Exercises around M&M §4.1 -- Probability

1 [Exercise 1 from Colton Ch 3] Each time an individual receives pooled blood products, there is a 2% chance of his developing serum hepatitis. An individual receives pooled blood products on 45 occasions. What is his chance of developing serum hepatitis? (Note that the chance is *not* 45x0.02=0.9) To keep it simple, assume that there is a 2% chance that a unit is contaminated and calculate the chance that at least one of the 45 units is contaminated. The 2% shows how old Colton's book is!

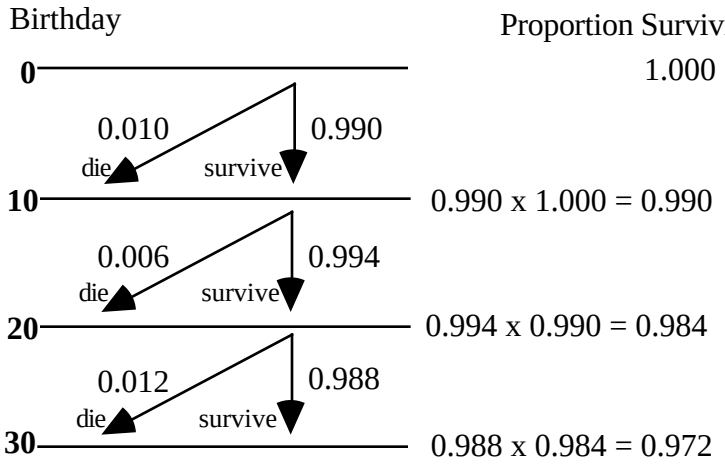
2 A Santé Quebec survey found the prevalence of 4 heart disease risk factors in a certain age-sex group to be: smoking: 32%; family history: 32%; SBP>155mmHg: 12%; diabetes: 5%. If risk factors are distributed independently of each other, what is the proportion of the age-sex group with (a) 4 risk factors (b) 0 risk factors (c) 1 or more risk factors?. A tree diagram may help.

3 The following [conditional] probabilities are taken from the abridged life tables for males (M) and females (F) computed from mortality data for Québec for the year 1990, and published by the Bureau de la statistique du Québec: [the probabilities for 90-year olds have been modified slightly!]. In a (current) life table, one takes current (in this case 1990) i.e. cross-sectional death rates and applies them to a fictitious cohort to calculate what % of the cohort would survive past various birthdays and to calculate the average age at death (also known as life expectancy).

x	prob that person who lives to his / her xth birthday will die during next 10 years: M	F
0	0.010	0.008
10	0.006	0.002
20	0.012	0.004
30	0.016	0.007
40	0.031	0.017
50	0.080	0.042
60	0.211	0.104
70	0.448	0.259
80	0.750	0.585
90	1.000	1.000

a Complete the following tree diagram and calculate the proportions of males who survive past their xth birthday (x = 0, 10, 20, 30, ... 100). Do likewise for females. Plot the proportions vs. x (these plots are called survival curves).

b Calculate, by successive subtractions or otherwise, the [unconditional] proportions [i.e. proportions of the entire cohort] who will die between their xth and x+10th birthdays (x = 0, 10, 20, 30, ... 90). Plot them as histograms. We will use these proportions to calculate life expectancy in a subsequent exercise.



4. Duplicates: To appreciate why the probability of duplicate birthdays is high, take a simpler case of drawing single digit numbers at random from the Random Number Table or spreadsheet until one obtains a duplicate. (also, try actually doing it to see how many draws it takes)

a Calculate the probability that in 5 draws one will not obtain a duplicate, ie. the probability of a sequence
1st# ; 2nd# [≠ 1st#] 3rd#[≠ 2nd# ≠ 1st#]
4th#[≠ 3rd# ≠ 2nd# ≠ 1st#] 5th#[≠ 4th# ≠3rd# ≠ 2nd# ≠ 1st#]

b Calculate, by successive subtractions or otherwise, the probability that the first duplicate will show up on the 2nd, 3rd, ...10th draw. Plot the frequency distribution of the # draws until a duplicate.

