Suggested Exercises from M&M Chapter 4 *[Homegrown exercises begin on next page]*

These pages were updated on September 16

To start with, do some of the odd-numbered exercises. answers to all odd-numbered exercises are given on textbook pages S-1 onwards.

Do some or all of the following even-numbered exercises. You are asked to hand in answers to designated ones.. see the list, and the deadline, on the main course page. Some of these will be discussed in tutorials or answers to them posted on the course web page

§ 4.1	§ 4.2	§ 4.3	§ 4.4		§ 4.5	§ 4
4.8 *	4.12	4.38	4.50	4.52	4.80	4.104
4.10 *	4.14	4.40	4.54	4.56	4.82	4.112
	4.16	4.44	4.58	4.60	4.84	
	4.26	4.46	4.62	4.64	4.88	
	4.30	4.47	4.66	4.68	4.90	
	4.32	4.48	4.70	4.72	4.96	
	4.34	4.49	4.74			
	4.36					

* Simulating random phenomena using Excel

Excel has two ways of generating random numbers that fit a certain pattern or "distribution."

One is the Random Number Generation Analysis **Tool**, which you may have to ask Excel to load each time. It is more extensive, and does more of the work for you if you want a specific distribution, but it is not as flexible if you want to repeat the process with a new sample [i.e. you cannot save the process by typing a formula into a cell]

A **second** way is to **build them yourself**. The building block is the RAND() function under the 'Math and Trig' category of Functions.

Just type **=rand()** into a cell [or ask the wizard!]

The formula will yield a random number between 0.00000.. and 0.99999... (effectively between 0 and 1). You cannot tell what the result will be because it starts the stream of random numbers using a seed which is determined (I suspect) by how long your computer has been running since you last started up. Unlike if you use the Tool, there is no way to control the starting value ("seed") if you use the function -- see the Help on the Random Number Generation Analysis Tool]. If you press the F9 key, it will recalculate and give you another! [you can control the re-calculation using the Calculation Tab under Preferences under the Tools menu]

How to turn this random number (uniform on 0 to 1) into the result of a Bernoulli (0/1) trial where you expect 73% to be 1 (yes) and 27% to be 0(no)?

First, see the simpler example 3.21 on p 269 of text, simulating 60% 1's and 40% 0's using single-digit 'pre-drawn' random numbers from Table B.

To simulate 73% yes and 27% no, you need to take 2 digit random numbers from table B. [you can map the 73 possible 2 digit numbers 00 to 72 into 'yes' and the 27 remaining ones (73 to 99) into 'no'].

With Excel you do likewise: you take advantage of the fact that a proportion 0.73 or 73% of the "Uniform on 0 to 1" distribution is in the interval 0 to 0.73, and 27% between 0.73 and 1.0 (see "spinner" on p 317). If the random number drawn falls in the interval 0 to 0.73 (as it will in some 73% of draws!), you will call it a 'yes"; if it falls in the interval 0.73 to 1.0 (as it will in some 27% of draws!) you will call it a 'no'.

So you just get Excel to determine whether the drawn number is above or below 0.73. i.e. use the

IF(expression, result if true, result if false)

function

=IF(RAND() <0.73 , "yes", "no")

[or do it in 2 steps -- using a cell for the random number and another for the result of the IF .. that way you can check that it is doing what you want!]

For **more complex examples**, see some of the 'simulate' spreadsheets on the course 323 web page. A good one is the "Gambling 17th century: deMéré 100 games" , where I simulate the result of rolling a die (actually several dice all at once!). There, one needs an integer between 1 and 6. As you will see from the formula in say the A1 cell, I take a random number uniform on 0.0 to 0.999.. , multiply it by 6 so that it is uniform on 0 to 5.999.. , add 1 so it is uniform on 1.0 to 6.999.. , then take the integer part of it, so it is uniform on the integers 1 to 6 inclusive.

"Homegrown" Exercises around M&M Chapter 4

-1- Pooled Blood [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- "Clustering" of Cardiovascular Risk Factors ?

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- "Duplicate Numbers" [mini-version of birthday problem]

To appreciate the high probability of duplicate birthdays, take a simpler case of drawing single digit numbers at random from a Random Number Table or spreadsheet until one gets 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, i.e., 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 [Y =] 2nd, 3rd, ...11th draw. Plot the frequency distribution of the # draws, Y, until a duplicate.

*_Hint:_ Let Y = which draw produces 1st duplicate (so Y = 2,... , 11).

```
e.g.  Pr[Y>6] =  Pr[Y = 7 or 8 or 9 or 10 or 11]
      Pr{Y>7] =  Pr[Y =      8 or 9 or 10 or 11]
-------------------------------------------------
      diff.   =  Pr[Y = 7]
```

-4- Errors caused by rounding

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 <-- 460 461 462 463 464
465 466 467 468 469 --> 470.

