Distributions and sampling

Markus Peuhkuri Esa Hyytiä Seyud Mortezaei
Tran Thien Thi

Introduction

This exercise will cover the main topics related to sampling, data aggregation technics, fitting different distributions and more. After completion of this exercise the students must have good understanding of how to sample their data set for simplicity in computations, how to estimate the true mean by sample mean, how to find a distribution of unknown dataset by fittings and more.

This work contains four tasks:

- Task 1: Sampling
- Task 2: Sampling and distributions
- Task 3: Distributions
- Task 4: High variability

All data can be found from `sampling-data.zip` archive from the assignment page and from `$RDATA/sampling-data.zip` (extracted to `$RDATA/sampling/` directory) at Aalto IT computers.

Prerequisites

This exercise requires students have prior knowledge about how to use R, statistics, probability distributions, Linux shell and how to leverage shell for writing basic shell script to automate the measurement process.

More information about R, statistics, probability and shell scripting can be found from the links of *Analyzing data using R or python* assignment or course web page.

Task 1: Sampling

In this task, you will practice random sampling and analyze its results.

In general, sampling serves some purposes, such as handling big amount of data or balancing amount of class instances for machine learning.

File `flowdata.txt` contains the following information for set of flows:

- Source IP (Anonymized)
- Destination IP (Anonymized)
- Protocol
- Is the port number valid
- Source port
- Destination port
- Number of packets
- Number of bytes
- Number of flows
- First packet arrival time
- Last packet arrival time

Complete the following tasks:

I. Select 1000 random sample data and produce a parallel plot to get an overview of the data.

II. Repeat the same procedure for flows with source port 80 (WWW).

III. Create a scatterplot based on number of bytes against packets and use logarithmic data if necessary. How are they related? What is the maximum average packet size?

IV. Study the average throughput of the connections. Average throughput is the number of bytes transferred divided by the transfer time. Clock resolution introduces some challenges, what can be said on the throughput of the flows that are transferred in zero time?

V. State your own observations on the data.

Tips: Useful R functions could be `sample()`, `ggparcoord()`, `plot()`.

Report, task 1

- Plots requested above with commands used to generate them
- Analysis of how bytes and packets are related.
- Throughput analysis.
- Conclusions

Task 2: Sampling and distributions

The goal of this exercise is to familiarize oneself with sampling and sampling distributions. The file `sampling.txt` contains certain session inter-arrival times. The goal is to study estimation of the mean inter-arrival time based on different sample sizes.

First you will generate histogram and compute mean with original data and after that produce histogram and mean using sampling specified.

1. Plot the histogram of the original data and compute the mean.

2. Select 5000 random samples from original data (i.e., you should have a vector of length 5000 values). Plot its histogram and compute the mean.

Following we generate 10000 random samples of different sizes (**n**) from the data. Select 10000 times n random elements from the data and compute mean of these n values. As a result, you should have an vector of 10000 values, each of them is mean value of n random elements. These values represent different results you could get for your statistic in a random sample and can be seen as samples from the sampling distribution of the sample mean statistic for n samples.

In addition to histogram of these 10000 values, study the values in a Q-Q plot against normal distribution and compute the mean and standard deviations of these 10000 values.

3. n=5
4. n=10
5. n=100

Tips: Useful R functions could be hist(), fitdistr(), rnorm(), qqplot(), mean(), and sd().

Report, task 2

- Histogram plots of each case
- Mean values of each case
- Q-Q plots, mean and standard deviations for cases 3-5.
- Discuss the effects of sample size to the sampling distribution and to the accuracy of the estimate.

Remember to add commands generated the plots and how statistics are computed.

Task 3: Distributions

This exercise addresses the modeling of measurement data with distributions.

There are several benefits to find suitable distribution to fit the data. For example, distributions will briefly describe the underlying data values and distributions could also utilized to generate new data to have larger dataset in certain cases. Furthermore, some learning algorithms assume some distribution to fit the data, which can help us understanding the low level details how the learning algorithms work.

The following three data sets are drawn from certain distributions presented at the lectures. The data sets are:

- `distr_a.txt`
- `distr_b.txt`
- `distr_c.txt`

Study each dataset to choise a good distribution for it. Estimate the parameters with software. Finally, validate your model by using appropriate plots and explain your modeling choices.

Report, task 3

For each dataset

- What distribution was choisen and why.
- Parameters of distribution.
- Validation with plots.
- Explain your choises.

Remember to document operations.

Task 4: High variability

This task attempts to demonstrate the effects of high variability in network measurements, by estimating means with on-line sampling. High variability can, for example, make them unpredictable in long term. File `flows.txt` contains once more values of flow lengths in packets and in bytes captured from a network.

1. Plot the data and compute its mean and median for both packets and bytes.
2. Let mean_n be the sample mean of the first n flow lengths in bytes. Derive an expression for running mean, i.e., write as a function of and the length of flow (y-axis has mean values and x-axis has amount of flows passed). This mimics a kind of an on-line measurement; we assume that the flows depart one-by-one and our estimate of the mean flow size in bytes is updated each time
3. Using the running mean, plot the mean estimate after each flow, i.e., plot the mean statistic for first observations as a function of n. Explain your observations.
4. Suppose that the interesting statistic is the median instead of the mean in an on-line scenario where a measurement system provides you with a large number of samples every second. How would you proceed in subtask 4.2 above?

Report, task 4

- Plots and values from point 1.
- Expression for running mean.
- Plot of mean estimate. Explain your observations.
- Computing median instead of mean. Derive expression.

Remember to document operations and reason your answers.