Single and multivariate statistics, and confidence intervals

Laboratory assignment #2 (deadline: Tue, Feb 24, 2026 at 11:59 pm)

Learning objectives

After completing this laboratory assignment students should be able to:

- Compute common statistics (including weighted mean and its standard deviation) for given univariate data sets in C++
- Compute variance-covariance and correlation coefficient matrices for a given multivariate data set in C++
- Extract critical values for normal distribution in MATLAB / Python given a desired percent probable error
- Generate clearly annotated observation, residual, histogram, and scatter plots in MATLAB / Python
- Analyse numerical and graphical output
- Write a technical report

Hints: students are advised to explore the help / MATLAB documentation for functions such as: `cdf()`, `icdf()`, `ones()`, `ceil()`, `floor()`, `fix()`, `sqrt()`, `histogram(…, 'Normalization', …, 'BinMethod', …)` /alternatively you can manually set `'NumBins'` and/or some of the other histogram properties/

Task 1

The same distance observable was measured multiple times by two different observers "A" and "B". The data sets acquired in the field by the two observers are provided to you in text files. Note that the measurement units are metres. Please include units in your results below.

In C/C++:

Perform the following tasks in C/C++:

- For each of the two random samples of observations compute the range, median, mean, variance, standard deviation, and standard deviation of the mean (show up to four decimal places)
- Using the mean as the best estimate, compute the residuals, their range, median, sum, mean, variance, and standard deviation (again, show up to four decimal places)
- Why are the range, variance and standard deviation values for the observations and their respective residuals the same? Why are the sum and the mean of the residuals zero?

- From the results for both observers, compute the weighted mean of the distance together with its standard deviation

In MATLAB / Python:

Perform the following tasks in MATLAB / Python:

- State the best estimate for each of the two random samples with its corresponding standard deviation at 95% confidence level
- For each data set plot a "time series" of the observations, its best estimate, and the weighted mean; also, make sure to show the 95% confidence interval for the best estimate
- On separate figure(s), plot the residual probability density histogram for the given samples; this time make sure to show the residual mean and its 99% confidence interval
- Do the plots exhibit any systematic issues? Why or why not?
- Are there any outlying residuals? Why or why not? Explain what you would do to mitigate such a problem.

Task 2

For the purpose of setting specifications when choosing new members on a football team, a random sample of ten forward players is to be analyzed. The weight [kg], height [m], speed [m/sec], and number of goals scored [g] in one season for each of the players is provided to you in text files. Again, please include units in your results below.

In C/C++:

Perform the following tasks in C/C++:

- Compute the mean, variance, and standard deviation for each of the single-variate samples $L_1 = [w_1 \quad w_2 \quad \dots \quad w_n]^T$, $L_2 = [h_1 \quad h_2 \quad \dots \quad h_n]^T$, $L_3 = [s_1 \quad s_2 \quad \dots \quad s_n]^T$, and $L_4 = [g_1 \quad g_2 \quad \dots \quad g_n]^T$ in the multi-variate data set $N = [L_1 \quad L_2 \quad L_3 \quad L_4]$
- Compute the variance-covariance matrix, C_N, for the multi-variate data set
- Compute the correlation coefficient matrix, R_N, for the multi-variate data set

In MATLAB / Python:

Perform the following tasks in MATLAB / Python:

- Produce a scatter plot for each pair of single-variate samples
- Given the correlation matrix and the scatter plots, discuss the degree of correlation between each pair of single-variate samples

Write-up and deliverables

The format and list of deliverables are explained in the Additional Course Information file uploaded on D2L.

Notes on source code

- This is an individual lab assignment and as such all results presented in the write-up must be obtained from your own source code.
- Computations must be performed in the C/C++ language using the Eigen library.
- Data visualization should be performed in MATLAB (or Python). Information related to statistical distributions should also be extracted via MATLAB (or Python) functionality.
- Source code will be evaluated on modularity, readability, use of comments, and style consistency. The use of functions is mandatory. Do not underestimate the value of well-organized and documented code in terms of long term usefulness.
- Data for your program(s) must come from external files. Hard coding of data should be avoided.