# Introduction to Probability and Statistics
# AIMS 2025/26 – Group Assignment

Submission deadline: Saturday 11$^{\text{th}}$ October, 6am.

## Instructions

- Each group has been assigned a different dataset. All datasets include:
  - At least one categorical variable that divides the data into two or more groups.
  - A few continuous variables.

  Details about your specific dataset are provided in the sheet `DataDescription` of your assigned spreadsheet.
- Your task for the presentation is to analyse the data. Below you will find some guidance on the type of questions you can address.
- The presentation should be divided into two parts:
  1. Exploratory Data Analysis (EDA): Descriptive Statistics of the dataset.
  2. Inference on appropriate questions of your choice.

  Make sure to link the two parts appropriately in your presentation.

## General guidance

Below are some general guidelines on the kinds of questions you might explore. Adapt or extend them according to your specific dataset.

### 1. Exploratory Data Analysis

- Use both numerical and graphical summaries to describe your data.
- When reporting numerical summaries, consider whether a table would make them clearer.
- If you create more than one plot for the same variable, relate the plots to each other and to the numerical results.
- Use the available arguments in your code to customise your plots (axis labels, font size, colours, legends, etc.). Plots should be clear and informative for the people who are listening to you!

**2. Inference**

- In your presentation, clearly state the questions you wish to answer, and accordingly justify the method(s) you have chosen.
- You may construct confidence intervals for some parameters, possibly comparing results across different groups of the categorical variable(s).
- Remember that confidence intervals can also be applied to proportions. Consider whether this applies to your dataset.
- You could formally test appropriate hypotheses in comparing different groups.
- If appropriate, explore relationships between numerical variables using the tools from Chapter 6 (correlation, linear regression, ...).
- Report which formulas you used and the results. Be prepared to comment on them, and on what each formula means.

The points above are only suggestions. The most meaningful questions will depend on your specific dataset, so feel free to explore beyond what I listed above.

However, don't overdo it. You only have 12 minutes to present. Focus on questions that are relevant to your dataset and that can be explained clearly within this available time.

## Additional Considerations

- During the EDA, you may consider whether any of the distributions we studied fit your data reasonably well. You can check this visually by overlaying appropriate theoretical density curves on your histograms.
- If a particular distribution appears suitable, go beyond graphical checks. For example, compare selected sample percentiles with the corresponding theoretical percentiles: do they match reasonably well?

**Final note:** Focus on clarity and interpretation rather than quantity. A concise, well-reasoned analysis is far more effective than an overcomplicated one.