

Homework No. 7

Statistical Analysis and Forecasting of Arrival

Note: Some questions are fully solved - these should not be handed in. For other questions, denoted by *, a partial solution is provided - the rest of the solution to these * questions should be submitted. As for the rest of the questions, fully solve them and hand in your solution.

Submit questions:

- Part 1: 1,2, and 3.
- Part 2: 1*,2*,3,4*,5*, and 6

Part 1. Basic Statistical Analysis

Description of data

The file **HW7_stat.xls** contains two worksheets.

- ❑ Worksheet “**main**” contains an information on arrival times of some particular working day, taken from the **Telephone Call-Center** Database (with all the irrelevant columns erased). This data is now used in Questions 1-3 below.
- ❑ Worksheet “**main_built**” is helpful for solving Question 3.

Questions

1. **(Submit)** Do you think that the following hypothesis is plausible: “The arrivals between 7:00 and 24:00 constitute a homogeneous Poisson process (with a *constant* arrival rate)”?
2. **(Submit)** Consider the interarrival data between 14:00 and 18:00. Assume that it is a sample from the exponential distribution. Construct a 95% confidence intervals for the average of interarrival times using two approaches:
 - confidence interval based on the exponential distribution;
 - normal approximation.

Compare the results of the two approaches. What is the better approach if you are not sure that interarrival times are exponential?

3. **(Submit)** Use the Test described in Recitation (with $L=0.12$) to check the hypothesis: “The arrivals between 7:00 and 24:00 constitute a *non-homogeneous* Poisson process?”

Part 2. Forecasting Arrival Rate at a Call Center.¹

Description of data

The file **HW7_forecast.xls** contains three worksheets:

- Worksheet "**call_counts**" contains the number of calls to the call center of an Israeli cellular phone company. The time period between 02/05/04 and 31/08/04 is studied. Saturdays and other three days with unusual pattern are excluded². For each day from the data set, number of calls per 33 half-hour intervals between 7:00 and 23:30 is considered. Each row contains intraday information for a specific day. Specifically, the first column contains date, and the second column contains day-of-week (1-Sunday, 2-Monday,...,6-Friday). The following 33 columns contain call volumes (number of calls) during the half-hour intervals. ("7:00" headline corresponds to 7:00-7:30 interval, "7:30" to 7:30-8:00 etc.)
- Worksheet "**forecast_day**" has the same structure as the worksheet "**call_counts**". It contains forecasts of an arrival rate for the days and the half-hour intervals in consideration. The time series model that was used for this purpose was presented at lectures. We assume that forecasting is performed **a day ahead**, i.e. number of calls for all previous days is known.
- Worksheet "**forecast_week**" contains forecasts of arrival rates, where each Thursday we perform forecast for the week that starts 10 days afterwards. (This procedure was used in the call center in consideration.) The same model as in the worksheet "**forecast_day**" is used.

Questions

1. **(Submit*)** Calculate overall daily call volumes for 02/05/04-31/08/04 and plot them (daily call volume versus date). Which conclusions can you derive from this graph?
2. **(Submit*)** Using 02/05/04-31/08/04 data, calculate average intraday arrival volumes during the half-hour intervals for Sundays, Thursdays and Fridays. Compare them in a graphical way. Which conclusions can you derive?

Below we shall study the quality of our forecasts for **the eight-week period between 20/06/04 and 13/08/04** (48 days overall). We shall use the following two measures of the forecast quality.

¹ We are deeply thankful to Sivan Aldor for her part in the preparation of this homework.

² The excluded days are 25-26/05/04(Shavuot holiday) and 22/08/2004 (unusually large arrival rate).

Root Mean Square Error (RMSE)

For each day j calculate $RMSE_j = \sqrt{\frac{1}{K} \sum_{k=1}^K (N_{jk} - F_{jk})^2}$, where $K=33$ is the number of intraday intervals, N_{jk} is the call volume in the time interval k on day j and F_{jk} is the forecast for this interval. Then calculate $RMSE = \frac{\sum_{j=1}^J RMSE_j}{J}$, where $J=48$ is the number of days.

