6 Mostly probabilistic models

Group 1
Author: Min Wu(8603148365, DIT025)
Email:minwuh081@gmail.com

I hereby declare that all solutions are entirely my own work, without having taken part of other solutions.

The number of hours spent: 20hours (Min Wu)
The number of hours has been present in supervision for this module: 0h

(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?

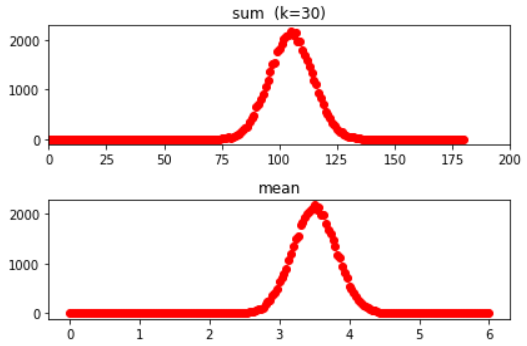
No, I don't think it will a way to improve the poll result by taking into account this information. 53/47 distribution is very close to 50/50. In the problem, there isn't mention how many people in vote either. The variance may exist. The data need to be normalized for an unbalanced dataset.

(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.

Understanding the program dice.py:

1. 30 dice roll 50000 times
2. calculate the sum of those appearing values
3. plot the distribution of each sum

thirty dices



The number of times vs of the sum of 30 dices(up panel), The number of times vs the average value of 30 dices (down panel). The highest probability is located at 105 which appears 2182 times and its mean value is 3.5.

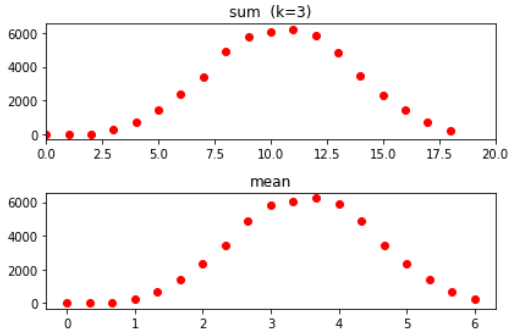
- the minimal number of one dice is 1, so the minimal sum of 30 dices is 30.
- the maximal number of one dice is 6, so the maximal sum of 30 dices is 180.
- The mean value is in the range of 1-6.

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)

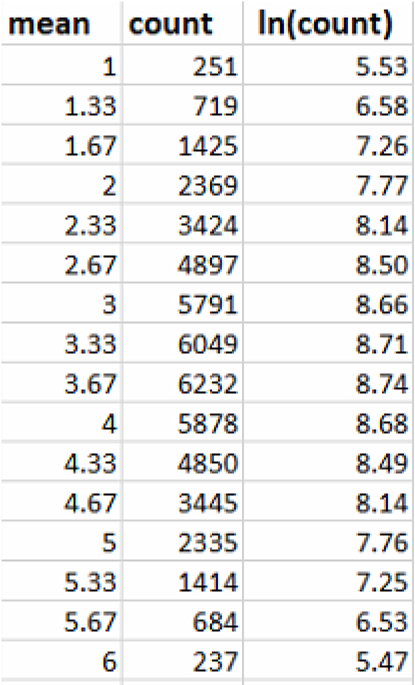
Assume k dices
The expected value with the highest counts is located at (6*k+k)/2
The variance is ∓2.5k

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.

