ISM 270 Homework 3 Hints

February 8, 2007

1 Question 1

"Fitzsimmons, page 347, Problem #11.7." This problem is straightforward; just implement Table 11.4 on page 331 using the equations given, and recalculate the MAD with $\alpha = 0.3$.

2 Question 2

"Fitzsimmons, page 347, Problem #11.10." Like Question 1, implement the spreadsheet using the equations given for α, β, and γ. Try to build your intuition about what the different terms are for. The equations at the bottom of page 339 are helpful, but watch out for the two errors they contain. Also, you must copy the equations up to row 19 as well as down from row 20.

3 Question 3

"Use the Excel solver to determine the best value of alpha to use in Q1 (problem 11.7)." Recall how we used the solver in Homework 1. Here, only two constraints are needed, one for the largest allowed value of the smoothing constant α, and one for the smallest α. Consider carefully: which cell should be the target cell, and which should be the cell that changes in order to influence the target cell? Do the problem for two ranges of α: once for the range .01 to 0.3, and once for 0.01 to 1. What MAD values do you get for each range?

The following quote is from Montgomery et al. [4], pages 102-104, regarding the choice of α: "The smoothing constant controls the extent to which past realizations of the time series influence the forecast. Small values of the smoothing constant give significant weight to many prior observations and result in a slow response of the forecasting system to changes in the parameters of the time series model. Larger values of the smoothing constant give weight only to the more recent historical data and cause the forecasting system to respond more rapidly to parameter shifts. However, a large smoothing constant may cause the system to respond to random variations in demand, when actually the model parameters have not changed. Such oversensitivity is not desirable..."

They note that if the results of a set of trials indicate a value for α that is larger than 0.3, the validity of the forecasting model needs to be examined. The data may be significantly autocorrelated, and a more detailed model which takes into account cyclical patterns or other unanticipated trends should be considered. For example, in Question 2 a more complex model was adopted to account for seasonal variations in demand.

4 Question 4

"Download the spreadsheet of call center worker data from the course web site, and observe the worksheet called 'day0108.' Apply Excel sorting functions to extract these fields: the calls of type NE handled by worker Avi. There should be fewer than twenty rows. For those rows, consider the cells under the column heading 'ser_time.' Sort these values in ascending order, and then do a chi-square goodness of fit test to them to see if an exponential distribution fits these values."

The purpose of this question is to review an important concept from statistics: the goodness of fit test of a chosen probability distribution to experimental data. Our textbook, page 399, the continuous time exponential distribution is given as

$$f(t) = \lambda e^{-\lambda t} \tag{1}$$

Here λ is the average arrival rate, with units that are the inverse of those of the time variable t. The mean of $f(t)$ is $\frac{1}{\lambda}$. Another way of writing $f(t)$, or of indicating an exponential random variable, is **Exp**(β), where $\beta = \frac{1}{\lambda}$. The cumulative distribution function for $f(t)$, for $t \geq 0$, is

$$F(t) = 1 - e^{-\lambda t} \tag{2}$$

F(t) gives the probability that the time between arrivals will be t or less. A discrete-time distribution that is closely related to $f(t)$ is the Poisson distribution, shown on page 400. The exponential and Poisson distributions are fundamental to the study of queueing theory and capacity management in Chapters 13 and 14. In this problem, we examine some real arrival time data of customers to a call center. So the underlying problem that motivates this question is: given this data, is the exponential distribution a good fit, and therefore does the Markovian arrival assumption of our queueing models apply?

Our thanks to Avi Mandelbaum and his group at Technion University in Israel for the bank call center data.

http://iew3.technion.ac.il/serveng/

4.1 Goodness of fit test for an exponential distribution

Pal, Jin, and Lim [5] is the primary reference for this section. Spiegel, Hsu, and Garcia all give more details on how to derive the chi-square distribution [2, 3, 7].

Suppose a random sample $X_1, X_2, ..., X_n$ is to be tested to determine if the underlying distribution could be an exponential random variable with parameter λ, or **Exp**(β). Create the null hypothesis

$$H_0 : X_i \text{ are iid } \mathbf{Exp}(\beta) \text{ random variables.} \tag{3}$$

The alternative hypothesis is

$$H_1 : X_i \text{ are NOT iid } \mathbf{Exp}(\beta) \text{ random variables.} \tag{4}$$

The following is a known result from the theory of statistics:

Pal, Theorem 6.2.4, page 160. Let $X_1, X_2, ..., X_n$ be iid nonnegative random variables with mean $\beta > 0$. Define $S = \sum_{i=1}^{n} X_i$ and $Z_j = \sum_{i=1}^{j} \frac{X_i}{S}$, with $i \leq j \leq n-1$. Then the random variables Z_j act like $(n-1)$ ordered random variables from a continuous uniform distribution over the interval $[0, 1]$ if and only if the random variables X_i are **Exp**(β).

