ELEC-E7130 Assignment 6. Distributions and sampling

Markus Peuhkuri Esa Hyytiä Seyud Mortezaei
Tran Thien Thi Weixuan Jiang
César Iván Olvera Espinosa

Prerequisites

1. To complete this assignment, students are required to have prior knowledge about how to use R or Python, statistics, and how to leverage the libraries available to process, analyze and plot data.

   If you lack the relevant skills, you may want to

   a. Take a look through related slides

   b. Refer to earlier assignments where you can learn some R and Python knowledge.

   c. Read the supporting materials.

Learning outcomes

At the end of this assignment, students should be able to

1. Understand the purpose of using distributions
2. Get to know about the different probability distributions available
3. Find the best distribution of unknown datasets by fittings and more
4. Validate the distribution through appropriate plots
5. Have a good understanding of how to sample their data set for simplicity in computations

Introduction

The present assignment covers the main topics related to probability distributions, fitting different distributions, validating the model of the distribution and the first steps to sampling to be aware of the utility and simplicity of computations. This assignment contains three tasks:

- Task 1: Introduction to distribution and sampling
- Task 2: Distributions
- Task 3: Sampling

Thereby, the students must understand how to find a distribution of unknown datasets by fittings and more as well as learn the importance, advantage, and disadvantage of using samples taken from the data set.

All data can be found from `sampling-data.zip` archive located in the directory `/work/courses/unix/T/ELEC/E7130/general/r-data` or, using the environment variable, `$RDATA/sampling-data.zip` (extracted into the directory `$RDATA/sampling/`) at Aalto IT computers.

> **Note:** You may type the command `kinit` before accessing to the directory to avoid issues related to the permissions.

```
$ source /work/courses/unix/T/ELEC/E7130/general/use.sh
$ cd $RDATA
$ ls
...
```

Task 1: Introduction to distribution and sampling

In the first task, answer the following questions:

1. What is **sampling** in statistics, and how does it help us understand data distributions?
2. Choose at least **three distributions** from the following options and explain their respective parameters and typical applications.
   - Normal Distribution
   - Beta Distribution
   - Exponential Distribution
   - Weibull Distribution
   - Log-gamma Distribution
   - Pareto Distribution
3. What are the components of the following **goodness-of-fit plots** used to validate a model?
   - Probability-Probability (P-P) plot
   - Quantile-Quantile (Q-Q) plot
   - Comparison of probability density (empirical and theoretical)
   - Comparison of cumulative distribution function (empirical and theoretical)

Report, task 1:

- Elaboration on the concepts of distribution and sampling.

Task 2: Distributions

The present task addresses the modeling of measurement data with distributions.

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

- **Download the three data sets** are drawn from certain distributions presented at the lectures, which are as follows:
  - `distr_a.txt`
  - `distr_b.txt`
  - `distr_c.txt`
- **Study each dataset to choose a good distribution for it.**
  - *Estimate the parameters* of distribution with software.
  - *Validate your model* by using appropriate plots
  - Explain your modeling choices and why.

  Tips: - Useful Python functions could be `distfit()`. - Useful R functions could be `fitdist()`.

Report, task 2

For each dataset:

- What distribution was chosen and why.
- Parameters of distribution.
- Validation with plots.
- Explain your choices.

Reminder: Document the process and operations.

Task 3: Sampling

This task provides an opportunity to practice random sampling, analyze the results, and understand the significance of sampling techniques in data analysis.

Download the file `flowdata.txt` which contains the following information for a set of flows as seen before:

- 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:

1. *Overview of the data set*
   - Select 1000 random sample data and produce a parallel plot to get an overview of the data.
2. *Number of bytes against packets*
   - Create a scatterplot (bytes vs packets) of the original data set and use logarithmic data if needed.
   - Create a scatterplot (bytes vs packets) of 1000 random sample data and use logarithmic data if needed
   - How are they related?
   - What is the maximum average packet size for both (original data set and 1000 random sample data)?
     Note: The average packet size of a flow is calculated with the number of bytes in a flow divided by its number of packets, that is, as the formula below:

$$Average\ packet\ size\ of\ a\ flow = \frac{total\ bytes\ of\ a\ flow}{total\ packets\ of\ a\ flow}$$

3. *Average throughput*
   - Calculate the average throughput of the connections. Clock resolution introduces some challenges, what can be said on the throughput of the flows that are transferred in zero time?
     Note: The average throughput of a flow is the number of bytes transferred divided by the transfer time, that is, the difference between the arrival time of the last packet and the first packet.
   - Study the average throughput of both the original dataset and the 1000 random samples data. State your own analysis of the data.

- **Draw conclusions about your own observations** on the data analyzed (original and random samples) and the usefulness of the graphs used.

  Tips: - Useful Python functions could be `pandas.sample()`, `pandas.plotting.parallel_coordinates()`, `matplotlib.pyplot.plot()`.
  - Useful R functions could be `sample()`, `ggparcoord()`, `plot()`.

Report, task 3

- Plots requested above with commands used to generate them

- Analysis of how bytes and packets are related.
- Throughput analysis.
- Conclusions

Grading standard

To pass this course, you need to achieve at least 15 points in this assignment. And if you submit the assignment late, you can get a maximum of 15 points.

You can get up to 30 points for this assignment:

Task 1

- Answer the questions appropriately. (9p)

Task 2

- Choose suitable distributions for the three data sets and explain why. (5p)
- Determine their parameters. (3p)
- Use graphs to verify your conjecture and explain them. (3p)

Task 3

- Plot the parallel plot as required. (2p)
- Draw the scatter plots as required and analyze. (4p)
- Calculate average throughput and analyze. (2p)
- Summary of observations on the data. (2p)

The quality of the report (bonus 2p)

The instruction of assignment

For the assignment, your submission must contain (Please don't contain original data in your submission):

- A zip file that includes your codes and scripts.
- A PDF file as your report.

Regarding the report, your report must have:

- A cover page indicating your name, student ID and your e-mail address.
- The report should include a description of measurements, a summary of the results and conclusions based on the results.
- An explanation of each problem, explain how you solved it and why you did it.