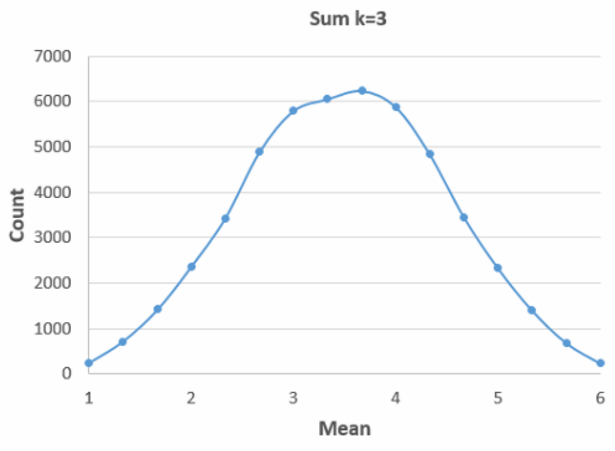
k= 3



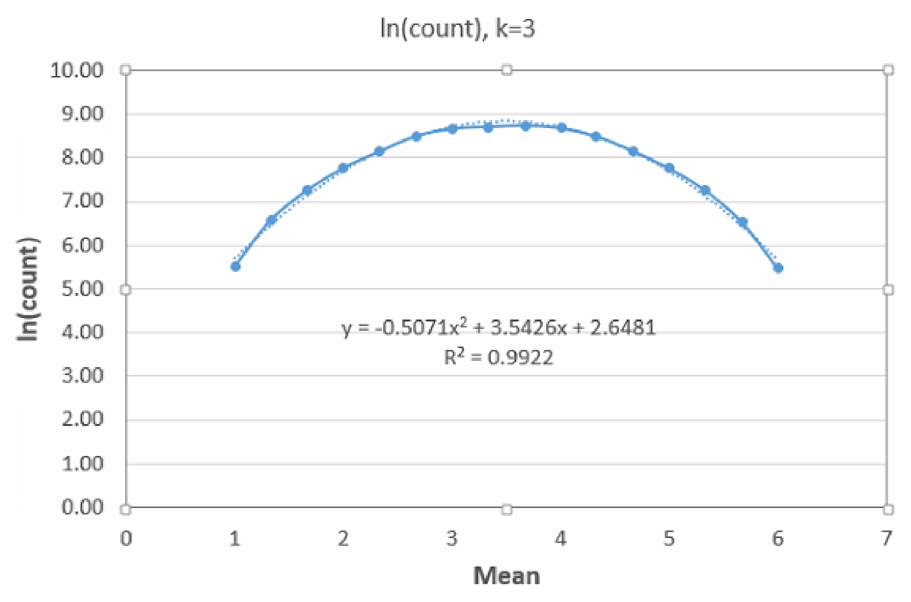
The minimal sum of three dices is 3 and the maximal sum of three dices is 18. The mean value is in the range of 1-6.



mean	count	ln(count)
1	251	5.53
1.33	719	6.58
1.67	1425	7.26
2	2369	7.77
2.33	3424	8.14
2.67	4897	8.50
3	5791	8.66
3.33	6049	8.71
3.67	6232	8.74
4	5878	8.68
4.33	4850	8.49
4.67	3445	8.14
5	2335	7.76
5.33	1414	7.25
5.67	684	6.53
6	237	5.47



Take the ln(count), so the plot becomes:



The fitted function is -0.5071x^2+3.5426x+2.6481 with R^2 equals 0.9922.

It seems quite good.

The final formula is :

y=exp(-0.51x^2+3.54x+2.65)

(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).

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

Those simulations are based on random number generation. In real life, is the decision of the driver who drives a car from one place to the other place made randomly?

(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.

I think each atom should randomly emit a particle in a certain period of time and in the end. After the particle has been emitted, the atom is no longer radioactive. After all the atoms become no longer radioactive, the radioactivity will be zero follow over time. At beginning all the atoms have radioactivity, the probability to emit a particle is highest, with the time increase, more and more atom is no longer radioactive, the radioactivity may be largely decayed. I guess it will be related to an exponential decay function.

(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).

- If the person has a disease, it gives the correct result will be a probability of 99%.
- If the person has no disease, the correct result will be a probability of 97%.
- it has been estimated that 0.33% of the individuals actually in the group

The problem is if the person has a positive result, what is the probability that the person actually has the disease?

if there are 100000 persons in the group, 330 of them actually have disease.327 of them are predicted to be positive which is right and 3 of them predict to be negative which are not right.

99670 person is health, 96680 of them predict to be negative which are right and 2990 of them are predicted to be positive which are not right.

The list is as followed:

	actual positive	actual negative
predict positive	327 (right)	2990 (wrong)
predict negative	3 (wrong)	96,680 (right)

It turns out that totally 327+2990=3317 samples are predicted to be positive, the correct results are only 9.9%.

Totally, 96683 samples are predicted to be negative, the correct results are 100%.

For a particular person, if the test has indicated to be positive, the probability of actually has a disease is about 9.9%.

Using Bayes' theorem

If A is for the people has disease

P(A)=0.0033 means 10000 person, 33 of them has disease.

