General Feedback Assignment 1

For question 1, most students failed to understand for what was the language python designed for and whether it is a compiled or interpreted language. A lot of students answered that python is designed to be a better language than the ABC language while working with the Amoeba system (some even answered that Guido van Rossum designed python because he was bored during Christmas!) which completely missed the point of the question. The correct answer is that python is designed as a scripting or general-purpose language. Some students also seem to be confused regarding the compiled/interpreted issue, which is understandable due to the ambiguity of the execution process. Strictly speaking, a compilation means a translation of one language to another language, in this case, python has a compilation process from the source code to a bytecode, however, in contrast with other compiled languages, this bytecode is not run directly by the CPU but instead is to be interpreted by a virtual machine. Another contrast is that the compilation process of python is done implicitly and you practically can instruct the python interpreter to avoid writing a .pyc file if you want it by setting a PYTHONDONTWRITEBYTECODE environment variable. In other compiled languages, like Java, the source code has to be compiled explicitly into a bytecode before you can run it. The official statement from the python website (https://www.python.org/doc/essays/blurb/) is that python is an interpreted language.

For question 4, we expected students to focus on the bias and sampling size issues. The estimate is wrong because the sample is too small and it is also biased (most of the data are films with high death rate). Answers that are too subjective to personal experience without mentioning those two issues get a mark deduction. Usually, these two issues are the most common causes of wrong estimations. Please pay attention to these two issues when you are doing work with your research or real data.

For question 5, some students forgot to handle years that have missing data (and got a mark deduction): few students handled the case for which the number of films in a year was zero and prevented dividing by zero which gives NaN. We also noticed that most students do not understand what “a good estimate” means. A good estimate means that the prediction is as close as possible to the actual value (which is hidden from us), but we can draw a basic assumption by looking closely at the data itself. The data given is sparse for some years, having only one data point per year. As such, it is not possible to give a good estimate using such unique data point.

For question 6, the most common mistake was using the wrong equation. Some of you didn't sum up the values over years but just stored each year's value in an array and just printed these values out.

You could have actually verified your answers by doing simple counts as demonstrated in the previous cell in the same assignment notebook:

```
deaths = (film_deaths.Body_Count>40).sum()  # number of
positive outcomes (in sum True counts as 1, False counts as 0)
total_films = film_deaths.Body_Count.count()
prob_death = deaths/total_films
```

For question 7, the most common mistake was that students didn't carefully assess what other information was given from the previous cell prior to the question. I.e. we saw that some students missed that they can actually use other information such as film's length, genre and so on. Several answers were quite good, some made a sound reasoning and explained how to use conditional distributions and Bayesian inference using other different features.

--

This feedback was written by Hardy Hardy, Nada YEM AbdelRahman, Fariba Yousefi and Mauricio A Alvarez.