Arithmetic mean and normal distribution

Rolling ten dices gave this picture

Fig. 1 Result of rolling 10 dices

Adding the number of eyes on each face and dividing the sum by 10 gives the arithmetic mean

$$\frac{1+1+3+6+1+5+5+5+1+6}{10} = \frac{17}{5} = 3.4$$

One can do this a lot of times. Values of the mean will be numbers belonging to the set

{1, 1.1, 1.2, 1.3, ..., 5.8,5.9,6}

Using Excel we shall explore patterns in distribution of means like these.

Rolling 100 "dices" using Excel

Fig. 2 shows the outcome of simulating the roll of 100 dices and computing the arithmetic mean of each roll

	A	B	C	D	E	F	G	H	I	J
1	Arithmetic means of 100 rolls with 10 dices									
2	4,0	3,0	2,4	2,8	2,6	3,2	2,9	3,2	3,3	3,1
3	2,7	3,3	3,4	3,7	3,1	5,0	3,7	3,2	3,1	3,6
4	4,4	3,3	3,8	3,5	3,4	3,2	3,2	2,9	4,2	3,4
5	4,1	3,1	4,0	4,3	3,5	4,5	4,1	3,2	2,8	3,7
6	3,0	3,1	3,6	3,6	3,3	3,6	3,4	4,6	3,4	2,5
7	3,5	4,6	3,0	3,8	3,8	3,9	4,1	4,1	3,0	2,7
8	3,7	3,5	4,0	2,8	3,3	3,4	3,1	2,6	3,6	3,9
9	3,5	3,0	3,9	3,0	3,6	3,3	2,5	3,4	3,6	2,6
10	4,4	3,4	3,6	4,0	3,9	3,7	4,3	4,5	4,8	4,4
11	3,2	4,2	4,2	4,6	3,8	3,2	3,3	4,1	4,0	4,1

Fig. 2 Screen capture from Excel

The formula used in each cell is shown below. SLUMPMELLEM is the Danish word for RANDBETWEEN. Notice that each call of the function acts independent of the others. Therefore you can have a formula for 10 independent dices and the calculation in one cell.

=(SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6))/10

Formulas for processing the data is shown in Fig. 3.

K	L	M	N
Number of rolls	Possible values	Frequency	Reletive frequency
100	1	0	0
	1,1	0	0
	1,2	0	0
	1,3	0	0
	1,4	0	0
	1,5	0	0
	1,6	0	0
	1,7	0	0
	1,8	0	0
	1,9	0	0
	2,0	0	0
	2,1	0	0
	2,2	0	0
	2,3	0	0
	2,4	0	0
	2,5	0	0
	2,6	1	0,01
	2,7	4	0,04
	2,8	3	0,03
	2,9	3	0,03
	3,0	4	0,04

K	L	M	N
Number of rolls	Possible values	Frequency	Reletive fr
=TÆL(A2:J501)	1	=TÆL.HVIS($A$2:$J$502;L2)	=M2/$K$2
	=L2+0,1	=TÆL.HVIS($A$2:$J$502;L3)	=M3/$K$2
	=L3+0,1	=TÆL.HVIS($A$2:$J$502;L4)	=M4/$K$2
	=L4+0,1	=TÆL.HVIS($A$2:$J$502;L5)	=M5/$K$2
	=L5+0,1	=TÆL.HVIS($A$2:$J$502;L6)	=M6/$K$2
	=L6+0,1	=TÆL.HVIS($A$2:$J$502;L7)	=M7/$K$2
	=L7+0,1	=TÆL.HVIS($A$2:$J$502;L8)	=M8/$K$2
	=L8+0,1	=TÆL.HVIS($A$2:$J$502;L9)	=M9/$K$2
	=L9+0,1	=TÆL.HVIS($A$2:$J$502;L10)	=M10/$K$2
	=L10+0,1	=TÆL.HVIS($A$2:$J$502;L11)	=M11/$K$2
	=L11+0,1	=TÆL.HVIS($A$2:$J$502;L12)	=M12/$K$2
	=L12+0,1	=TÆL.HVIS($A$2:$J$502;L13)	=M13/$K$2

Fig. 3 Screen capture showing formulas for counting and calculating statistical parameters. TÆL.HVIS is the function COUNTIF

