Supplementary material for ...
Build your own Monte Carlo spreadsheet

Derek Christie
Toi Ohomai Institute of Technology, NZ
thechristiesok@gmail.com

1. Introduction

This supplementary material includes Monte Carlo applications which have been used by students, teachers, and researchers at tertiary level. Some activities need data which can be found in the Excel file Supplementary Data. A complete copy of the Monte Carlo Master sheet has also been included as supplementary data. The instructions below have been abbreviated in many cases, and fuller instructions can be found in the main article. The screen shots have been made from Excel after pressing Crtl+~ which shows the formulas. The order of topics follows roughly those in the main article.

2. Poisson random variables

One more randomization function will extend the usefulness of the Monte Carlo Master sheet. The Poisson distribution models the number of times a random event will occur if you know how many you would have expected on average. Unfortunately there is no Excel function which gives random Poisson variables directly. The following user defined function =Nrand() can be pasted below the other VBA functions. Access these functions through Developer - Macros - MonteCarlo - Edit.

Note that this function =Nrand() has already been included in the Monte Carlo Master sheet. It uses an exact method due to Knuth [1] for values of N below 50, and then uses the normal approximation to the Poisson beyond that.

```
Function Nrand(n) As Integer
Application.Volatile
If n < 50 Then
    L = Exp(-n)
    k = 0
    p = 1
    Do
         k = k + 1
         p = p * Rnd()
    Loop While p > L
    Nrand = k - 1
    Else
    Nrand = Application.WorksheetFunction.NormInv(Rnd, n, n ^ 0.5)
    End If
End Function
```

The =Nrand() function is used in many of the examples that follow.

3. Simulations

3.1. Buffon's needle

Buffon (1707-1788) was a French mathematician who, along many other things, worked out the probability that a needle dropped at random on a plane with ruled parallel lines would intersect one of the lines. He calculated that $p = \frac{2L}{t\pi}$ where L is the length of the needle, and t is the spacing between the lines. L must not be longer than t. This is easily simulated by setting the lines to unit spacing t = 1 as in the sheet below. Set the needle length. Pick a random point in the unit square. Pick a random orientation. (Excel uses radian measure.) Find the other end of the needle and see if it is outside the lines y = 0 and y = 1. The Mean (B10) is the probability of crossing a line. An estimate of π can then be found from this probability.

	F	G	H	I
1		x	y	
2	Random start	=RAND()	=RAND()	
3	Needle length	0.9		
4	Random orientation	=RAND()*2*PI()		Intersect line?
5	Needle end	=G2+G3*COS(G4)	=H2+G3*SIN(G4)	**=IF(OR(H5<0,H5>1),1,0)**
6	Monte Carlo probability	=B10		Linked cell
7	Monte Carlo pi	=2*G3/G6		
8	True pi	=PI()		

Figure 1: Buffon's needle formulas

For needle lengths greater than t, a correction factor is needed. The expression $p = \frac{2L}{t\pi} - \frac{2}{t\pi}\left[\sqrt{L^2 - t^2} + t \cdot \sin^{-1}\left(\frac{t}{L}\right)\right] + 1$ can be left as a challenge for students.

3.2. The Outer Mongolian soft drink factory

VSA (Volunteer Service Abroad) has asked you to manage a soft drink factory in Tuva in Outer Mongolia. Empty bottles go through a bottle washing machine (Washer) then through a combination filler/topper machine (Combo) which fills the bottles and puts the tops on. Both these machines have a 90% chance of working on any particular day. There is nothing you can do if the washer breaks down, but there is a backup to the combination filler/topper. You have an old Filler and an old Topper which you can use instead. Each of these machines has an 80% chance of working. What proportion of days the plant is likely to be working?

