Square	>1000	41	<.0001

__Note__: probability density function formulas for 50 continuous distributions can be found under the "Options" tab, by clicking the "Distributions Description" button.




Windows Help

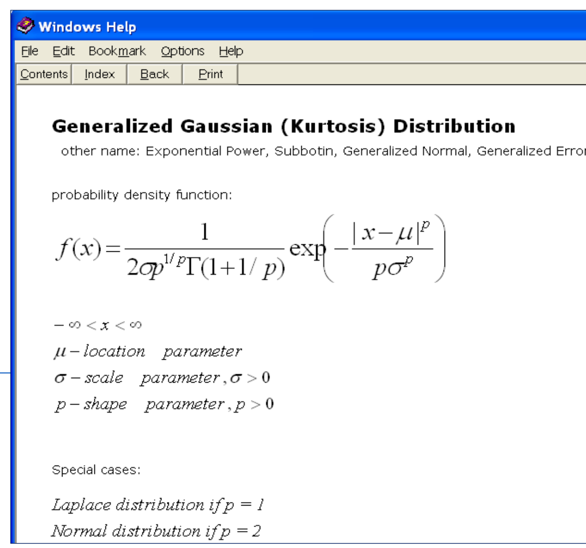
File Edit Bookmark Options Help

Contents Index Back Print

Skew Normal
Weibull
Weibull (Shifted)
Weibull Min

Statistical Continuous Distributions Names in SEEStat Interface

Distribution Name	*SEEStat Interface Distribution Name*
Asymmetric Laplace	Laplace Asymmetric
Beta II	Beta II
Beta Prime	Beta II
Beta-K	Dagum
Birnbaum-Saunders	Fatigue Life
Burr	Burr XII


Windows Help

File Edit Bookmark Options Help

Contents Index Back Print

Generalized Gaussian (Kurtosis) Distribution

other name: Exponential Power, Subbotin, Generalized Normal, Generalized Error

probability density function:

$$f(x) = \frac{1}{2\sigma p^{1/p}\Gamma(1+1/p)} \exp\left(-\frac{|x-\mu|^p}{p\sigma^p}\right)$$

$-\infty < x < \infty$

μ – *location parameter*

σ – *scale parameter*, $\sigma > 0$

p – *shape parameter*, $p > 0$

Special cases:

Laplace distribution if $p = 1$

Normal distribution if $p = 2$

Part 6: Medical Wards

Example 6.1: *LOS in Internal Medicine (in days) – Distribution fitting*

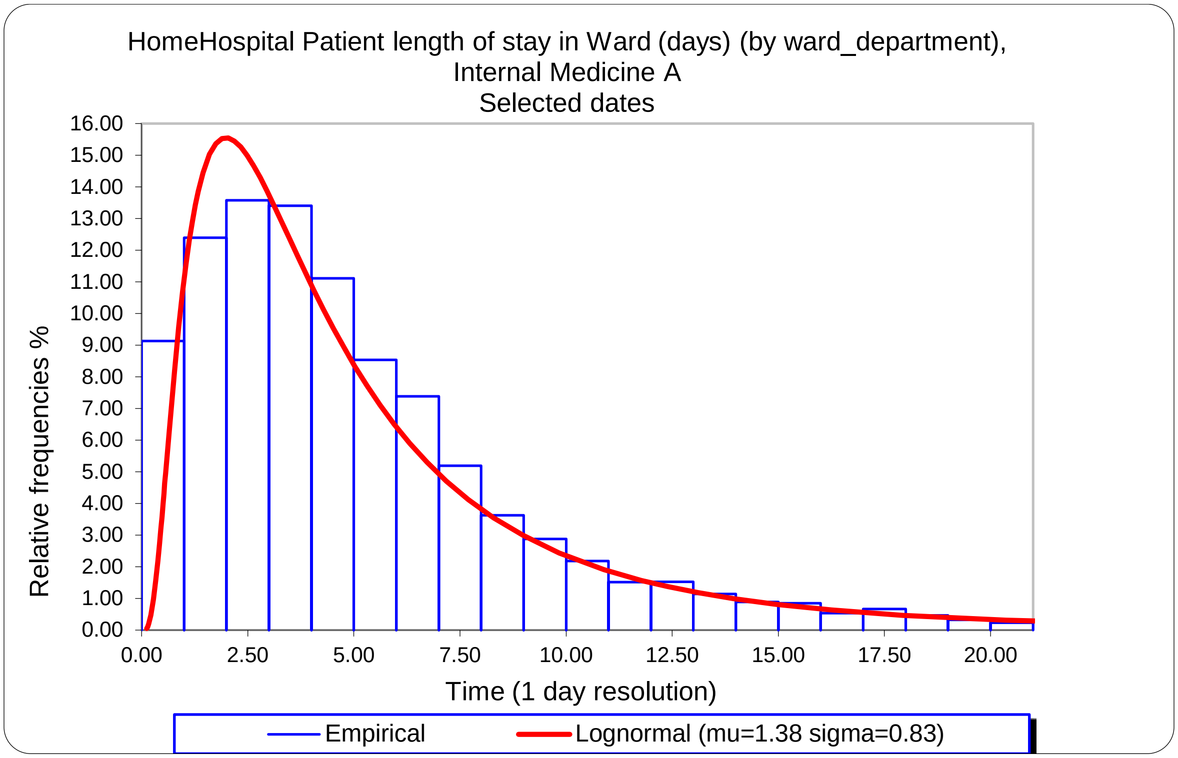
Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Fitting"**.
In the **“Variable”** tab, select **"Patient length of stay in Ward (days) (by ward_department)"**.

