Statistical Methods for the Physics Sciences: Assignment Grading Standards

Each assignment is broken into a number of tasks, worth the points mentioned at the beginning of each task. The tasks are designed to lead you through the statistical reasoning, and are not necessarily all the same in difficulty, in order to provide a range of challenges.

The points are then split between code and reproducibility (1 point), numerical results and figures (2 points) and prediction/interpretation (3-4 points). The smallest increment for any grade awarded to each of these aspects is 0.5 points.

Code and Reproducibility

1 point is given for producing code that executes I full and produces the correct outputs – less is given for a less complete or correct task, depending on whether a mostly correct but not optimal approach was demonstrated, or whether a correct approach was demonstrated but it was not successfully completed. Task completion can be determined objectively depending on what was asked for, and should apply the methods learned during the course. Here you will also be graded on reproducibility: make sure to include the cluster names, random seeds, etc. in your outputs.

For coding, we are applying the following grading guidelines. We are not focusing on code quality in this course so this will contribute at most a point to the points available for a task, but your code should demonstrate some important good practices for efficient and reproducible data analysis:

- ***Commenting and docstrings***: make sure you explain what your code is doing with brief commenting in the code cells (typically only once per block of a few lines to perform a task). Use docstrings to explain the purpose of any functions you write, their inputs and outputs (e.g. see https://philuttley.github.io/prog4aa/14-style/index.html).
- ***Code provenance***: include in comments/docstrings where your code came from if you did not write it yourself, e.g. from the course, from stack exchange (with link), from a colleague (state who), from chatGPT/Bing (then also give the prompt if it is small enough to include). Re-using code to do specific parts of the analysis is fine as long as you clearly state the source and it is clear you did most of the thinking into how to solve the problem, and did not just copy someone else (AI or human's) whole approach. Otherwise it will be considered as plagiarism.
- ***Use of functions and general efficiency***: to avoid repetitive and difficult-to-read code you should creating functions to do calculations or make plots that must be repeated multiple times. To speed up your code you should also avoid using loops in cases where it is simple to implement the same thing numpy/scipy arrays or Pandas dataframes, especially in cases where you would have to indent multiple levels of loops so that the computing time becomes prohibitive.

Numerical Results and Figures

2 points are given for the presentation of results: this includes both numerical results and figures You should demonstrate here key skills that go into the successful presentation of results so that others can understand: what you did and why, what your results are and be able to replicate successfully what you did, i.e. the essentials for presenting scientific research. This includes:

- ***Brief explanation*** (couple of sentences) of what you are doing and why (even if this is given in the question!), this should be in a markdown cell not the code cell.
- ***Figure quality***: The most important thing is that the figure is clear and can be easily interpreted, by you or someone else. There are a number of aspects to consider:
  - ***Axes*** must be clearly labelled (with units!) to show what is being plotted and with sufficiently large text to read easily (ideally when seen in a paper or on a slide in a presentation), also for the axis values as well as the axis label!
  - ***Markers and lines***: not too big/thick that they obscure each other too much, but not too small to see. If plotting multiple marker/line styles or colours, include a legend if possible to make it clear what is being plotted.

   - ***Scales***: use log or linear x- and y-scales as appropriate, e.g. data covers a factor 10 range or more on that axis: you should probably use log scale, also use log-log to highlight whether a relationship is power-law, log-linear if exponential.
   - ***Ranges***: the axis ranges plotted should be appropriate to the data values: make sure all the data are shown, but there should not be too much white space in your plot, e.g. if the data are all cluttered at one end your range is likely too big.
   - ***Visibility***: make sure that all the data shown on the plot is visible and if possible, that different components do not cover other components. E.g. use transparency (alpha parameter in matplotlib) and consider whether a histogram should be solid or stepped-line. Choose colours which are distinct but colour-blindness friendly (the default matplotlib palette is designed this way).
- **Clearly presented numerical results**: numerical results should be clearly identified (for example, for statistical tests that return multiple outputs, clearly state which is the p-value, which the test statistic) and labelled.

Predictions + Interpretation

A key part of statistical analysis is to state a hypothesis before performing the analysis, and interpreting the results. **Note that this is where a lot of points are lost by students, even when the answers given are fully correct!** This might seem subjective at first, but what we look for here first and foremost is a coherent statistical argument that correctly applies the concepts learned during the class.
Some questions ask you to make a prediction about what you will see in the data before you look at it: **there will be no points deducted for the prediction being incorrect**! In research, we make predictions that turn out to be wrong once you look at the data all the time! You can achieve full points with an incorrect prediction: you will be judged based on whether you applied valid statistical reasoning learned in class to motivate your prediction.

The specifics which you need to demonstrate depend on what the task asks of you (e.g. whether it only requires applying a statistical tests, produce plots, make predictions etc.).
The aspects we will consider here are:

- ***Assumptions*** being made if you are using some statistical test or specific approach to the analysis, especially the assumptions that must be made about the data for any statistical approach you are using to be valid. State the assumptions that the analysis you are using explicitly. E.g. should the data be normally distributed, or just the sample mean, should there be a certain minimum number of data points? The assumptions behind most statistical tests or approaches used are described in the course material (and for other methods, see online or in textbooks). Should be a few sentences at most.
- ***Validation of assumptions*** if needed, at most a few sentences or an extra figure. E.g. if the data themselves need to be normally distributed does this seem to be the case? This may need a plot to check, or at least an argument that the data 'look' normally distributed (which may or may not be correct). In many cases you are relying on the central limit theorem, i.e. the 'law of large numbers', are there enough data for this, e.g. given how skewed the data distribution is (see Episode 3 in the online materials, and lecture week1_l2)? Most of the time you just need to demonstrate awareness of these aspects before proceeding with your approach, but sometimes you may be able to demonstrate with an additional plot or test provided this does not take much effort compared to the task itself.
- ***Specificity of prediction and interpretation:*** your predictions/interpretations should correctly apply the statistical concepts learned in class, and they should agree with your own results obtained during the analysis. In almost all cases, we ask you to be specific with both, and to motivate your answer based on specific numerical results and features in Figures you've produced during the homework. Note that usually this involves figures/results produced in the same exercise, but using Figures/results from previous exercises can be equally valid. Important is that the results you use to motivate your argument are (1) relevant, and (2) the statistical reasoning is valid.
- ***Validity of the statistical argument:*** does the prediction/interpretation correctly apply the statistical concepts learned in class?