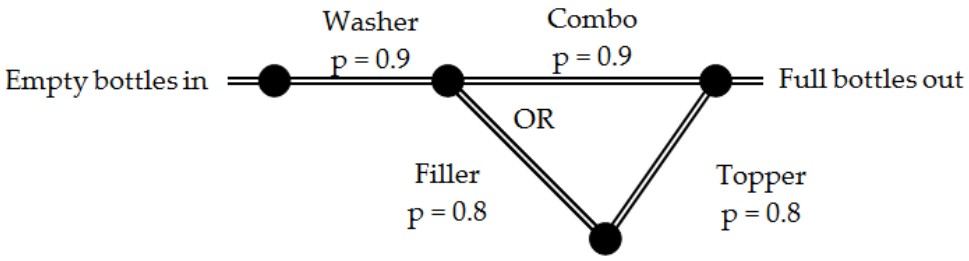

Figure 2: Flow diagram for the Mongolian soft drink factory

	F	G	H
1			p
2	Washer going	=IF(RAND()<H2,1,0)	0.9
3	Combo going	=IF(RAND()<H3,1,0)	0.9
4	Washer + Combo going	=G3*G2	
5	Filler going	=IF(RAND()<H5,1,0)	0.8
6	Topper going	=IF(RAND()<H6,1,0)	0.8
7	Washer + Filler + Topper going	=G2*G5*G6	
8	Plant going	**=MAX(G4,G7)**	
9		Linked cell	

Figure 3: Typical formulas for the Mongolian soft drink factory

The exact answer using a probability tree is 0.874.

3.3. Coprime numbers

Two numbers are coprime if they have no divider greater than 1 in common. For example, 35 and 12 are coprime. A famous result in number theory states that the probability that two integers chosen at random are coprime is $p = 6/\pi^2$. Ignoring the problem of how to pick a random integer over an infinite range, the sheet below picks numbers up to 100 million. The GCD function finds the greatest common divider. If this is 1, the numbers are coprime.

	F	G	H
1	Number 1	=RANDBETWEEN(1,H1)	100000000
2	Number 2	=RANDBETWEEN(1,H2)	100000000
3	Greatest Common Divisor	=GCD(G1,G2)	
4	Coprime?	**=IF(G3=1,1,0)**	Linked cell
5	p	=B10	
6	Margin of error	=1.96*B11/SQRT(B4)	
7	Theory p = 6/pi^2	=6/PI()^2	

Figure 4: Typical formulas for testing coprimality

The Mean (B10) gives the probability that two numbers will be coprime. Will the result still be true if you use numbers from a triangular distribution? Try using

=INT(tri(1,100000000,1000000)) and =INT(tri(1,1,1000000)) In H1:H2. (Yes.)

3.4. Making a shuffled pack of cards

In F1:F4 type 1 1 1 1. In G1:G4 type C D H S. In F5 type =F1+1 and in G5 type =G1. Copy F5:G5 down to row 52. You now have an ordered deck of cards. Highlight the two columns I2:J52, type =perm(F1) and ***Ctrl+Shift+Enter*** as usual. You now have a shuffled deck of cards. Press the F9 key a few times to check.

A variety of problems involving a shuffled deck can now be proposed. For example, what is the probability of being dealt three of a kind in a five card poker hand?

Make your hand the first five cards in the shuffled deck. In J1 type =COUNTIF($H$1:$H$5,H1). (Note the $ signs.) This counts the number of cards in the first five which match the number in H1. If there are three of them the result will be three. Copy down to J5. Add the other formulas.

	F	G	H	I	J	K
1	1	C		=perm(F1)	=perm(F1)	=COUNTIF($I$1:$I$5,I1)
2	1	D		=perm(F1)	=perm(F1)	=COUNTIF($I$1:$I$5,I2)
3	1	H		=perm(F1)	=perm(F1)	=COUNTIF($I$1:$I$5,I3)
4	1	S		=perm(F1)	=perm(F1)	=COUNTIF($I$1:$I$5,I4)
5	=F1+1	=G1		=perm(F1)	=perm(F1)	=COUNTIF($I$1:$I$5,I5)
6	=F2+1	=G2		=perm(F1)	=perm(F1)	=SUM(K1:K5)
7	=F3+1	=G3		=perm(F1)	=perm(F1)	**=IF(K6=11,1,0)**
8	=F4+1	=G4		=perm(F1)	=perm(F1)	**Linked cell**

Figure 5: Looking for three of a kind in a poker hand

Students can work out why a total of 11 in K6 indicates three of a kind. The probability is about 2%. Some other totals are no pair 5, two pairs 9, and full house 13.

In a similar vein, what is the probability of getting a void (no cards in a suit) in a bridge hand of 13 cards? (About 5%.)