B is for the positive results.

Problem is to find out the A when the results are positive P(A|B)

According to Bayes' theorem,

P(A|B)=P(A)*P(B|A)/P(B)

Where P(B) is the probability of positive results and P(B|A) the probability of positive results when the person has disease.

P(B)= P(B|A)*P(A)+P(B|A')*P(A')

where P(B|A’) is the probability of positive results when person has no disease. P(A’) is the probability of person has no disease.

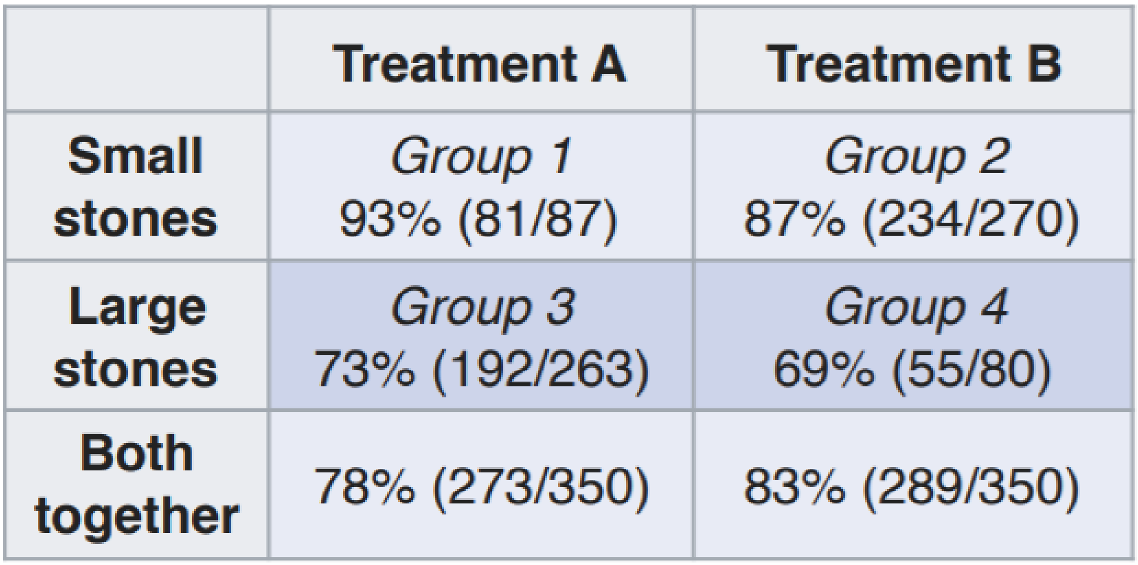
$P(B|A)=0.99$ (a correct result with a probability of 99% if the person has the disease**)**
$P(B|A')=1-0.97=0.03$
$P(A')=1-0.0033=0.9967$

$P(B)=0.99*0.0033+0.03*0.9967=0.033168$

$P(A|B)=0.0033*0.99/0.033168=0.0984=9.84\%$

For a particular person, if the test has indicated to be positive, the probability of actually has a disease is about 9.84%.

(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)

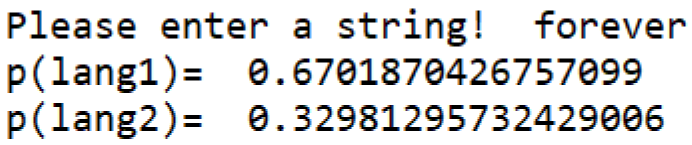
The success rates for both together are not comparable.
Because for treatment A, the data is largely distributed on large stones which have lower success rates compared to its small stones treatment.

For treatment B, the data is largely taken from the small stones which have a higher success rate. In order to have a comparable results, the number of samples for different sizes of stone should be similar for both treatments.

(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!