If the ending numbers of the unrounded data were uniformly distributed (each ending digit has a probability of 1/10), calculate

a the average error per (rounded) number

b the average absolute error per (rounded) number

c the square root of the average squared error per (rounded) number ['root mean squared error', or 'RMSE' for short]

-5- Saving on Binary Tests by Pooling (More Advanced)

When a binary blood test [one that yields a positive ("+ve") or negative ("-ve") result] gives +ve 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 π = 0.1, calculate the expected number of tests required to determine the status of m blood samples.

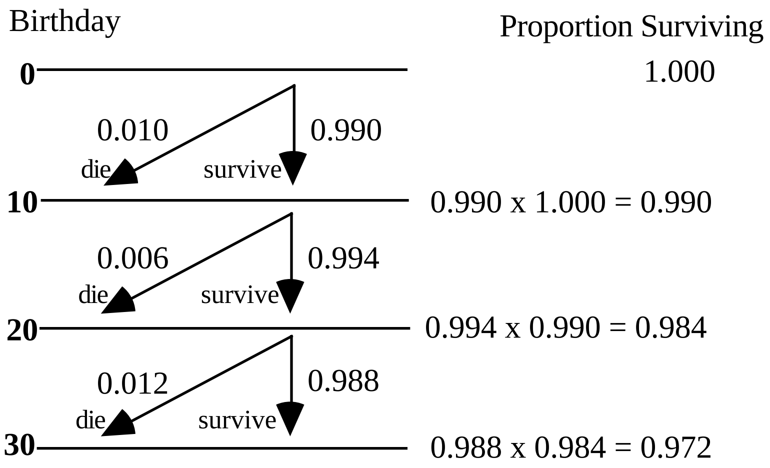
(Drawing a tree diagram may help to track the scenarios)

-6- Life Tables

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 -- if these rates persisted -- and to calculate the average age at death (also known as life expectancy at birth).

	prob that person who lives to his / her xth birthday will die during next 10 years	
x	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). Make sure to label the axes correctly.



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.

```
* e.g. Pr[Die after 70th birthday] = 0.wwww
       Pr{Die after 80th birthday] = 0.zzzz
       Pr[die between 70 and 80]  = difference
```

-7- Life Expectancy at birth vs. at attained ages

From a lifetable, one can also restrict attention to (or "condition on"] those who have already survived to a certain age and calculate the average (or expected) longevity from that point onwards. Full lifetables from the Vital Statistics sections of national statistics agencies (Statistics Canada, ...) usually have a column showing these as a function of attained age.

Web-based health promotion groups have recently started using on-line life-expectancy calculators, some general (only input is age and sex), and some very specific (user must input additional information on life-style, family history, etc.). During the May 2001 version of Course 323, I put some links to some of these on the 323 web page, but some of the links have become out of date rather quickly! You can always find new ones by searching on key words such as 'life expectancy'. Of course, I cannot vouch for how good they are.

Here is one such calculator that I found [the link worked in mid September]:
`http://www.retireweb.com/death.html`

I am struck by the morbid file name -- even though how long one *lives* and the age at which one *dies* are the same quantity!

Unlike others I have seen, the youngest attained age it accepts is 5 -- but (unlike others) it does accept attained ages higher than 100.

a Use it (or any other one you find to calculate life expectancy for persons of your age and sex [Is it OK to call it your life expectancy?] as well as for persons who are older than you by 10, 20, 30, ... years

b Explain to an older relative why -- in these calculations -- he/she has a higher life expectancy than you *[some variant of the word 'conditional' might be helpful]*

[the Spring 2001 article by D Redelmeier, U of Toronto -- web search for Redelmeier Oscar -- on longevity of screen actors and actresses who win an Oscar has a trick analysis issue revolving around this point! It is the same issue that makes bishops live longer than priests -- and cardinal longer than bishops -- and full professors longer than associate professors -- and jazz musicians longer than persons just born!]

c In his "More information on the life expectancy calculation" the author of the above-cited link states that

"Any given individual will have a 50% chance of living longer than the life expectancy and a 50% chance of dying earlier than the life expectancy."

Is the first of these sentences correct? Or is it the case that a good deal more than half of us will live longer than average?

You might find it helpful to use the rough distribution you derived from part b of -6- to straighten him out!

He also states that

"On average, however, the table will produce the correct value."

Is he correct in this?

d [Advanced, optional, should be of interest to Epidemiology and Biostatistics students] From the life expectancy at each attained age, could you reconstruct the life table itself? Hint: Think of the area under the lifetable curve as the total (or average, if the curve begins at Proportion Alive =1.0 at Age =0)

"Homegrown" Exercises around M&M Chapter 4

-8- Correcting for guessing on multiple choice exams

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- Galton's way of showing that the heights of the married couples in his dataset were virtually uncorrelated

See Q2 of "Exercises around Chapter 5" opposite F 18 in the Course 323 web page.

-10- Other Exercises from Course 323

"Exercises around Chapter 2" opposite J 3 in the Course 323 web page.

-11- Testing for HIV

Opposite M 7 in the Course 323 web page.