In the **"Options"** tab, under the **"Convert to"** select **"Relative frequencies".**

In the **"Select Categories"** tab, select **"Internal Medicine A"**.

In the **"X Properties"** tab, under **"Range to Display"** change the **Upper Quantile** limit to **97.5.** Click the **"Range to Compute”** button and choose "**Select Range"**, change the **Low Limit** to 1 and the **Upper Quantile** to 100.

Click the **"Dates->"** button.
Mark **“Dates totals only”**. Select months from **January 2004** to **October 2007** (using the shift key).
Open the **"Days"** tab and select **"All days"**.
Click **"OK"**.



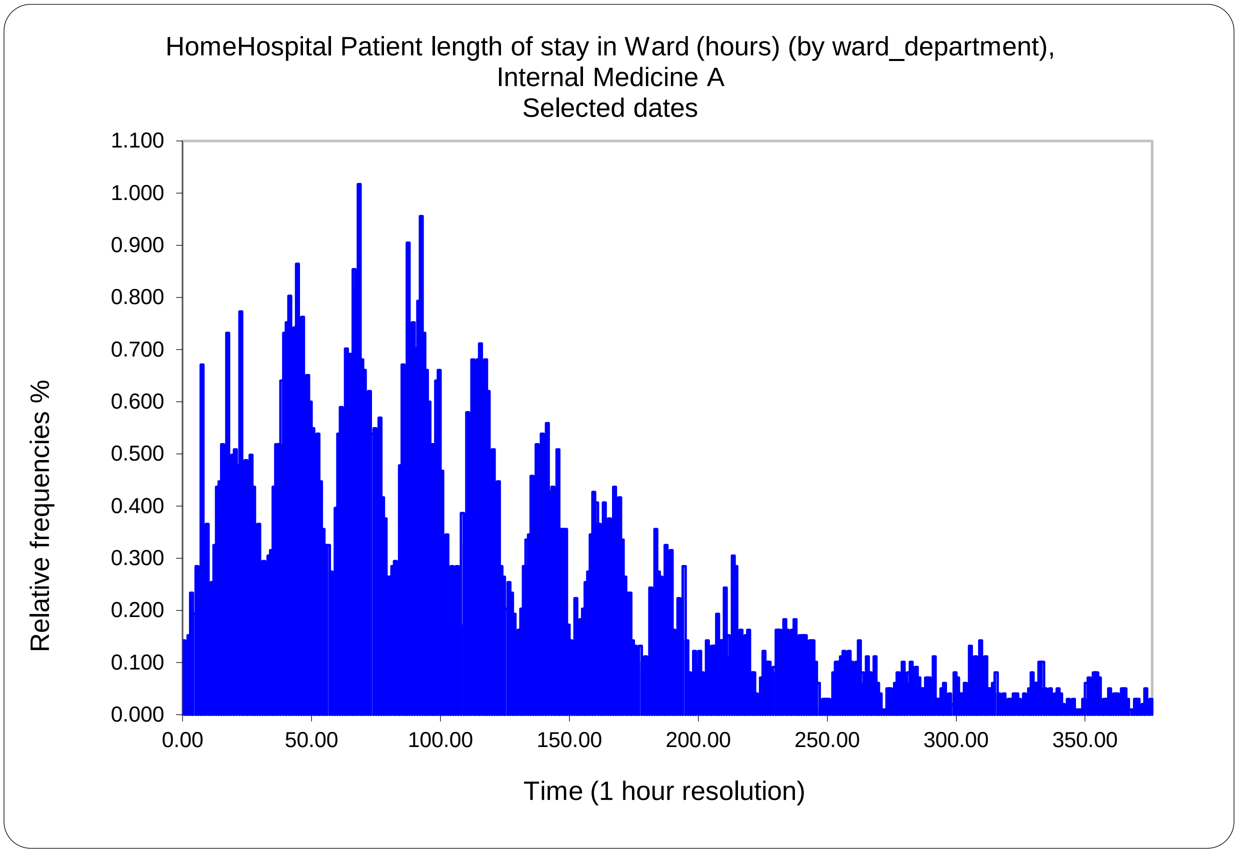
When considering daily resolutions, the Log-Normal distribution turns out to fit the LOS distribution well.

