Department of Mathematics and Computer Science
University of Southern Denmark, Odense

April 21, 2006
Marco Chiarandini

DM86 Local Search Methods – Weekly Notes

Week 11, Spring 2006

Lecture April 20

This lecture, as well as the following one, is aimed at providing the basics of statistics which are useful for the analysis of LS algorithms. We stated why the analysis of LS algorithms needs Statistics and pointed out which are its main goals. We addressed the issue of determining a performance measure and introduced random variables and probability distributions. We reviewed few known theoretical probability distributions which are relevant for the study of algorithms or for the inferential statistical theory which we will discuss later.

Next we recognized two main branches in the subject: Descriptive Statistics and Inferential Statistics and focused the remaining part of the lecture on the former. We reviewed data summary measures such as location measures: average, quantiles, mode, and dispersion measures such as range variance and inter-quartile. Using the software R and data generated by three construction heuristics for graph coloring, we went over the graphical representations of data, such as histograms, barplots, smoothed probability density, empirical cumulative distribution functions, boxplots and qqplot. An introduction to R showing the methods used at the lecture can be found in the manual "Simple R" by John Verzani (link available from the course Section Literature). Finally, we mentioned about the difference between correlation and linear regression analysis. The use of these two kinds of analysis is left to be explored in R.

The material discussed at the lecture can be found in any manual of applied statistics such as, for example, the book by Zar "Biostatistical analysis" or by Petruccelli and Balgobin, "Applied statistics for engineers and scientists", available for consultation from the IMADA library. Chapter 4 of the text book is a suggested reading.

Exercises

The following exercises are to be carried out using R.

Exercise 1

Consider the distribution of solution quality produced by several runs of one construction heuristic (developed in the exercises of the first lectures of the course) on the TSP, GCP or SMTWTP on one single instance. Produce all the statistics introduced above. In particular, report the summary measures and the graphics describing the algorithm performance.

Exercise 2

Consider now for the algorithms of Exercise 1 both solution quality and computation time distributions. Generate hypothesis on the resemblance of these distributions with the theoretical ones introduced at the lecture (use `qqplot` or `qq.plot` from the `car` library). Motivate your hypothesis with a reasoning detached from the graphical results.

Exercise 3

Use the methods in Exercise 1 to compare two or more construction heuristics according to the solution quality criterion only. In particular, focus on empirical cumulative distribution functions (ECDFs). Enlarge then the analysis to more than one instance and produce a matrix of ECDF plots, each plot reporting the curves of the algorithms considered on one instance (use the `ecdf` function from the library `Hmisc`).

Exercise 4

Consider a metaheuristic algorithm developed in the past lectures for a problem at choice and run it several times on a single instance. Report the cumulative distribution function of the two random variables *solution quality* and *computation time*, i.e., the solution quality distribution at a fixed computation time, $F(q, T) = P[x \leq q, t = T]$ and the computation time distribution consequent to a fixed level of solution quality attainment, $F(Q, t) = P[q = Q, x \leq t]$.

On the data obtained by a single run, study the correlation between the variables computation time and solution time. Is there a dependency between them? How could you use all the data collected, i.e., the data from all the runs, in order to obtain a more robust analysis?