Laboratory Assignment 1

Statistics and Python Familiarization

This assignment covers the introductory statistics that you've learned. The exercises should be solved within the https://hub.cse.kau.se environment, using jupyter notebook.

When you have completed the exercises, you should upload the following on Canvas:

- one lab report, template available here
- the jupyter notebook document with all assignments solved, template available here.

For this assignment you need to download two data files that you will use in your calculations. You can find these files here:

- labdata.dat
- labdata2.dat

Please note that all laboratory assignments are individual, so each person should hand in their own individual solution.

Problem 1: Average values

You have performed measurements on a computer system using five different workloads, which in this case means five different benchmark programs. Each program reports a MFLOPS-16 value. The reported values are 654, 633, 684, 662, and 622 MFLOPS

(a) What is the average MFLOPS that can be calculated from these vaules?

You are evaluating the performance of a web server, and in particular its speed during the subsequent transfer of three images embedded in the web page. You study the transfer logs and find that the images are sent back-to-back on the same connection and that the time to transfer the three images are 13.32, 15.12, and 14.83 milliseconds. The images are 1.2 MebiByte each.

(b) What is the average throughput obtained during this back-to-back transfer expressed in bits per second?

You change a configuration setting in the web server, and consider another webpage which also has three images, but now the images are of different size namely 0.5, 1.6, and 2.4 MebiByte. The transfer times are 4.21, 20.33, and 29.83 millisecond. Again the images are transferred back to back.

(c) What is now the average throughput for the back-to-back transfer (again in bps)?

You do some further experiments and find that there are four important configuration options for the server that can be tweaked, and that they can be applied independently of each other. The options increase the image transfer rate by 18%, 27%, 12%, and 42% respectively.

(d) What is the average increase for a configuration tweak?

We are evaluating Starlink throughput performance by performing many repeated measurement runs. Each measurement run has a $8 - 15$ second duration, and report the average throughput during the run. Results for 6 runs spaced 10 minutes apart are 236, 157, 149, 201, 213, 189 Mbps.

(e) What is the estimated mean throughput during the measured hour?

Problem 2: Distributions

`ping` is a program that can be used to measure the round-trip time (RTT) between two hosts on a network. If two hosts are called **A** and **B**, then the RTT between them (from **A**'s point of view) is the time required for a packet to go: $\mathbf{A} \rightarrow \mathbf{B} \rightarrow \mathbf{A}$. The RTT is an important network metric as it can be used in calculation to estimate the average throughput for TCP connections, and so on. Below is an example of running `ping`, to measure the RTT between one of the university computers and `www.sunet.se`:

```
per [~/Courses/DVAD27/Lab-1] → ping -c 10 sunet.se
PING sunet.se (37.156.192.50): 56 data bytes
64 bytes from 37.156.192.50: icmp_seq=0 ttl=55 time=4.405 ms
64 bytes from 37.156.192.50: icmp_seq=1 ttl=55 time=4.399 ms
64 bytes from 37.156.192.50: icmp_seq=2 ttl=55 time=4.392 ms
64 bytes from 37.156.192.50: icmp_seq=3 ttl=55 time=4.442 ms
64 bytes from 37.156.192.50: icmp_seq=4 ttl=55 time=4.479 ms
64 bytes from 37.156.192.50: icmp_seq=5 ttl=55 time=4.433 ms
64 bytes from 37.156.192.50: icmp_seq=6 ttl=55 time=4.433 ms
64 bytes from 37.156.192.50: icmp_seq=7 ttl=55 time=4.365 ms
64 bytes from 37.156.192.50: icmp_seq=8 ttl=55 time=4.486 ms
64 bytes from 37.156.192.50: icmp_seq=9 ttl=55 time=4.476 ms

--- sunet.se ping statistics ---
10 packets transmitted, 10 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 4.365/4.431/4.486/0.039 ms

per [~/Courses/DVAD27/Lab-1] →
```

In the `ping` output above the RTT is the last field of each line, denoted time. In the file `labdata.dat`, which is available online, 1000 RTT samples has been gathered between two different hosts. The samples are located in the first column of the file, and it is you job to answer the following questions:

(a) Using an appropriate measure of central tendency, how much time is on average required for a `ping` probe to travel from **A** to **B**?

Depending on a number of different factors (e.g. network congestion), the RTT between two hosts can vary extensively. Using the gathered data, answer the following questions:

(b) Using standard deviation as an measure of dispersion, how much variation does the RTT samples display?

(c) Create a histogram, with appropriate axis labels, to visualize the distribution of RTT samples.

(d) How would you like to describe the shape of the distribution, and why do you think it has this shape?

Problem 3: Confidence intervals

To be able to do some more statistical analysis of network performance, we need to be able to calculate confidence intervals around the mean based on a given set of data samples. You should now implement your own python function that is able to compute confidence intervals. this. If so, a call to your function could look like:

```
c1, c2 = myConfInt(X, clevel)
```

where `X` is the sample set you want to calculate a confidence interval around the mean for, and `clevel` is the desired level of confidence $(1 - \alpha)$.
First of all, test your function on the RTT samples that you used in the previous exercises:

(a) Calculate $[c_1, c_2]$ for a confidence level of 90%.

(b) Calculate $[c_1, c_2]$ for a confidence level of 95%.

(c) Why is the interval larger for 95% than for a 90% interval? Should not a 95% interval be a *more precise* measure of the interval containing the "real" value?

Using this new function, you would like to investigate how the Linux TCP stack performs. When analyzing TCP behavior several methods can be used, let's say that we have used network emulation in this case. Two computers have been connected to a network emulator, which controls the bandwidth between the both computers, the delay, packet loss rate and so on. Let's say that the network emulator is configured so that the communication between the two computers is similar to communication between two relatively nearby computers on the Internet.

We would now like to evaluate the performance of TCP, so we design an experiment in which one of the computers (**A**) downloads a set of 33 files from the other computer (**B**). The time required to download the files is recorded in the first column of `labdata2.dat`.

(d) Between what values can we expect the mean of the transmission time to be? Choose a 95% level of confidence.

As we are gifted in TCP and kernel programming we optimize one aspect of the TCP stack, in the computer that transmits the data (**B**). The experiment is then repeated, so that the same files are downloaded in the same order. The results stored in the second column of `labdata2.dat`.

(e) Compare the achieved results to the unmodified TCP stack. What can you say about the difference in terms of confidence intervals?

(f) Do one set of calculations according to the paired measurements (aka Means of Difference) approach, and one set of calculations according to the Difference of Means approach, both using the same 95% confidence level. Do they show the same results? Why/Why not?

Problem 4: ANOVA

The throughput of three different cellular network providers were tested by transmitting a 10 Mbyte file five times, first in Network I, then in Network II and finally in Network III. The average completion times are provided in the table below. Using ANOVA, are there any statistically significant differences between the network providers, i.e., what can you say about the difference with 95% confidence? What can you say with 90% confidence? Calculate the results using numpy matrix operations first, and then if you wish you can verify your calculations using an alternative approach.

Compute the confidence intervals of the mean for each of the networks. What picture does the confidence intervals give regarding the difference between the networks, and how does that relate to the ANOVA results?

	Network I	Network II	Network III
Test 1	12.9	14.4	14.2
Test 2	13.9	13.2	14.7
Test 3	14.2	13.5	14.4
Test 4	13.8	14.5	14.9
Test 5	13.3	13.9	13.8