Example 6.2: ***LOS in Internal Medicine (in hours) - Distribution***

Return to the **"Statistical Models (Summaries)"** window.
Click the **"New Model"** button. Select **"Distributions"**, then **"Estimates"**.
In the **"Variable"** tab, select **" Patient length of stay in Ward (hours) (by ward_department)"**.
In the **"Select Categories"** tab, select **"Internal Medicine A"**.
In the **"X Properties"** tab change the **Upper Quantile** limit to **95.**

Click the **"Dates->"** button.
Mark **"Dates totals only"**. Select months from **January 2004** to **October 2007**.
Open the **"Days"** tab and select **"All days"**.

Click **"OK"**.



In the 1-hour resolution we observe a completely different LOS distribution, with peaks that are periodically 24 hours apart. The reason for this phenomena is the discharge procedure that is performed over "batches" of patients and, hence, takes a few hours. This results in a very low variance of the discharge time, as most patients are released between 3pm and 4pm (see Example 6.3).

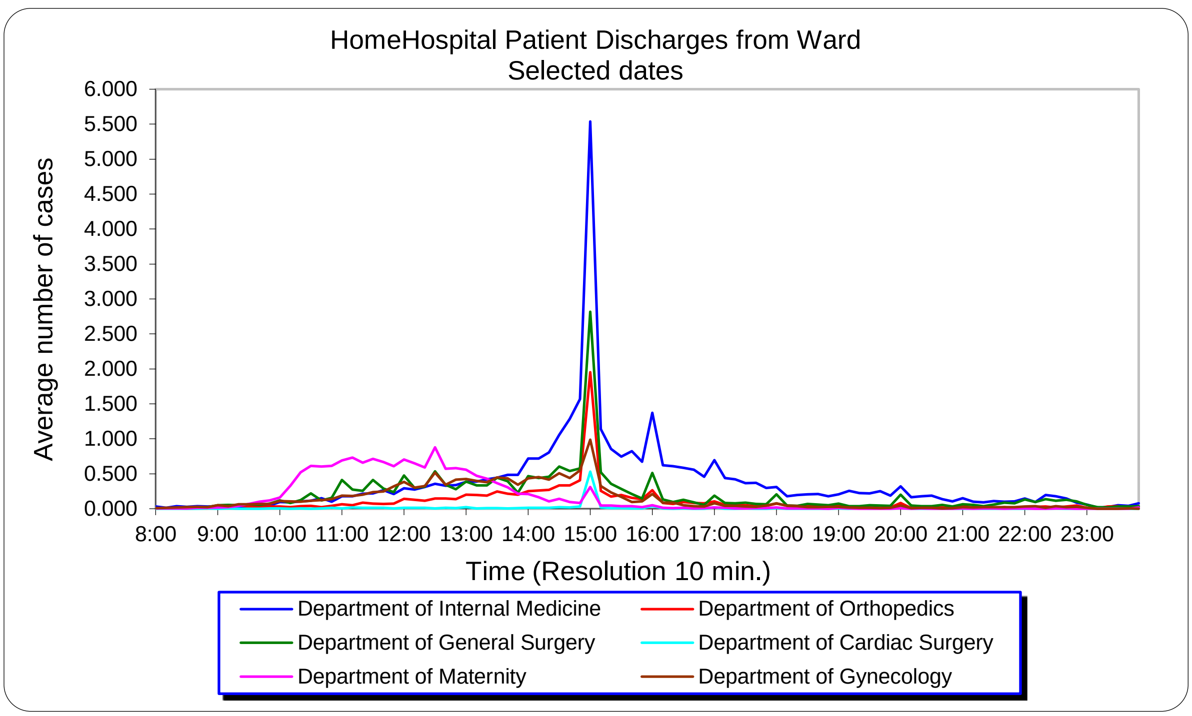
Example 6.3: ***Patient Discharges from Ward - Intraday time series***

Return to the **"Statistical Models (Summaries)"** window; click **"New Model"**, select **"Time Series"** and **"Intraday"**.

