TECHNISCHE UNIVERSITÄT BERLIN
Fakultät IV – Elektrotechnik und Informatik
Fachgebiet Intelligente Netze und Management verteilter Systeme
Prof. Anja Feldmann, Ph.D.
Gregor Maier, Dr. Jörg Wallerich

1st Work Sheet Internet Measurement SS 08

Question 1: (100 points) *Basic Statistics* The goal of this assignment is to make you familiar with with handling measurement data and using basic statistics on the data. Please note, that some of the statistics will be explained in next week's lecture (e.g., confidence intervals). You can wait with solving these part until they have been explained in the lecture.

The measurement data is in CSV (comma separated values) format (space delimited). You can download them from the lecture's webpage (in section 'Homework'). These files can be processed using a variety of tools like R, gnuplot or similar tools. For some exercises you might want to use a program/tool to preprocess the data. When we present the solution for the assignment we will use R as our tool. Therefore we *suggest* that you also use R for the work sheets (but it is not required).

(a) The first data set is derived from a packet level trace. The file (`packet-size.dat`) is a one column file containing the size of each packet measured in bytes.

Your goal is to derive the following statistic values from the packet sizes:

1. Mean
2. Standard Deviation
3. 95%-Confidence Interval for Mean
4. Minimum, Maximum, 25%-, 50%- (= Median) and 75%-Quantiles[1]
5. Create a scatter plot (X-Y-plot)
6. Plot the Histogram

(b) The second data set is derived from the same packet level trace but contains the sum of bytes passing the link for time spans of 1 second (`bytes-1sec.dat`).

Your goal is to derive the same statistic values as listed in part (a).

(c) The third data set is derived from a proxy log (`proxy-log.dat`) containing for each transferred file the

- start time of the transfer in seconds since epoch (1.1.1970)
- duration of the transfer in seconds
- size of the transferred file in bytes

The data file thus consists one transfer per line with three space separated columns.

Your goal is to derive the same statistical values as in part (a) for the following data:

1. inter-request times (time from the start of one request to the start of the next one).
2. file sizes
3. transfer durations

Information on R

Introduction material:

- R Introduction slides: `http://www.net.t-labs.tu-berlin.de/teaching/ss08/IM_lecture/PDF/Introduction_to_R.pdf`[2]

[1] Look at the `summary` function in R
[2] Upon request we can give a small introduction to R outside of the normal lecture hours

- R Tutorial: `http://www.cyclismo.org/tutorial/R/`
- to read our one-column sample data into R use the following R command:
  `size <- scan(file="packet-size.dat")`
- to read a multi-colum file, use `proxy <- read.table(file="proxy-log.dat")`.
- You might the following R-functions useful: `plot, summary, mean, var, sd, density, ...`

Getting R

R is an open source program under the GPL. You can get it for Linux, Unix, Windows, MacOS, ... from the R project homepage (`http://www.r-project.org`).

Submission Details:

Due Date: 14-May-2008, 10:00 c.t. (just before the lecture) In paper form (i.e., print out the plots; print or write down values like mean). We might also offer "digital submission". If we do, you will find more information on the homepage.
Note, on 14-May we will also present a reference solution of the work sheet.