STA 247 — Assignment #2. Due in class on October 28.

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: [35 marks] Suppose that we roll two six-sided dice, one red and one green, with all outcomes of these rolls being equally likely. Let R be the random variable equal to the value showing on the red die, and let G be the random variable equal to the value showing on the green die. Define the random variable X to be $|R - G|$.

a) Give a table of the joint probability mass function of R and X.

b) Find $P(X \geq R)$.

c) Give a table of the probability mass function of X, using the table of the joint distribution for R and X that you found in part (a). Say how you obtained it.

d) Find $P(X > 1)$.

e) Find $E(X)$.

f) Give a table of the probability mass function for the conditional distribution of R given $X = 2$. Say how you obtained it.

g) Is R conditionally independent of G given X? Show why or why not.

This question is to be solved using pencil, paper, and perhaps a simple calculator, not by writing an R program. Include enough details in your answers to make clear how you obtained them. You must include the actual numerical answers (for example, 15/12 or 1.25) not just formulas that could be evaluated to give the answers.

Question 2: [25 marks] You roll a fair six-sided die 100 times. Let X be the random variable that is the number of these rolls that show 1, 2, or 3. You then roll the die another 100 times. Let Y be the random variable that is the number of these rolls that show 1 or 2. With simple calculations in R, using the `dbinom` and `pbinom` functions, find the following:

a) $P(X \leq 60)$.

b) $P(Y \geq 60)$.

c) $P(X = Y)$.

d) $P(X > Y)$.

Hand in your R commands (which should consist of one line for each question) and their numerical output (which should have at least five digits after the decimal point).

Question 3: [40 marks] A software company with three programmers (who we will identify by the numbers 1, 2, and 3) specializes in writing fairly small programs for custom database conversion. The company also employs a program tester, who attempts to find the bugs these programs. We assume that any bugs found are fixed.

These programs are sufficiently similar that it makes sense to consider how good the three programmers generally are at writing reliable programs of this sort, with the programmer assigned and the program that this programmer writes being seen as varying randomly. Let the random variable A be the programmer assigned to write one of these program (either 1, 2, or 3), let the random variable B be the number of bugs in the program before testing (assumed to range from from 0 to 5), and let the random variable R be the number of bugs in the program after testing (ie, the number of remaining bugs that the tester did not find).

Assume that the three programmers are equally likely to be assigned to write a program, so that $P(A=1) = P(A=2) = P(A=3) = 1/3$.

Assume that the number of bugs in a program (before testing) depends on the programmer assigned to write it according to the following table for $P(B=b \mid A=a)$:

b:	0	1	2	3	4	5
$P(B=b \mid A=1)$:	0.55	0.35	0.04	0.03	0.02	0.01
$P(B=b \mid A=2)$:	0.10	0.30	0.35	0.15	0.06	0.04
$P(B=b \mid A=3)$:	0.05	0.25	0.35	0.20	0.10	0.05

We assume that R, the number of bugs remaining after testing, is conditionally independent of A given B — that is, how many bugs the tester finds depends only on how many bugs there are, not on which programmer was responsible for these bugs. We also assume that whether the tester finds one bug in the program is independent of whether the tester finds other bugs. Finally, we assume that the probability that the tester will find any particular bug is 0.6.

Supposed we would like to compute the following three things:

a) The distribution for the number of bugs, R, remaining in a program after testing — that is, a table of values for $P(R=r)$ for $r = 0, \ldots, 5$.

b) The expected value of R — that is, $E(R)$.

c) The conditional distribution for which of the programmers (A) wrote a program, given that it has no bugs remaining after testing ($R=0$) — that is, a table of values for $P(A=a \mid R=0)$ for $a = 1, 2, 3$.

Rather than try to compute these quantities exactly, you should write an R program to estimate them by simulation. You should write an R function called `sim` that takes as arguments a random number seed (`s`) and the number (`n`) of simulated values for A, B, and R to produce. It should return a list with an element named `R.pr` that is a vector of estimated probabilities for R to have the values 0 to 5, an element `E.R` that is the estimated expected value of R, and an element `A.givenR0.pr`, that is a vector of estimated conditional probabilities for A to have the values 1, 2, or 3, given that R has the value 0.

You should hand a listing of this function, and its output when called with `n` set to 1000 and `s` set to 1, with `n` set to 1000 and `s` set to 2, with `n` set to 1000 and `s` set to 3, and with `n` set to 100000 and `s` set to 4.

You may need to use the `sample`, `rbinom`, and `sum` functions, along with arithmetic on vectors and the `for` and `if` statements.