STA 247 — Assignment #3. Due in class on November 21.

Late assignments will be accepted only with a valid medical or other excuse.

Worth 9% of the course mark.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. Handing in work that is not your own is a serious academic offense. Fabricating results, such handing in fake output that was not actually produced by your program, is also an academic offense.

Question 1: [40 marks total, 8 for each part] You roll a red die, a green die, and a blue die. These are fair dice, with equal probability of landing on any of their six sides. The red die has the usual numbers 1, 2, 3, 4, 5, 6 marked on its six sides. But on the green die, the sides are marked with the numbers 1, 1, 3, 4, 5, 6 (note that two sides have the number 1), and on the blue die, the sides are marked with the numbers 1, 1, 1, 4, 5, 6 (note that three sides have the number 1). Let X be the sum of the numbers showing on the red, green, and blue dice. Answer the questions below, showing your work. These questions are to be answered with paper, pencil, and perhaps a simple calculator, not by writing an R program (though you can use R as a simple calculator if you like).

a) Find $E(X)$.

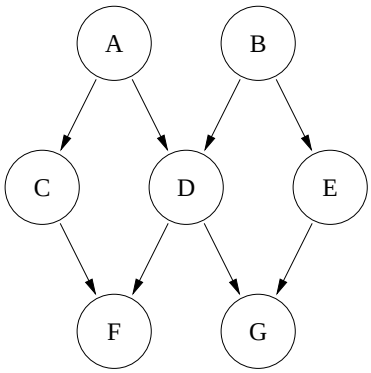
b) Find $\mathrm{Var}(X)$ and $\mathrm{SD}(X)$.

c) If Markov's inequality is applied to the random variable X, what upper bound will it give for $P(X \geq 14)$?

d) Use Chebyshev's inequality to obtain an upper bound for $P(X \geq 14)$.

e) Use Markov's inequality in a cleverer way than in (c) to obtain an upper bound on $P(X \geq 14)$ that is lower than the upper bounds in (c) and (d).

Question 2: [20 marks total, 4 for each part] Consider the following directed graphical model for seven random variables, A, B, C, D, E, F, and G:



Which of the following independence and conditional independence relations can be inferred from this graphical model? Briefly explain why the relationship can or cannot be inferred, with reference to the paths between the variables.

a) A is independent of E.

b) G is independent of C.

c) D is conditionally independent of E given B.

d) F is conditionally independent of G given D.

e) F is conditionally independent of A given C and D.

Question 3: [40 marks total, 20 for each part] Consider again the directed graphical model from Question 2, and suppose that the seven random variables all have range $\{1, 2, 3, 4\}$.

Define the conditional distribution for one of these variables given the values of its parent variables as follows:

- For a random variable with no parents (A or B), the distribution is uniform over $\{1, 2, 3, 4\}$, with each having probability 0.25.
- For a random variable with one parent (C or E), the probability that the child has the same value as its parent is 0.85, and the probabilities of the other three values are each 0.05.
- For a random variable with two parents (D, F, or G), if the parents have the same value, the value of of the child is equal to this value with probability 0.85, with the other three values having probabilities of 0.05, and if the two parents have different values, the child has probability 0.45 of having each of these values, and has probability 0.05 of having each of the other two values.

Equivalently, a child with one or two parents gets one of its parents' values with probability 0.8, choosing between two parents with equal probabilities, and otherwise gets any of the four values (including the values that its parents have) with equal probabilities. For example,

$$P(B = 3) \;=\; 0.25, \quad P(C = 2|A = 2) \;=\; 0.85, \quad P(C = 2|A = 3) \;=\; 0.05,$$

$$P(F = 1|C = 1, D = 1) \;=\; 0.85, \quad P(F = 1|C = 2, D = 1) \;=\; 0.45, \quad P(F = 1|C = 2, D = 4) \;=\; 0.05$$

There are two parts to this question:

a) Write an R function called `sim.ABCDEFG` that simulates from the joint distribution for A, B, C, D, E, F, and G defined by the directed graphical model above, with conditional probabilities as specified. It should takes as arguments a random number seed and the number of sets of values for these seven variables to simulate. It should return a data frame with columns named `A` through `G`, with one row for each set of simulated values. Hand in a listing of your program.

b) Use your simulation program from part (a) to generate 100000 random values for the seven variables, using random seed 1. Use these simulated values to estimate the following conditional probabilities:

$$P(A = 1 \,|\, G = 1)$$
$$P(B = 1 \,|\, G = 1)$$

Then, using the `table` function in R, display tables with estimates of the following joint probability mass functions:

$$P(A = a, B = b)$$
$$P(F = f, G = g)$$

Finally, use `table` to display estimates of the following conditional joint probability mass functions:

$$P(A = a, B = b \,|\, D = 2)$$
$$P(F = f, G = g \,|\, D = 2)$$

Hand in the above results, along with a short discussion (about half a page) of whether they are what you would expect from the graphical model. You may want to run your simulation program one or more additional times with different random number seeds to see how much the results vary randomly.

I will soon post some hints on using R for this question on the web page.