Probability and Statistics
Lab n^o 1

G. Urvoy-Keller - urvoy@unice.fr

November 21, 2012

1 Random Variables

1. Generate 3 random vectors of size 10000 from different distributions :
   - A uniform distribution between 0 and 1.
   - A normal distribution $\mathcal{N}(0, 10)$
   - A exponential distribution of parameter $\lambda = 2$

   (a) What is the number of bins to be used to represent the corresponding histograms according to Sturge's rule?

   (b) What is the bin size according to the Normal Reference rule?

   (c) What is the number of bins for each sample vector you have generated according to the Normal Reference Rule ?

   (d) Represent the histograms (R is using Sturge's rule with improvements, hence you can just use hist(X)) , cdfs and boxplots of each random vector.

2. Consider one random vector of size 1000 for each normal distribution $\mathcal{N}(0, v)$ for $v \in 1 \ldots 1000$ with steps of 50. For each random vector, compute the empirical variance and the empirical IQR and plot those pairs in a graph.

   The objective here was to highlight the fact that both the IQRs and the variances can be used to measure the variability of a random variable.

2 $E[1/X]$ vs. $1/E[X]$

We observed in course/recitation that the discrepancy between $E[1/X]$ vs. $1/E[X]$ is a function of the variance of X. This is what we are going to illustrate.

Let us consider the family of uniform distributions in the interval $[100 - v, 100 + v]$ for $v > 0$

1. What are the mean/variance of the family?
2. For each $v \in \{1, 2, \ldots 30\}$, draw a random vector of size 1000, compute its empirical variance $v[X]$ as well as $E[1/X]$ (simply mean(1/x) in R). Plot the pairs $(E[1/X] - 1/E[X], v[X])$ and comment.

3 Dependence vs. similar distribution

1. Draw a random variable X and a random variable Y (both of size 10000) from the same exponential distribution of parameter $\lambda = 2$. Plot the qqplot and the scatterplot of X and Y. The scatterplot is simply obtained by plot(X,Y). In the scatterplot, it might be useful to zoom in where the mass is. You can adjust the x-axis (resp. y-axis) between the 10-th and 90-th quantiles of X (resp. Y) with the command :

```
 plot(X,Y,xlim=cbind(quantile(X,0.1),quantile(X,0.9)),
ylim=cbind(quantile(Y,0.1),quantile(Y,0.9)))
```

   Comment the results

2. Let $Z = log(X) + 5$. Plot the qqplot and the scatterplot of X and Z. Comment the results

4 Loss events

The characterization of the loss process in the Internet is an highly debated topic in the research community (including ISPs).
In the present lab, we focus on losses observed by a sender machine S, involved in a BitTorrent session. In a BT session, a machine connects to and is contacted by other machines with whom it exchanges data over TCP connections.
The 2 files IP_address_loss.txt contain the time intervals between consecutive losses for the traffic sent by S to two different hosts.

4.1 Data Cleaning

We are interested in consecutive loss periods in a TCP stream of packets. Loss periods are not exactly loss events: if ever two packets sent back-to-back are lost, we count only a single period of loss. Thus, a loss period is a period where one or more consecutive losses occur. Due to the way BitTorrent and TCP are working, data needs to be cleaned in two ways: (i) removing too small values and (ii) too large values. The reason why we need to get rid of too large and too small values are the following:

- Too small values might reveal a high degree of dependence among the packets that were lost. To put it differently, losses that occur in the same time window should be discarded (counted only once, as we concentrate on loss periods and not on loss events).
- Too large values might/can be due to the way BT is working and not to the absence of losses. Indeed, BT can generate "long" idle periods of time during which hosts do not exchange data. This means that a BT transfer looks like an ON/OFF FTP process, see Figure 1.

1. Clean the traces by using the 10-th and 90-th quantiles (i.e.,keep only the values in between those two quantiles) of the distributions and show how the boxplots of data evolves before and after cleaning for the two TCP transfers.

4.2 Assessing the exponential hypothesis

Some researchers have observed during large measurement campaigns that inter-arrival times are exponentially distributed. We investigate here if it is a reasonable assumption for our two traces.

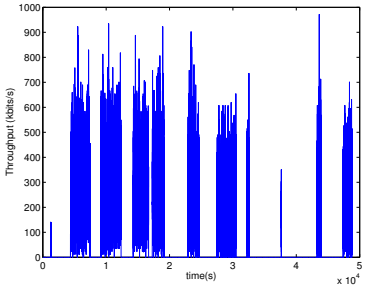

Figure 1: Example BT connection between two hosts

1. For each of the 2 connections (the cleaned versions obtained from the previous question), estimate the parameter of the exponential distribution that should model it.

2. For each of the 2 connections, generate a random vector following the exponential distribution of size 1000, represent the qqplot of each vector and the corresponding trace. Comment.

5 Central limit theorem

We are going to illustrate here the central limit theorem, which can formulated as follows (source wikipedia.org):

Let $X_1, ..., X_n$ be a random sample of size n – that is, a sequence of independent and identically distributed random variables drawn from distributions of expected values given by μ and finite variances given by σ^2. Suppose we are interested in the sample average

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

of these random variables. By the law of large numbers, the sample averages converge in probability and almost surely to the expected value μ as $n \to \infty$. The classical central limit theorem describes the size and the distributional form of the stochastic fluctuations around the deterministic number μ during this convergence. More precisely, it states that as n gets larger, the distribution of the difference between the sample average S_n and its limit μ, when multiplied by the factor $\sqrt{n}$ (that is $\sqrt{n}(S_n - \mu)$), approximates the normal distribution with mean 0 and variance σ^2. For large enough n, the distribution of S_n is close to the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The usefulness of the theorem is that the distribution of $\sqrt{n}(S_n - \mu))$ approaches normality regardless of the shape of the distribution of the individual X_i's.

Lindeberg-Lévy CLT: Suppose $\{Xi\}i \in \mathbb{N}$ is a sequence of i.i.d. random variables with $E[Xi] = \mu$ and $Var[Xi] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{(}n)(Sn - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}\Big(S_n - \mu\Big) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2).$$

where

$$S_n := (X_1 + \cdots + X_n)/n$$

To illustrate the theorem, let us generate 3 random vectors of size 1000 from different distributions :

- A uniform distribution between 0 and 1.
- A normal distribution $\mathcal{N}(0, 10)$
- A exponential distribution of parameter $\lambda = 2$

Questions:

1. Report in a table the empirical (resp. theoretical) mean and standard deviation for each random vector (resp. random variable).
2. Prove that we are in the conditions of the theorem for each vector.
3. Towards which distribution should $\sqrt(n)(Sn - \mu)$ should converge in each case.
4. Represent in a table with three columns (one for each original distribution) and two rows corresponding to:
   - the histogram of the original distributions
   - S_{10}
5. Report also the empirical mean and standard deviation for S_{10} for all cases.