Investigate



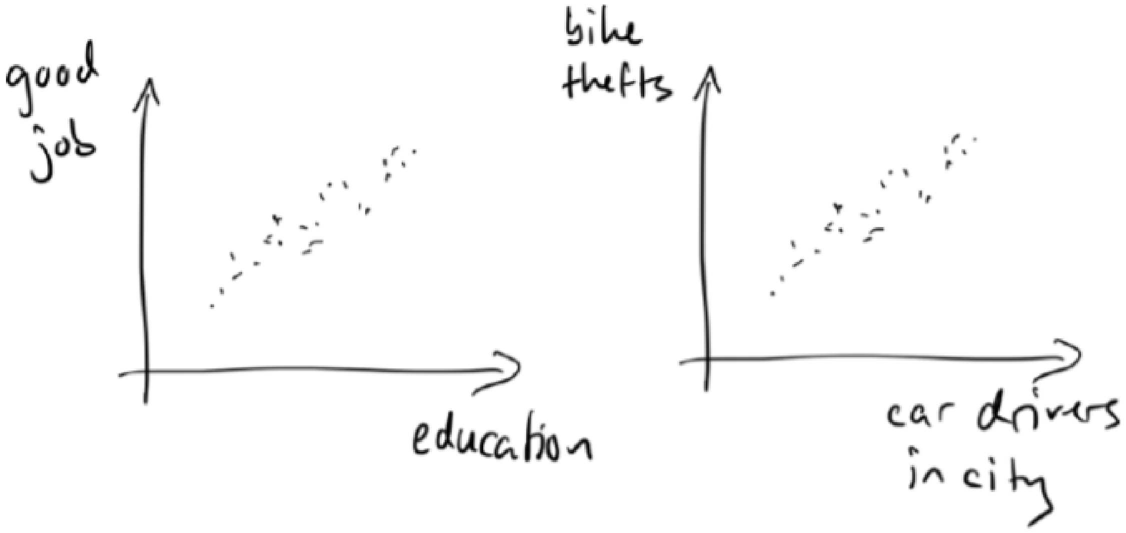
```
Please enter a string!  forever
p(lang1)=  0.6701870426757099
p(lang2)=  0.3298129573242906
```

explain

1. The program get words from different texts written using different language.
2. Calculate the distribution of each character in a language
3. Read a string
4. compare the characters of the string to the character in a language to predict the word coming from which language.

Comments: the correlation between two characters or even more characters are missing in this programme. The results are thus not very believable.

(CORRELATION AND CAUSALITY)
Study the two graphs. What are your observations?



Higher education level, better jobs. Education is correlated with education because high education is required in some good jobs.

More car drivers in the city, more bike thefts. However, the car drivers in the city seem not correlated to the bike thefts. A car driver is not a necessary condition to steal a bike.

There is more people living in the city, which could cause higher bike thefts and more car drivers.

(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)

I will choose statistics for all days in May during these years. The weather seems related to a season, so it will not make sense to statistics for all days in a year. Statistics for May 19 during these years will be so specific to a certain date. The samples will only be 5 if the years are 5. I think May will be in the middle of spring so that I choose statistics for all days in the middle of a season. I could get the weather distributions for May. However, it still hard to evaluate May 19th.

If I had 100 years of weather statistics. I could try statistics for May 19th for 100 years.

(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).

In a stochastic mode is a model that has some random factors, because of its uncertainty. Stochastic models are used to identify unknown potential relationships among variables. for a deterministic model, the relationships are already determined, which thus predicts results with 100% certainty.

b) Does randomness exist? If so, what is it?
No. I don’t think the exact randomness exist.

c) Is there a “right” probability e.g. for rain tomorrow? Or even for a die?

No.

(SELF-CHECK)
pass

Reflection

I. (SUPERVISION AND FOLLOW-UP LECTURE)

a) Did you have your checkpoint meeting for this module?

No

b) Did both of you attend the compulsory follow-up lecture? If you already talked to us about this, please explain.

Yes, I attend the compulsory follow-up lecture.

c) If you were asked to talk to a supervisor about the main submission, who did you talk to?

No

Reflect on your experiences from working with the module and try to make the most out of them. You are also encouraged to discuss your experiences with other groups.

If you reflect around individual problems (which is good), try to also draw general conclusions that may be helpful for you going forward in this course and long-term.

(Time spent in the reflection is time well spent, as it maximizes the learning from the significant effort you already made when working with the problems.)

