Probability methods in built environment

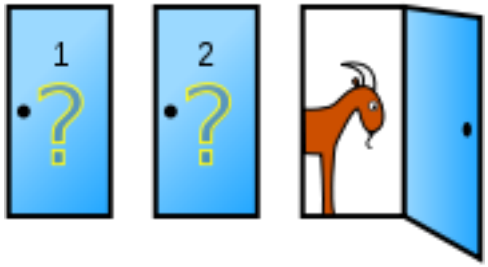
Take home problems for PhD students (2019/20)

Teacher: Goran Turk

April 11, 2020

1 Car or goat

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?



Solve this problem theoretically and by simulations.

2 Birthday problem

Suppose we have 365 days in a year. Suppose you are in a class of n students. What is the probability that at least two persons in a class have the same birthday?

Solve this problem theoretically and by simulations. (n = your age)

3 The Cliff-Hanger

From where he stands, one step toward the cliff would send the drunken man over the edge. He takes random steps, either toward or away from the cliff.The probability of taking a step toward the cliff is equal p, the probability of taking a step away from the cliff is (1-p). What is the probability p_{safe} of coming home safely as a function of p.

Assume two possibilities: the number of steps the drunken man needs to get home is equal your age, the number of steps is infinite. Draw a graph $p_{\text{safe}}(p)$.

Solve this problem theoretically and by simulations.

The problem can be extended to the case, where a drunken man starts k steps from the edge. Draw several graphs $p_{\text{safe}}(p, k)$.

4 Approximate values of mean and variance

We have to find the mean m_Y and variance σ_Y^2 of a random variable Y which is a function of several other random variables X_i. We know the means, variances and covariances of random variables X_i, we know the function $Y = g(X_1, X_2, \ldots)$.

Please, find an example from your field of studies which includes a non-linear relation between random variable Y and at least two random variables X_i. Obtain or assume the means, variances and covariances of these random variables. Derive the approximate equation of the mean and variance of Y and evaluate them. Results should be compared to results obtained by random simulations assuming multi-normal distribution. Before you find, describe and solve your own problem you can try to solve one of the following two examples:

Example #1:

The use of Manning's equation for the average velocity of the water V in an open channel flow:

$$V = \frac{(A/P)^{2/3}(S/100)^{1/2}}{n},$$

where S is the slope of channel bottom $[\%]$, n is the Manning coefficient depending on the material, A is the flow cross-sectional area, determined normal (perpendicular) to the bottom surface $[m^2]$, and P is is the contact length (in the cross-section) between the water and the channel $[m]$. Assume the following characteristics:

Variable	Mean	Standard deviation	Correlation matrix			
A	20	1.0	1	0.7	0	0
P	14	1.5	0.7	1	0	0
S	1	0.1	0	0	1	0.5
n	0.013	0.004	0	0	0.5	1

Example #2: Characteristic lateral bearing capacity f_k of nails in fibreboard is (according to EN 622-2):

$$f_k = 30\, d^{-0.3}\, t^{0.6},$$

where t is the board thickness $[mm]$ and d is the nail diameter $[mm]$. Assume the following characteristics:

Variable	Mean	Standard deviation	Correlation matrix	
t	30	2	1.0	0.1
d	4	0.1	0.1	1.0

5 Random variate generation

Generate a sample X_i corresponding to random varaible X distributed according to beta distribution with parameters ($\alpha = 4, \beta = 6$). Sample size is 10000.

1. Use three methods: inverse method, accept/reject method and a built-in function for random variate generation in Mathematica or any other similar software.
2. Compare CPU times.
3. Draw a histogram and compare it with theoretical PDF.

Assume that $Y = \max_{i=1}^{5} X_i$. Use a generated sample and produce a sample of 2000 Y_is which corresponds to maximum of five random variables X. Draw a histogram of sample Y_i and compare it to a histogram of random sample X_i.

Derive a PDF of Y in draw PDF graphs of X and Y.

6 Random vector generation

PDF of random vector X, Y is defined as

$$f_{XY}(x, y) = a\, x, \qquad 0 \le x \le b, \quad 0 \le y \le x$$

where b is the month of your birthday.

1. Draw a graph of PDF $f_{XY}(x, y)$.
2. Derive marginal and conditional PDFs.
3. Generate a sample of 10000 X_i, Y_i using inverse method.
4. Generate a sample of 10000 X_i, Y_i using accept/reject method.
5. Draw a scatter plot of both samples and compare them. Compare CPU time, too.