Think about why this might be true, given the fact that the exponential distribution is the only continuous-time distribution with the property of being memoryless.

The term "ordered random variables" just means the samples are rearranged into ascending order. Take the samples $X_1, X_2, ..., X_n$, which could be for example $\{$11, 5, 93, 1, 100, 54$\}$; then if they are ordered, we have $\{$1, 5, 11, 54, 93, 100$\}$.

Now the hypothesis H_0 can be transformed into an easier operation:

$$H_0^* : \text{ the random variables } Z_i \text{ are samples of a continuous uniform distribution over [0,1].} \tag{5}$$

A sufficient statistic derived by Pearson [6] for testing the hypothesis H_0^* is

$$T^* = -2\sum_{i=1}^{n-1} ln(Z_i) \tag{6}$$

Look carefully at the definition of Z_i, and write out a few terms to be sure you understand what it means. Now, if T^* is too large or too small, H_0^* is rejected and the random sample is not well modelled by an exponential distribution. We say the probability distribution of the test statistic T^* under hypothesis H_0^* is $\chi^2_{2(n-1)}$, the chi squared distribution with 2(n-1) degrees of freedom. If T^* is too large, the points have shifted too far towards zero, and if T^* is too small, the points have shifted too far towards one in the uniform distribution over $[0,1]$.

So, we reject H_0^* (and H_0) at significance level ϕ if T^* is "significant": $T^* > \chi^2_{2(n-1),(\phi/2)}$ or $T^* < \chi^2_{2(n-1),1-(\phi/2)}$. We accept H_0^* (and H_0) otherwise.

To complete this procedure in Excel:

- Extract Avi's type NE job service times from the big spreadsheet using Excel sorting operations.
- Take that list of service times, and sort them into ascending order. Perhaps it's best to paste the list over to a fresh worksheet.
- The service times are the X_i values; compute the Z_i and $ln(Z_i)$ values, and then find T^*.
- Use a chi square distribution calculator, such as the following:

  `http://www.stat.sc.edu/~west/applets/chisqdemo.html`

  The number of degrees of freedom is $2 \cdot (n-1)$, and we are interested in the area right of T^*. After hitting Compute!, record the number you see, which is the level of significance.

  In testing a hypothesis, a *type one error* occurs if we reject the hypothesis when it happens to be true. The maximum probability with which we would be willing to risk a type one error is called the *level of significance* of the test.

  For this problem, choose a level of significance of 0.05. Should we accept hypothesis H_0^*?

For further reading:

`http://mathworld.wolfram.com/Chi-SquaredDistribution.html`

`http://en.wikipedia.org/wiki/Chi-square_distribution`

References

[1] J. Fitzsimmons and M. Fitzsimmons. *Service Management: Operations, Strategy, and Information Technology*, 5th edition. McGraw-Hill/Irwin, 2006.

[2] A. Leon-Garcia. *Probability and Random Processes for Electrical Engineering* , 2nd edition. Addison-Wesley, 1994.

[3] H. Hsu. *Probability, Random Variables, and Random Processes*. Schaum's Outline Series, McGraw Hill, 1997.

[4] D. Montgomery, L. Johnson, and J. Gardiner. *Forecasting and Time Series Analysis*, 2nd edition. McGraw-Hill, 1990.

[5] N. Pal, C. Jin, and W. Lim. *Handbook of Exponential and Related Distributions for Engineers and Scientists*, pages 167-170. Chapman and Hall/CRC, Boca Raton, FL, 2006.

[6] K. Pearson. “On a Method of Determining Whether a Sample of Size n Supposed to Have Been Drawn from a Parent Population Having a Known Probability Integral has Probably Been Drawn at Random.” Biometrika, Vol. 25, No. 3/4 (Dec., 1933), pp. 379-410.

[7] M. Spiegel, J. Schiller, and R. Srinivasan. *Probability and Statistics*, 2nd edition. Schaum’s Outline Series, 2000.