	F	G	H	I	J	K	L
1	1	C		=perm(F1)	=perm(F1)	C	=COUNTIF($J$1:$J$13,K1)
2	1	D		=perm(F1)	=perm(F1)	D	=COUNTIF($J$1:$J$13,K2)
3	1	H		=perm(F1)	=perm(F1)	H	=COUNTIF($J$1:$J$13,K3)
4	1	S		=perm(F1)	=perm(F1)	S	=COUNTIF($J$1:$J$13,K4)
5	=F1+1	=G1		=perm(F1)	=perm(F1)		**=IF(MIN(L1:L4)=0,1,0)**
6	=F2+1	=G2		=perm(F1)	=perm(F1)		**Linked cell**

Figure 6: Looking for a void in a bridge hand

3.5. Chuckaluck

Chuckaluck is a gambling game played on a table marked into six areas labelled 1 to 6. A customer is encouraged to put a dollar on any number, say 5. The banker now rolls three dice. If the customer's number fails to come up, he loses his dollar. Otherwise he gets his dollar back plus as many dollars as his number shows. He reasons thus - six numbers and three dice, so in the long run half the dice will show 5 so at least I will break even. Sometimes, however my number will appear twice or even three times so in the long run I should make money. Who has the advantage, the bettor or the bank, and by how much?

	F	G	H	I
1	Chosen number	5		
2	Three dice	=RANDBETWEEN(1,6)	=RANDBETWEEN(1,6)	=RANDBETWEEN(1,6)
3	Successes	=COUNTIF(G2:I2,G1)		
4	Payout	=IF(G3=0,-1,G3)		
5		**Linked cell**		

Figure 7: The Chuckaluck formulas

Somewhat unexpectedly, the bank has an edge of about 8% on each bet.

3.6. The seven letters problem

There are seven different letters and their addressed envelopes. The letters are mixed up and put at random into the seven envelopes, one per envelope. What is the probability that all seven letters will be in the *wrong* envelope? At first sight it looks as if it might be straightforward to work this out analytically, but it isn't as easy s it looks.

It isn't hard to make a Monte Carlo estimate. Set up the numbers 1 to 7 in column F. Make a matching permuted set in column G using =perm(F1).

	F	G	H	I
1	1	=perm(F1)	=IF(F1=G1,1,0)	
2	2	=perm(F1)	=IF(F2=G2,1,0)	
3	3	=perm(F1)	=IF(F3=G3,1,0)	
4	4	=perm(F1)	=IF(F4=G4,1,0)	
5	5	=perm(F1)	=IF(F5=G5,1,0)	
6	6	=perm(F1)	=IF(F6=G6,1,0)	
7	7	=perm(F1)	=IF(F7=G7,1,0)	
8		Matches	=SUM(H1:H7)	**=IF(H8=0,1,0)**
9				**Linked cell**

Figure 8: Formulas for the seven letters problem

The exact probability for n letters is $p = \frac{(n!/e) \text{ to the nearest whole number}}{n!}$ which for n = 7 gives p = 1854/5040 = 0.367857143...

3.7. Missing oystercatcher flocks

A wind farm was proposed near Taharoa Beach, NZ. As part of the site investigation, flocks of migrating oystercatchers were counted by both observers and radar. 821 flocks were seen by observers only, 209 flocks trailed by radar only, and 596 flocks observed by both observers and radar. The problem is to estimate how many flocks were missed by both, and make a 95% confidence interval for total number of flocks.

If we assume independence, then a reasonable estimate for the missing flocks = 209/596x821.

The potential numbers observed flocks can be modelled by Poisson random variables.

	F	G	H	I	J	K	L
1		Radar	No radar			Radar	No radar
2	Seen	596	821		Seen	=nrand(G2)	=nrand(H2)
3	Not seen	209	=G3/G2*H2		Not seen	=nrand(G3)	=K3/K2*L2
4	Total	=SUM(G2:H3)			Total	**=SUM(K2:L3)**	**Linked cell**

Figure 9: Formulas for the total number of flocks

A 95% confidence interval is roughly between 1810 and 2020 flocks.

3.8. The sex ratio of coconut crabs

In 2008, two students from the Bay of Plenty Polytechnic, NZ, did a survey of coconut crabs on Niue Island. Altogether they trapped and measured 49 males and 69 females for a total of 117 crabs.

If the students had done the identically designed survey at slightly different places, or visited their traps on different nights, the figures for males and females would very likely have been close, but different. This would have made their calculation of the sex ratio different as well. Find a 95% confidence interval for the Female/Male sex ratio. Put 1 into the Test value, to see if the sex ratio is significantly different from 1.