In the **"Variables"** tab, select **"Patient Discharges from Ward"**.

In the **"Select Categories"** tab, select
"Department of Internal Medicine",
"Department of Orthopedics",
"Department of General Surgery",
"Department of Cardiac Surgery",
"Department of Maternity",
"Department of Gynecology".

Open the **"X Properties"** tab and change the low limit to **08:00.**
Click the **"Dates->"** button. Mark **"Dates totals only"** and select months from **January 2004 to October 2007**. Open the **"Days"** tab and select **"Weekdays"**.
Click **"OK"**.



For additional reading on a data-based perspective on patient flow in hospitals, using the HomeHospital database, see the following paper: "On patient flow in hospitals: A data-based queueing-science perspective" by Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N. Marmor, Yulia Tseytlin, and Galit B. Yom-Tov. *Stochastic Systems*, 5, (2015), 146–194. http://ie.technion.ac.il/serveng/References/Short_Patient_flow_main.pdf
or the extended version on:
http://ie.technion.ac.il/serveng/References/Patient_flow_main_EV.pdf

Addendum

What you have experiences is only the beginning of what SEEStat can offer.

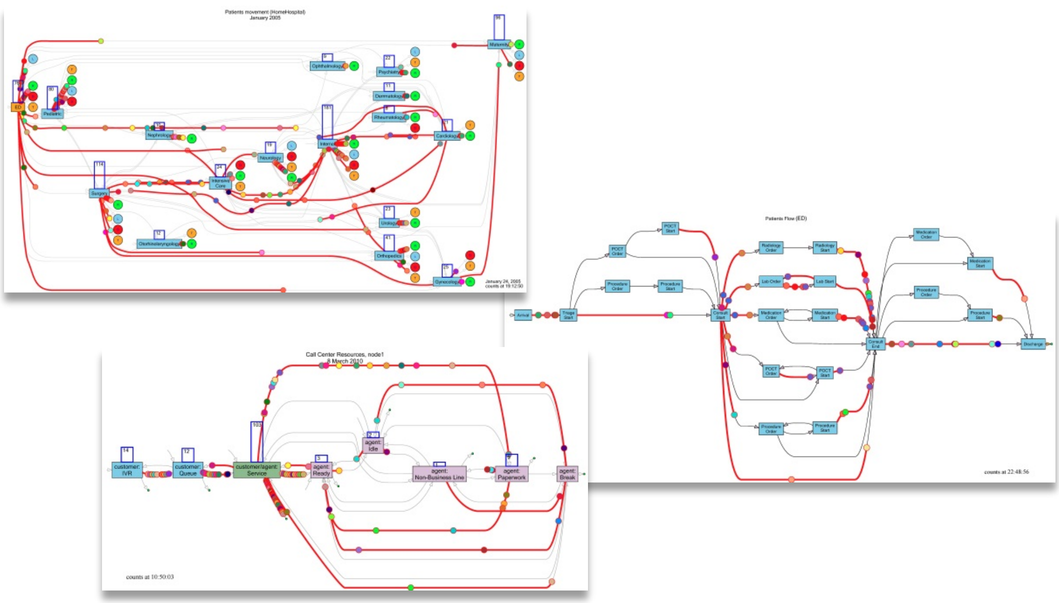
In fact, SEEStat also has “relatives” that could take you way further. For example, consider SEEGraph which creates semi-automatically data-animations. Examples are in

https://www.youtube.com/watch?v=1A6-jzS_scI

https://www.youtube.com/watch?v=jx3UUQCPODE

https://www.youtube.com/watch?v=-ik5kA7aLGg

where you can “observe” customers and agents in call centers, doctors and nurses in hospital, and more.



Appendix A: Mixture Fitting and Distribution Fitting

Mixture Fitting

The mixture distribution function:

$$\boxed{p(x) = \sum_{j=1}^{k} w_j \varphi(x;\theta_j); \qquad \sum_{j=1}^{k} w_j = 1; \qquad w_j > 0}$$

θ_j – distribution parameters (shape, scale or location parameters)

φ – given distribution function (Normal, Lognormal, Gamma, Exponential, Weibull or Invers Gaussian)

k – given number of components of the mixture

w_j – weighting coefficient or mixing proportions

Task: To find the optimal w_j and θ_j for sample X and given k and φ.

