6 Mostly probabilistic models

Depending on your background, some of these exercises may be easy, while others may be more difficult. We simply ask you to do your best with these problems, within the limits of the time allocated for the course.

(keeping track)

(DATA CALIBRATION)

a) Say that you have data from an opinion poll about how people vote. In your poll you happen to have a 53/47 distribution of men and women. However, in the population as a whole you know that the ratio is 50/50. Is it possible to improve the poll result, by taking into account this additional knowledge? If so, how would you suggest to do it?

b) (voluntary and more difficult) Suppose that you have two attributes for each person (male/female, and city/country) and that you know these ratios in the population as a whole, e.g. 50/50 and 70/30. How could you do then?

(investigating the abstract)

(DICE SIMULATION)

Have a look at the program **dice.py**, and study it, including the pseudo-random generator. (you may read about **about pseudo-number generation**)

Simulate the sum of two dice (Monte Carlo simulation = simulation based on random generation). Explain why some values appear to be more probable than others.

a) Simulate for more than two dice, i.e. as a function of k. For the sum, what can you observe about its expected value and its variance? What can you observe about the mean? At least, give a qualitative answer, and a more quantitative if you can (eg. exactly what is the expected value of the sum?). (I here assume that you have already refreshed the notion of mean and variance from your other course)

b) (voluntary) As a second step ask yourself if you can prove if any of your observations/hypotheses always hold. (even in the situation when you read mostly in a book, it

is really essential for your understanding of a topic to think about how you could in principle find things by investigating, and then attempt to prove them)

c) As for the shape of the curve apparent already for k=3, it is quite possible to discover a formula with the techniques of trying out that you have already practiced in the course. Even if you already know the answer from other courses, try to imagine and propose how the formula could be discovered with simple observations and testing. If you don't know the answer make an initial attempt to guess the formula and/or suggest some ideas, but don't spend a lot of time here unless you really want to. I ask this mostly to make you aware that by trying things out, you would probably after a while find the formula.

(investigating the world)

(STOCHASTIC TRAFFIC SIMULATION)

A classical application of statistical models is simulation of systems where things happen randomly based on certain probability distributions that are chosen to be as realistic as possible. This is called Monte Carlo simulation.

See for example these simple traffic simulation demos:

demo1, **demo2**, **demo3**, **wikipedia** (there are lots out there).

I intentionally chose some simple demos, so that you can more easily relate to actually creating them yourself. As in many other areas these days, there is a lot of sophisticated software with advanced models and beautiful graphics, where you almost forget that the same basic mathematical techniques are used.

All I ask you is to take a look so that you have seen it. Comments still welcome!

(RADIOACTIVE DECAY)

Radioactivity can be modelled by assuming that each atom (of a certain radioactive kind) has a given probability per time unit to emit a particle. When the particle has been emitted, the atom is no longer radioactive. Motivate what kind of function you can expect the radioactivity to follow over time.

(MEDICAL TEST)

A public screening is done of a group of people to find the persons who have the disease X. This is done with a medical test. As with most medical tests, the test is not 100% reliable. It gives a correct result with a probability of 99% if the person has the disease, and with 97% if the person does not have the disease. Prior to the screening, it has been

estimated that about 0.33% of the individuals in the group have the disease.

For a particular person the test has indicated a positive result. What is the probability that the person actually has the disease? (If you cannot solve the problem, at least try to explain why the answer is not simply 0.99 or 0.97!)

Hint: Begin by writing down in mathematical notation what you know from the start. Try also to think what would happen with very extreme or symmetric numbers to investigate and understand the problem (this is a good generally useful technique).

(KIDNEY STONE TREATMENT)

In a study, two treatments for kidney stones were considered (this is real data). Carefully study the table with the success rates and make any observations.