The actual number seen of each sex is a random number that closely follows the Poisson distribution, so we can use the =Nrand() function to model the numbers they potentially might have seen.

	F	G	H
1	Coconut crabs	Observed	Simulated
2	Males	49	=nrand(G2)
3	Females	68	=nrand(G3)
4	F/M sex ratio	=G3/G2	**=H3/H2**
5			**Linked cell**

Figure 10: Typical formulas for estimating sex ratio

The true sex ratio is very likely to be between about 0.96 and 2.0, and the two sided p value is about 0.09, so we are not yet convinced that the sex ratio is not 1.

4. Monte Carlo integration

4.1. The generalized sphere

The equation $|x|^n + |y|^n + |z|^n = 1$ is a sphere when n = 2. |x| means the absolute value or positive value of x. Other values for n give a variety of different shapes like a rounded cube (n = 4) or a 3D ninja star (n = 0.5) as seen below. A little unexpectedly, n = 1 gives the equation of an octahedron.

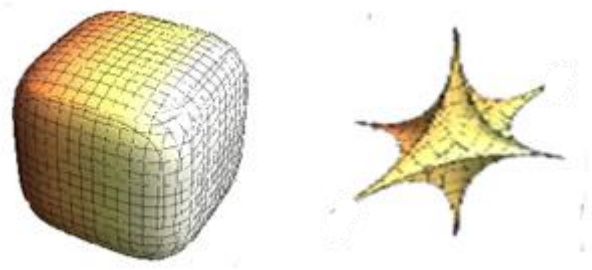

Figure 11: The generalized spheres for n = 4 and n = 0.5

Use Hit or Miss Monte Carlo integration to find the volume of the rounded cube. The solid fits neatly into a 2x2x2 cube.

	F	G	H	I	J
1	x	y	z	n	
2	=1-2*RAND()	=1-2*RAND()	=1-2*RAND()	4	=ABS(F2)^I2+ABS(G2)^I2+ABS(H2)^I2
3					**=IF(J2<1,1,0)**
4		Volume	=B10*8		**Linked cell**

Figure 12: The formulas for n = 4

The volume of the rounded cube is about 6.5, and that of the ninja star about 0.09. The exact volume for the sphere (n = 2) is 4/3π, and for the octahedron (n=1) is 4/3.

5. Resampling and permutation tests

5.1. A permutation anova. Macrophthalmus hirtipes, the stalk-eyed mud crab

Sand has been dredged to deepen a New Zealand marina, and has been dumped on a designated Disposal site. There is concern that a sensitive Impact site will be affected. A third Control site is chosen which is unlikely to be affected by the dredging or dumping. The question of interest is whether the numbers of stalk-eyed mud crab are different at the three sites one year after dredging. With normal data you would perform a one way anova. The crab data, which can be found in the Supplementary Data file, gives the number of crabs

seen in a series of 12 samples per site. It is obviously not normal so a standard anova is not valid. Nor in fact is the non parametric Kruskall-Wallis test applicable which assumes that the shape of the distribution at each site is the same except for change of median. (The same problem renders the Mann-Whitney two sample test invalid for many data sets.) After finding the mean crab numbers at each site, we need some way of measuring how much those means differ. The SD of the means works well. If the SD of the means is small, there is not much difference between the sites. If the SD is large then there probably is a difference. In this case, the SD of the differences is 0.59, but is this large or small? We will assume that there is no connection between the data and particular sites by permuting the data. Then we will see if the 0.59 we see is in the top 5% of the possible values. If it is, we will declare that there is a difference between the sites.

Copy the data from the Supplementary Data file onto the Monte Carlo sheet. Put the appropriate formulas in column I. Copy the data across to K:N. Permute the numbers in column L by highlighting L2:L37 and typing =perm(G2) Ctrl+Shift+Enter. Link N7 and make the test value equal to I7. Go. Use the one sided p value because only the top tail indicates a difference.