Solution:

- Specify initial values for w_j by using grid.
- For each point in a grid calculate mean, standard deviation (specify initial values for θ_j) and objective function (sum of squares of difference between empirical and fitted cumulative distribution functions).
- Select points with smallest values of objective function (best points).
- For each best point compute fitted distribution by using optimization algorithm.
- The best fit with smallest value of objective function is selected.

Distribution Fitting

Algorithms of distribution fitting (SEEStat software):

1. *Maximum likelihood estimates.*
   (11 distributions)
2. *L-moments estimates* (linear combination of order statistics).
   (31 distributions)
3. *Maximization of likelihood function.*
   (7 distributions)
4. *Optimization of the object function* (sum of squares of differences between empirical and fitted cumulative distribution functions)
   (1 distribution)

Appendix B: Smoothing References

Heft: Hazard estimation with flexible tails

[1] Charles Kooperberg, Charles J. Stone and Young K. Truong (1995). Hazard regression. *Journal of the American Statistical Association*, **90**, 78–94.
[2] Charles J. Stone, Mark Hansen, Charles Kooperberg, and Young K. Truong (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics*, **25**, 1371–1470.

Loess: Local polynomial regression fitting

[3] Cleveland, E. Grosse and W.M. Shyu (1992). Local regression models. Chapter 8 of *Statistical Models in S*, eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

Muhaz: Estimate hazard function from right-censored data

[4] H.G. Mueller and J.L. Wang (March 1994). Hazard rate estimation under random censoring with varying kernels and bandwidths, *Biometrics*, **50**, 61–76.
[5] O. Gefeller and H. Dette (1992). Nearest neighbour kernel estimation of the hazard function from censored data, *J. Statist. Comput. Simul.*, **43**, 93-101
[6] K.R. Hess, D.M. Serachitopol and B.W. Brown (1999). Hazard function estimators: A simulation study, *Statistics in Medicine*, **18** (22), 3075–3088.

Supsmu: Friedman's SuperSmoother

[7] J.H. Friedman (1984). SMART User's Guide. Laboratory for Computational Statistics, Stanford University Technical Report No. 1.
[8] J.H. Friedman (1984). A variable span scatterplot smoother. Laboratory for Computational Statistics, Stanford University Technical Report No. 5.

Pspline: Fit a polynomial smoothing spline of arbitrary order

[9] J.O. Ramsay, N. Heckman and B.W. Silverman (1997). Spline smoothing with model based penalties. *Behavior Research Methods, Instruments, & Computers*, **29** (1), 99–106.

Bspline: Fits a cubic smoothing spline to the supplied data

[10] J.M. Chambers and T.J. Hastie (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole.
[11] P.J. Green and B.W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Chapman and Hall.
[12] T.J. Hastie and R.J. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall.

dpill: Select a bandwidth for local linear regression

[13] D. Ruppert, S.J. Sheather and M.P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.
[14] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

dpih: Select a histogram bin width

[15] D.W. Scott (1979). On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

[16] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.

[17] M.P. Wand (1995). Data-based choice of histogram bin width. *University of New South Wales*, Australian Graduate School of Management Working Paper Series No. 95–011.

dpik: Select a bandwidth for kernel density estimation

[18] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.

[19] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

locpoly: Estimate functions using local polynomials

[20] M.P. Wand and M.C. Jones (1995). *Kernel Smoothing.* Chapman and Hall, London.

density: Kernel Density Estimation

[21] R.A. Becker, J.M. Chambers and A.R. Wilks (1988). *The New S Language*. Wadsworth & Brooks/Cole (for S version).

[22] D.W. Scott (1992) *Multivariate Density Estimation. Theory, Practice and Visualization*. New York: Wiley.

[23] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* **B**, 683–690.

[24] B.W. Silverman (1986). *Density Estimation*. London: Chapman and Hall.

[25] W.N. Venables and B.D. Ripley (2002). *Modern Applied Statistics with S*. New York: Springer.

bandwidth: Bandwidth selectors for kernel density estimation

[26] D.W. Scott (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley.

[27] S.J. Sheather and M.C. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society series B*, **53**, 683–690.

[28] B.W. Silverman (1986). *Density Estimation*. London: Chapman and Hall.

[29] W.N. Venables and B.D. Ripley (2002). *Modern Applied Statistics with S*. Springer.

[30] Adrian W.Bowman and Adelchi Azzalini (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Clarendon Press Oxford

[31] Jeffrey S.Simonoff (1996). *Smoothing Methods in Statistics*. Springer

[32] T.J. Hastie and R.J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.