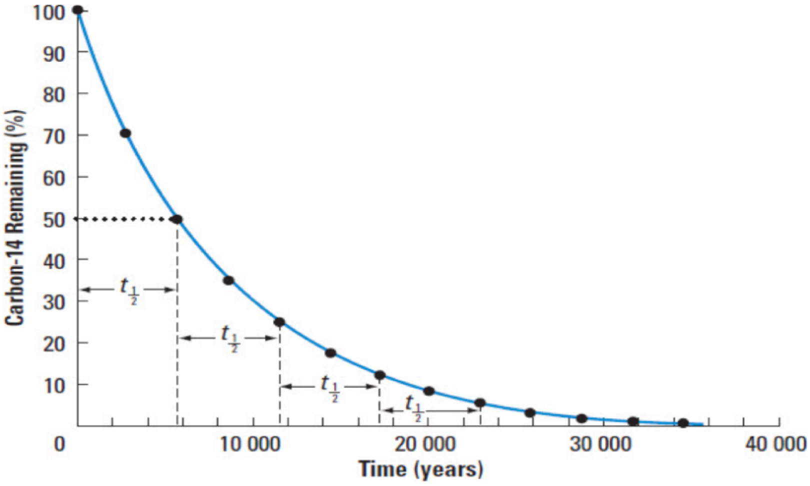
As the answer, give some well-motivated points summarizing your reflections.

I spent 20 hours and tried my best to solve those problems in module 6. I didn't go to supervision but I attended Monday's lecture.

For the first problem in module 6. In the beginning, I didn't misunderstand the meaning of k and n. After I got suggested from TA I become more clear of the meaning of k, and n. I could solve the problem straightforward. First, I analyzed the distribution by rolling 30 dices and then solve the normal distribution function using 3 dices.

For the second problem, I learned three simulations of transport systems using random number generation.

For the third problem, I didn't analyze the radioactivity decay deeply. After I attend the follow-up lecture, I could understand the problem and analyse it better.



The equation governing the decay of a radioactive:

$$C = C_0 e^{-��t}$$

where C0 is the number of atoms which have radioactivity at the beginning (at time t = 0) and C is the number of atoms left after time t
Lifetime (t0.5) is when radioactivity become half of it beginning (C=C0/2)

$$t = ��\ln 2$$

For the fourth problem. I spent a lot of time on the understanding the problem and solved it using the receiver operating characteristic curve (ROC curve) which gave me 9.9%. After I got feedback from TA and attended the follow-up lecture I learned and solved the problem using the Bayes' theorem. The results 9.8% are quite similar to it using the receiver operating characteristic curve.

For the fifth problem. I could solve it quite strightforward.

For the sixth problem, I spent some time to understand code and tried several words, for example, English, Swedish, German. It is quite interesting but the accuracy is not very high, it seems that some correlation lacking.

For the seventh problem, I solved the problem by thing about the possible correlation between variables. I think there is no correlation between the number of drivers and the bike thieves but I didn't find out the reasons for the relationships between these two variables. After I attended the follow-up lecture, I could understand better. The direct relationship between drivers and thieves does not exist. But in the city, there are living more people compared to the countryside. More drivers indicate more people living in the city leading to a result of more bike stolen.

For the last problem, in the beginning, I thought no randomness exist, I couldn't explain why, I thought everything in the world should have the potential relationship between each other, but most of them we could find out it yet. By attending to the follow-up lecture, I learned distribution, there might randomness exist before we could find all the roles in nature. For example, in the Quantum field, how the electron moves, we still can not detect where the electron is.

Summary

In this module, I learned probabilistic models: Monte Carlo simulation, in which using a random number generation. I learned Bayes' theorem and receiver operating characteristic curve (ROC curve) which indicates the reliability of the model. I learned the normalized method when compare two data sets with the different numbers of samples. I learned Normal distribution which could indicate relationships between variables if the physical relationship could not solve. In the final of this module, I learned what the randomness is and whether it exists.

III. (HOW WELL DID YOU SOLVE THE PROBLEMS?)

Give a single assessment for the whole module and motivate with a sentence or two. This is for your own practice.

Use the scale "insufficient/sufficient/good/very good", or a combination such as "between good and very good" or "good or very good". Use the grading criteria we have suggested, or clearly motivate your own.

(We as teachers will then set the grade for this module. We think it is better if you are able to make a fair assessment rather than an inflated one.)

I tried my best when solving the problems and do the reflection. Assessment for this module is very good.

(SELF-CHECK)

passed