	F	G	H	I	J	K	L	M	N
1	Area	Crabs				Area			
2	Disposal	0	Means			Disposal	=perm(G2)	Means	
3	Disposal	1	Disposal	=AVERAGEIF(F:F,H3,G:G)		Disposal	=perm(G2)	Disposal	=AVERAGEIF(K:K,M3,L:L)
4	Disposal	0	Impact	=AVERAGEIF(F:F,H4,G:G)		Disposal	=perm(G2)	Impact	=AVERAGEIF(K:K,M4,L:L)
5	Disposal	0	Control	=AVERAGEIF(F:F,H5,G:G)		Disposal	=perm(G2)	Control	=AVERAGEIF(K:K,M5,L:L)
6	Disposal	0				Disposal	=perm(G2)		
7	Disposal	0	SD	=STDEV(I3:I5)		Disposal	=perm(G2)	SD	**=STDEV(N3:N5)**
8	Disposal	1				Disposal	=perm(G2)		**Linked cell**

Figure 13: The formulas for the permutation anova

As it turns out there is insufficient evidence to claim a significant difference between the sites. The p value is about 0.1.

5.2. Outliers and correlation

The graph below is taken from a medical journal. The annotation on the graph claims a significant correlation between the variables r = 0.31, p = 0.037. However there is an obvious outlier at about (56, 24) which looks as if it might be distorting the results. Can the correlation and p value be trusted? It looks like the whole study may depend on one individual.

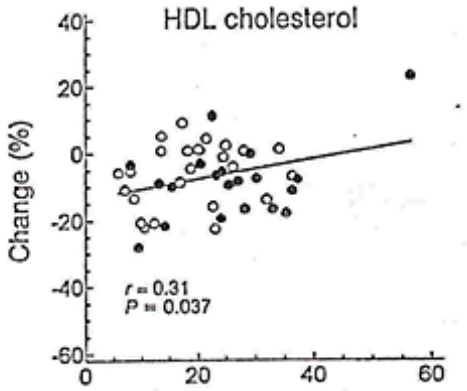

Figure 14: Is this outlier distorting the correlation?

The data can be found in the file Supplementary Data. Copy the data onto the Monte Carlo Master sheet at F1.

	F	G	H	I	J	K	L
1	X	Y			X	Y	
2	56	24			=pick(F2)	=pick(F2)	
3	20	13	Correlation		=pick(F2)	=pick(F2)	Correlation
4	27	1	=CORREL(F2:F43,G2:G43)		=pick(F2)	=pick(F2)	**=CORREL(J2:J43,K2:K43)**
5	37	-7			=pick(F2)	=pick(F2)	**Linked cell**

Figure 15: Resampling to test the significance of a correlation

In H4 type =CORREL(F2:F43,G2:G43). Make another copy in columns J to L. This is paired data, so highlight both columns J2:K43 and type =pick(F2) ***Ctrl+Shift+Enter*** as usual.

Set the test value to 0 and Go. The 95% confidence interval is roughly (-0.1, 0.6) and the two sided p value is more than 0.05 so we cannot accept the claim for a significant correlation.

5.3. Cronbach's alpha

As part of a larger project, a nursing lecturer composed a set of five questions, the total of which aimed to measure cultural identity among nursing students. The internal consistency of such a set of questions is often measured by Cronbach's alpha which is given by $\alpha = \frac{p}{p-1}\left[1 - \frac{\sum V_i}{V_T}\right]$ where p is the number of questions, Vi is the variance of the responses to the i^{th} question, and V_T is the variance of the total. For our data set p = 5.

Alpha can range from 0 (no consistency) to 1 (perfect consistency). A value of 0.8 is considered good and 0.7 usually considered acceptable. The value of alpha calculated for the researcher's set of questions using the responses from 50 students was 0.82 which is hopeful. However, she would be like to be 95% sure that the true value of alpha is greater than 0.7 before she publishes.

The data from the 50 students can be found in the Supplementary Data file. The formulas below show the formulas needed to calculate alpha.

	F	G	H	I	J	K	L	M
50	49	4	3	3	3	3	=SUM(G50:K50)	
51	50	3	2	2	3	3	=SUM(G51:K51)	
52								Alpha
53	Var	=VAR(G2:G51)	=VAR(H2:H51)	=VAR(I2:I51)	=VAR(J2:J51)	=VAR(K2:K51)	=VAR(L2:L51)	=5/4*(1-SUM(G53:K53)/L53)

Figure 16: Cronbach's alpha confidence interval

Copy columns F:M over to O:V in the normal way. Highlight the copied question response data and use =pick() to resample with replacement, keeping the matched nature of the data. Link the resampled data to B3. This will be a one sided test because the researcher wants to show that alpha is greater than 0.7. Set the lower percentile to 5% and the test value to 0.7. Use a one sided p.