Average Percent Error (APE)

For each day j calculate $APE = \frac{100}{K} \sum_{k=1}^K \frac{|N_{jk} - F_{jk}|}{N_{jk}}$. Then compute $APE = \frac{\sum_{j=1}^J APE_j}{J}$.

3. **(Submit)** Calculate RMSE and APE for the two forecasts summarized in the worksheets "forecast_day" and "forecast_week". Which forecast turns out to be more accurate? Why?

Let $N_j = \sum_{k=1}^K N_{jk}$ denote overall daily call volume on day j , and $F_j = \sum_{k=1}^K F_{jk}$ denote the forecast of daily call volume on day j .

Consider several simple forecasting methods that can be implied in EXCEL.

Method 1. Last observation. Let the forecast be equal to the last call volume of the same type, meaning the call volume during the same time interval of the same day-of-week on the previous week: $F_{jk} = N_{j-6,k}$.

Method 2. Moving average. Let the forecast be equal to the average of five last call volumes of the same type (previous 5 weeks).

Method 3. Using yesterday's call volume during the considered time interval. Implement forecasting formula:

$$F_{jk} = N_{j-1,k} + (F_{2,jk} - F_{2,j-1,k})$$

Where $F_{2,jk}$ and $F_{2,j-1,k}$ are forecasts by Method 2. We add the expression in brackets in order to incorporate difference between days-of-week in our forecast.

Method 4. Using yesterday's overall daily call volume and intraday arrival shape.

Use forecasting formula

$$F_{jk} = N_{j-1} \cdot p_{j-1,k} + (F_{2,jk} - F_{2,j-1,k})$$

Where N_{j-1} is the overall number of arrivals on day $j-1$ and $p_{j-1,k}$ is the fraction of interval j in the daily arrival rate. Estimate $p_{j-1,k}$ using moving average of last 5 intervals of the same type:

$$p_{j-1,k} = \frac{N_{j-1,k} + N_{j-7,k} + N_{j-13,k} + N_{j-19,k} + N_{j-25,k}}{N_{j-1} + N_{j-7} + N_{j-13} + N_{j-19} + N_{j-25}}$$

Method 5. Invent your own forecasting method and implement it. Provide formulae that you use.

4. **(Submit*)** Forecast the arrival volumes by each method described above (for the whole period). Consider the dates 01/08/2004 and 20/07/2004, compare in graphical way the arrival volumes in practice with the forecasted values (of the 5 methods and the time-series methods). What can you deduce from these results?
5. **(Submit*)** Calculate RMSE and APE for the five methods (for the eight-week period between 20/06/04 and 13/08/04). Compare the methods between them and compare them with the time-series methods considered in Question 3. Summarize your conclusions.
6. **(Submit)** Let the number of calls during the time interval k on day j be Poisson with parameter λ_{jk} . Assume that we succeeded to predict the values of parameters λ_{jk} **exactly** ($F_{jk} = \lambda_{jk}$). Would you expect RMSE and APE be equal to zero? Why?
7. **(Solved)** Consider the call volumes on 17 Tuesdays in May-August 2004, during 10:00-10:30 time interval. Is the following hypothesis plausible: "The call volumes in 10:00-10:30 interval on Tuesday are Poisson random variables with a common arrival rate". If not, can you think about an alternative model that will accommodate the phenomenon that you have discovered?

Hint. You are not expected to perform a complicated statistical analysis.

Note. The widely used mathematical technique for forecasting problems is *Time Series Analysis*. There is a dedicated course on Time Series, in which students are using the Time Series Modelling package ITSM2000. The software is included in a CD-ROM which accompanies the book "Introduction to Time Series and Forecasting" by Peter J. Brockwell and Richard A. Davis.