This readme file explains the difficult levels of the five lecture slide decks, as the post-workshop survey showed that "Due to the high school background of the students, and the workshop's limited time frame, some found the pace of the lectures to be overwhelming, particularly the statistical section of the lectures". To calibrate the learning expectation, we outline the main content of each slide deck.

stat-Sun-module1-intro.pdf (Introduction to data science and Covid-19 motivating example)

- Provides a brief introduction to data science, including some learning resources, including a link to Harvard Data Science Review (HDSR), a research and education journal.
- Uses Covid-19-related data science as motivating examples, including an illustration of the base rate fallacy.
- Outlines the learning goals of the workshop.
- 17 pages in total, and the overall content is suitable for all levels, including senior high school students.

stat-Sun-module2-data.pdf (Exploratory data analysis and the concept of regression)

- Discusses the main dataset used for the workshop analysis, including types of data.
- Shows some exploratory data analysis, including some basic descriptive statistics (e.g. mean and median) and graphs (e.g. histogram of the gene expression)
- Moves towards statistical inference and association analysis, including presenting some regression lines summarizing the relationship between gene expression data and SNP data.
- 25 pages in total, and the overall content is suitable for all levels, as the discussion focuses on intuition based on graphic representation. However, content on linear regression includes "arg min" type of math stat expressions, which are not suitable for a typical high school student.

stat-Sun-module3-pvalue.pdf (Introduction to p-values, using simulation of flipping a coin)

- Starts with a contextual question "A coin is tossed 100 times and 57 heads are observed. Is this a fair coin?", then discusses the key elements of conducting a test, including the null hypothesis, sample size, data, and parameter.
- Uses simulations to demonstrate p-values and again, focus on intuition based on graphic representation.
- To motivate students to further study, some formulations are given (e.g. on pages 10, 25-27), but these advanced materials are clearly marked by e.g. "Looking ahead" and "What's next".
- 31 pages in total, and the overall content is suitable for all levels, with the exception as noted in the bullet point above.

stat-Sun-module4-mht.pdf (Regression slope p-value and multiple hypothesis testing via simulation studies)

- Uses simulation to obtain p-value for the regression slope discussed in module 3.
- Provides intuition, graphs and step-by-step instructions on why and how to perform the Monte Carlo simulation.
- Towards the end of the slide deck (starting on page 29), briefly touch on multiple hypothesis testing
- 23 pages in total, and the overall content may be too advanced for senior high school students, while more suitable for junior undergraduate students majoring in statistics or graduate students from other disciplines but with some background training in data science.

stat-Sun-module5-misc.pdf (misc topics)

- Discuss two misc topics: principal component analysis and overfitting.
- 28 pages in total, and the overall content is even more advanced than that of module 4, i.e. too advanced for senior high school students.