The one sided p value is about 0.11, and the one sided 95% confidence interval for alpha is about 0.65 to 1. This means that there is insufficient evidence here to claim that the true value of alpha is greater than 0.7. More data is needed.

6. Monte Carlo Risk Analysis

The Monte Carlo risk analysis example in the main article investigated financial risk. The general idea of simulating scenarios need not necessarily involve the quantification of risk. Exactly the same process can be used to investigate the likely range of some other item of interest by generating and collating the range of plausible scenarios.

6.1. Coconut crabs revisited

Example 3.8 involved a student research project surveying coconut crabs on Niue Island in the South Pacific. One of their aims was to estimate the likely range of the population (a 66% confidence interval).

They put out 15 lines of 12 baits each, visiting each line on two different nights. The result of these 360 bait visits was a total of 117 crabs caught. This gives an average CPUE (catch per unit effort) of 0.325 crabs per bait visit. An earlier coconut crab survey on another similar Pacific island, Vanuatu, found that a factor of 12,000 allowed an approximate estimate of crabs per square km.

Density in crabs per square km = CPUE x 12 000.

Probably a similar factor will work for Niue. An experienced ecologist suggested the factor for Niue would lie somewhere between 9000 and 15000.

Coconut crabs inhabit a strip of land along the coast between 1.5 and 2.5 km wide and between 25 and 35 km long. Use this data to make a "likely" 66% confidence interval for the population. Use =Nrand() for the crabs seen and =tri() for the factor and the lengths.

	F	G	H	I	J	K	L
4	Total bait visits	=G1*G2*G3		=I1*I2*I3			
5	Crabs seen	117		=nrand(G5)			
6	CPUE Catch per unit effort Crabs/bait visit	=G5/G4		=I5/I4	Low	Mid	High
7	CPUE to Crabs/sq km factor	12000		=tri(J7,K7,L7)	9000	12000	15000
8	Crabs/sq km	=G6*G7		=I6*I7			
9	Suitable coast length	30		=tri(J9,K9,L9)	25	30	35
10	Inland range	2		=tri(J10,K10,L10)	1.5	2	2.5
11	Area sq km	=G9*G10		=I9*I10			
12	Total population	=G8*G11		=I8*I11			

Figure 17: Monte Carlo estimate of the coconut crab population

A 66% confidence interval is found between the 17th and 83rd percentiles, and is roughly between 190 000 and 270 000 crabs.

6.2. Estimating student hours

This example is a simplified version of a faculty planning exercise to estimate the probable number of total student hours of mathematics tuition needed the following year in a small polytechnic science department. This estimate directly affects departmental funding.

- Certificate Mathematics Semester 1. 6 hours per week for 15 weeks. About 16 enrolments expected. If there are less than 12 enrolments the class will not run.
- Certificate Mathematics Semester 2. 6 hours per week for 15 weeks. Probably about 4 students will not return from Semester 1. This class will run if the Semester 1 class ran, irrespective of numbers.

- Diploma Mathematics Year 1. 6 hours per week for 30 weeks. About 35 enrolments expected.
- Diploma Mathematics Year 2. 5 hours per week for 30 weeks. There are 32 enrolled in Year 1 this year. About 5 will not return.
- Pre Nursing Drug Calculations. 2 hours per week for 15 weeks. About 40 enrolments expected.

The actual number of students enrolling or not returning are assumed to be random, and can be modelled by the Poisson distribution using our function =Nrand(). Find a 66% confidence interval for the total number of student hours of maths teaching anticipated.

	F	G	H	I	J	K
1		Applied	Started	Hours	Weeks	Total
2	Certificate 1	=nrand(16)	=IF(G2<12,0,G2)	6	15	=H2*I2*J2
3	Certificate 2	=H2	=IF(G3=0,0,H2-nrand(4))	6	15	=H3*I3*J3
4	Diploma 1	=nrand(35)	=G4	6	30	=H4*I4*J4
5	Diploma 2	=32-nrand(5)	=G5	6	30	=H5*I5*J5
6	Nursing	=nrand(40)	=G6	2	15	=H6*I6*J6
7						**=SUM(K2:K6)**
8						**Linked cell**

Figure 18: Student mathematics hours risk analysis

Set the percentiles to 17% and 83%. The number of student mathematics hours is "probably" between about 13 300 and 16 200.

7. Reference

[1] Donald E. Knuth (1969). Seminumerical Algorithms. The Art of Computer Programming, Volume 2. Addison Wesley.