	Treatment A	Treatment B
Small stones	*Group 1* 93% (81/87)	*Group 2* 87% (234/270)
Large stones	*Group 3* 73% (192/263)	*Group 4* 69% (55/80)
Both together	78% (273/350)	83% (289/350)

(designing)

(LANGUAGE RECOGNITION)

Look at the simple language recognition program in **language.py**. It uses the probabilities of single letters for two languages (from reference texts **lang1.txt** and **lang2.txt**), to calculate if a string is likely to belong to the first or the second language. Investigate and explain!

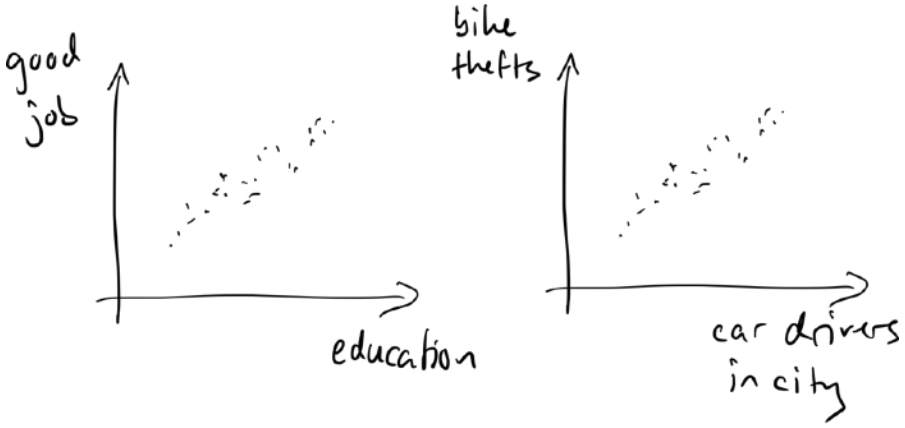
(New better program below!!!)

(voluntary) Given the interest in Markov chains in the introductory lecture, I decided to extend the first program so that it also calculates the probabilities according to a Markov chain between adjacent letters - this is a powerful way to model many phenomena. I chose to do it in the simplest way where the probability of the next letter depends on one previous letter. The new program is in **language2.py**. I also added a loop so that you can easily play around with different input strings.

(thinking)

(CORRELATION AND CAUSALITY)

Study the two graphs. What are your observations?



(WEATHER PREDICTION)

Consider the issue of predicting the probability of precipitation on a given future day. To your help you have weather statistics from the last five years. Now suppose you want to predict if there will be any precipitation on May 19. Should you base your prediction on a) statistics for May 19 during these years, b) statistics for all days in May during these years c) statistics for all days during these years? Before you consider anything else, assume that the probability is simply estimated as the relative frequency of precipitation for all days you choose to include.

Motivate your answer, and discuss the difficulties involved in choosing the model. Would the situation be any different if you had 100 years of weather statistics? Is choosing the model something that necessarily requires human judgement? Hint: while the specific question is not so difficult to intuitively answer, the general issues behind the question are deep. So think!

(I placed the problem in this section, since we are not really asking how to best predict the weather, but rather asking a fundamental question about modelling and prediction)

(NATURE OF RANDOMNESS AND PROBABILITY)

a) What kind of predictions can you draw from stochastic models, compared to when you have a deterministic model (like for example an astronomical model of planetary motion).
b) Does randomness exist? If so, what is it?
c) Is there a “right” probability e.g. for rain tomorrow? Or even for a die?

(mathematical knowledge)

(JOINT PROBABILITIES, MARGINAL DISTRIBUTIONS, CONDITIONAL PROBABILITES)

Look at my lecture notes about joint probability tables, marginal and conditional probabilities. This should mostly

be known already from the other course (for those of you in the ADS program), but I think it is good to see other explanations.

(LAW OF LARGE NUMBERS)

Law of large numbers on Wikipedia

(finally...)

(SELF-CHECK)

- Have you answered all questions to the best of your ability?
- Is the required information on the front page, file name correct etc.?
- Anything else you can easily check?

If you pass the self-check, simply write "Self-check passed!". Otherwise, fix your submission before you submit - do not submit an incomplete module! You can receive personal help and/or a short extension if you contact a supervisor.

Remember to confirm your successful self-check!