Assignment 2

Statistics exercises

Please report, for all tests you perform: which test you used (and why), the relevant statistics (t-values, F-values, z-values, correlation coefficients, intercept and slope etc.) and the significance level. Also include all relevant plots (with a brief explanation). For completeness' sake, specify which data file you used.
If you make plots, pay attention to details (axes labels, readability etc.).

If you use R, you may also specify the R-command you used (but this is *only* supplementary information to allow me to check what you did exactly).

Go to http://ai.vub.ac.be/~bart/statsnumbers.html and type in your student number. This will be converted to four numbers between 1 and 4
If you receive the numbers: 1, 3, 2, 2, use :
Question1_1.csv, Question2_3.csv, Question3_2.csv, Question4_2.csv
The data files themselves are all found in StatisticsAssignmentData.zip on the Canvas site.

The deadline is December 22, midnight.

Question 1
The data file contains the results from an experiment that measures the resonance frequencies of two types `x1` and `x2` of (hopefully) accurate crystal oscillators needed to provide timing information for high performance hardware.
The resonance frequencies for the first type of crystal are given in column `frequency`, the corresponding type of crystal used in column `crystal`.

- Make an *appropriate* plot to show how the data are distributed. Explain why you used this type of plot (2+2pts)
- Test whether the resonance frequencies of the two types could be identical. Also test whether they are different from the target frequency of 536 870 912 Hz. Explain why you used this test. (2+2 pts.)
- Would you trust these crystals in your application, and why (not)? (2 pts)

Question 2
The data file(s) contain the number of times visitors to two variants of a website (`site1` and `site2`) bought something or left without buying (as registered by the web server). The files whose name ends in `a` contain the results after 200 visits to each variant.

- Do an appropriate test to see whether these data indicate a difference between the two sites. Explain why you used this test and give the effect size. (2+1+2 pts)
- The files whose name ends in b contain data from a longer run (5000 visits) of the experiment.
- Do an appropriate test to see whether these data indicate a difference between the two sites. Explain why you used this test and give the effect size. (1+1+1 pts)
- What did this teach you about the importance of sample size (a.k.a. "big data")? (2 pts)

Question 3
The data file contains the grades of two groups of students, one of which followed remedial lectures (`experiment`) and one which did not (control).

- Investigate whether the remedial lectures had any effect, and explain what you did and why you did it. 2+2 pts
- What was the power of this experiment? What does this mean, and are you happy with this (and why)? (2+1 pts)
- How large would the groups need to be to achieve power 0.8? Explain. (2+1 pts)

Question 4
The data file contains data on whether two algorithms for deep learning model successfully solved a given task, and the number of parameters in each variant of the algorithm. It is clear that increasing the number of parameters increases the chance of success.

- Can you figure out whether one of the algorithms is better, controlling for the number of parameters? Explain how you determined this (3+3 pts).
- At which number of parameters does each of the algorithms have a 50% chance of succeeding? Explain how you calculated this. (2+2 pts).