Exercises around M&M §4.2 and §4.3 -- Expectation and Variance of Random Variables

5 Suppose one has to analyze a large number of 3 digit numbers. To make the job easier, one rounds each number to the nearest 10
e.g. 460 <-- 461,462, 463, 464 and 465, 466, 467, 468,469 --> 470. If the ending numbers of the unrounded data were fairly uniformly distributed, calculate (a) the average error per (rounded) number (b)the average squared error per (rounded) number.

6 (More advanced) When a binary blood test (one that yields a +ve or -ve result) gives positive results in only a small proportion π of blood samples, it may be possible to economize on the costs of testing by pooling m blood samples, according to the following procedure: each blood sample is divided into two portions; one portion is kept in reserve while the other is pooled with the corresponding portions from m-1 other blood samples; if the result of a single test on the pooled bloods is -ve, the m individual blood samples are considered -ve; if the result is +ve, then the m reserve bloods are individually tested.
With m = 20 and $\pi = 0.1$, calculate the expected number of tests required to determine the status of m blood samples.

7 (More advanced) Sometimes one wishes to estimate via a survey what proportion π of persons in a population have a certain attribute, but respondents may be sensitive to being asked directly about the attribute in an interview. The Randomized Response Technique has been suggested as a way to estimate π. In one suggested variant of the technique, each respondent is asked either Q: Do you have the attribute? or another (unrelated) question Q_u. Q_u must be such that (i) just like Q, it permits yes/no answers (ii) a known proportion θ of persons would answer yes to it and (iii) the interviewer or investigator could not predict individual responses to Q_u. One possible Q_u might be : Was your mother born in April? The interviewer gets the respondent to use a randomization technique to randomly select between Q [with probability φ] and Q_u [with probability 1-φ] and to answer yes or no to the selected question (without telling the interviewer which one is being answered). see also M&M3 page 369, Q4.111.

a If you used this technique on n=100 persons with $\varphi = 0.4$ and $\theta = 0.1$ and obtained p=0.25 or 25% 'yes' answers, what is your best estimate of π? Begin by finding an expression for the expected value of p and then solve for π.

b Rewrite the expression for the estimator of π in general algebraic terms i.e. in terms of the observed percentage and the quantities φ and θ.

c In the interview, when a respondent answers 'yes', what is the probability that (s)he is answering Q rather than Q_u? What if (s)he answers 'no'? Given your answers, what might be 'good' values of φ and θ to use in the technique?

8 (More advanced) Suppose one wishes to estimate via a multiple choice examination [with k answers to choose from for each question], what proportion π of questions a student knows the answer to (excuse the dangling preposition!).

a Show that the simple proportion p of correctly answered questions gives a biased (over)estimate of π if the student simply randomly guesses among the k answers on questions where (s)he doesn't know the answer. Do this by calculating the expected value of p (i.e. the average mark per question) when each answer is marked 1 if correct and 0 if not.

b One can "de-bias" the estimate by marking each correct answer as 1 and each incorrect one answer as m (where m is presumably a negative quantity). What value of m will provide an unbiased estimate of π? Begin by finding the expected mark per question, then set it to π and solve for m.

9 Effect of random error: If "true" values (t) are distributed over {8,9,10,11,12} with probabilities {0.10,0.25,0.30,0.25,0.10} and, independently of t, errors (e) are distributed over {-1,0,1} with probabilities {0.2,0.6,0.2}, then x=t+e will be distributed over {7,8,9,10,11,12,13} with what probabilities? Hint: enumerate all 15 possible configurations of {t,e}; calculate their probabilities by multiplying probabilities of 2 independent events; then for each value of x, gather (add) together the probabilities of the configurations yielding that value. A tree with t as 1st branches and e as its sub branches, or a 2-d table such as in top right side of p2.14, will help. Sketch & visually compare the shapes of the t & x distributions. Quantify your impression by comparing var(x) with var(t). Is var(x) correctly predicted by var(t) and var(e)?