7 Parameter estimation

There is a sample of 100 elements in the web-classroom (file "Y.dat").

Estimate distribution parameters using method of moments and maximum likelihood method for three of the following distributions:

1. normal distribution,
2. log-normal distribution,
3. Gumbel distribution of maximums,
4. Frechet distribution,
5. Weibull distribution,
6. Rayleigh distribution,
7. gamma distribution.

8 Robust statistics

The data of population by municipalities in Slovenia is given in the file: "H227 ang.xlsx" published in the web-class (Source: www.surs.si). Take a random sample of 20 elements from two different years.

For one of these two samples evaluate:

1. sample average and standard deviation,
2. median and MAD,
3. Windsorized average and median.

Determine the influence functions for all six above listed statistics and draw their graphs.

Perform a hypothesis testing comparing the mean values of two populations (each year corresponds to one population). Compare the results of t-test with the results of Mann-Whitney test. Was the use of t-test ot Mann-whitney test justified by the sample distribution? Comment the results.

9 ANOVA

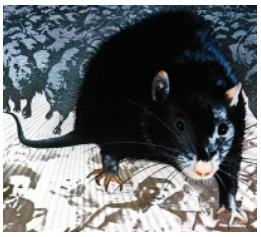
Tensile strength of the paper depends on wood fibre concentration. Four sets of experiments with four different fibre concentrations were completed. Measured data are in the following table:

Tensile strength (kPa)			
5 %	10 %	15 %	20 %
48	83	97	131
55	117	124	172
103	90	128	152
76	123	119	158
62	133	110	126
69	103	-	138

Execute ANOVA hypothesis testing exploring if wood fibre concentration influences the tensile strength of the paper. Perform a-priori and post hoc (posteriori) ANOVA tests. The significance level is 5 %.

10 Nested design - ANOVA

The study analysed the protein intake in the livers of experimental rats. We want to explore if the results depend on the person performing the experiments. Lets assume that we have two technicians: Brad and Janet. Were experiments completed consistently? Brad chooses three rats (A, B C), Janet the other three (D, E, F). Both technicians have measured the intake 10 times.
(Source: http://www.biostathandbook.com/nestedanova.html).



Technician:	**Brad**			**Janet**		
Rat:	Arnold	Ben	Charlie	Dave	Eddy	Frank
	1.119	1.045	0.9873	1.3883	1.3952	1.2574
	1.2996	1.1418	0.9873	1.104	0.9714	1.0295
	1.5407	1.2569	0.8714	1.1581	1.3972	1.1941
	1.5084	0.6191	0.9452	1.319	1.5369	1.0759
	1.6181	1.4823	1.1186	1.1803	1.3727	1.3249
	1.5962	0.8991	1.2909	0.8738	1.2909	0.9494
	1.2617	0.8365	1.1502	1.387	1.1874	1.1041
	1.2288	1.2898	1.1635	1.301	1.1374	1.1575
	1.3471	1.1821	1.151	1.3925	1.0647	1.294
	1.0206	0.9177	0.9367	1.0832	0.9486	1.4543

Perform the nested ANOVA analysis. The significance level is 1 %.

11 Kriging, co-kriging

The Jura data set includes the data on position and several categorical and numerical statistical data. The explanation of the data can be found here: (https://rdrr.io/cran/gstat/man/jura.html). The data is available on our web-class, too.

The basic statistical analysis shows that some of the data are strongly correlated with each other (e.g. Cu-Pb, Co-Ni, Cr-Ni, Zn-Cu, Zn-Pb, Zn-Cr, Zn-Ni).

Your task is to perform the ordinary kriging of one of the elements (chose anyone: Co, Cr, Cu, Ni, Pb or Zn) on a grid 10×10 points anywhere on the region where the data is available. Comment the results.

Your next and the last task is to perform the ordinary co-kriging of the same element which you used before but with the additional information of the other element which is in relatively strong correlation with the previously chosen one. Compare the results of kriging and co-kriging and comment.

12 The end

You have come to the end of it. Please write a report on the solution of all problems. You don't have to be to fancy about it. If you have written the solution by hand just take a photo or scan the paper and append it to the other files. If you typed the solution that's fine, too. Submit also all the programs (your code) which you have produced (Mathematica, Matlab, Phyton or whatever you have used to solve the problems.