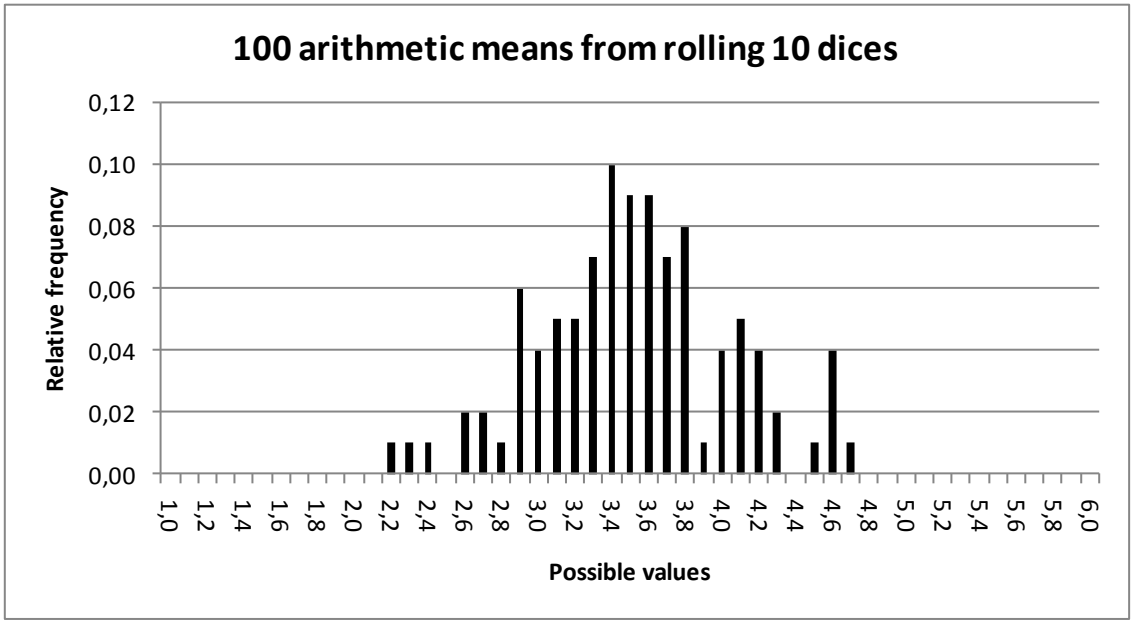

Fig. 4 Bar diagram visualising the statistics. Bar diagram is used since we know the the possible (discrete) values for the mean of 10 dices.

Having sat up the spreadsheet it is just a matter of pressing function key F9 to do another roll of 10 dices. And you can do this a lot of times studying the stochastic nature of the phenomenon. On the one hand very randomly distributed but anyway there seams to be some kind of pattern: They cluster in the middle.

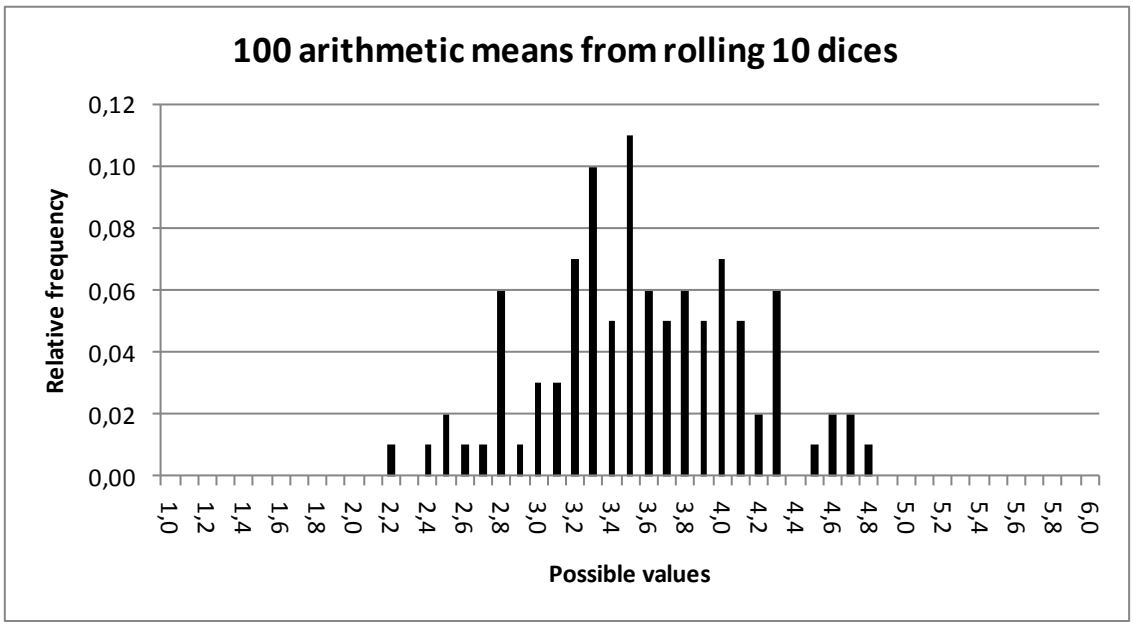

Fig. 5 One more roll of 10 dices

Computers and spreadsheets now a days can handle very large amount of data so why not go for it and see what can happen if we roll the dices many more times.

Since the structure and formulas for the spreadsheet is already keyed in it is a question of copying cells to appropriate regions.

Next two figures shows 5000 rolls with the 10 dices.

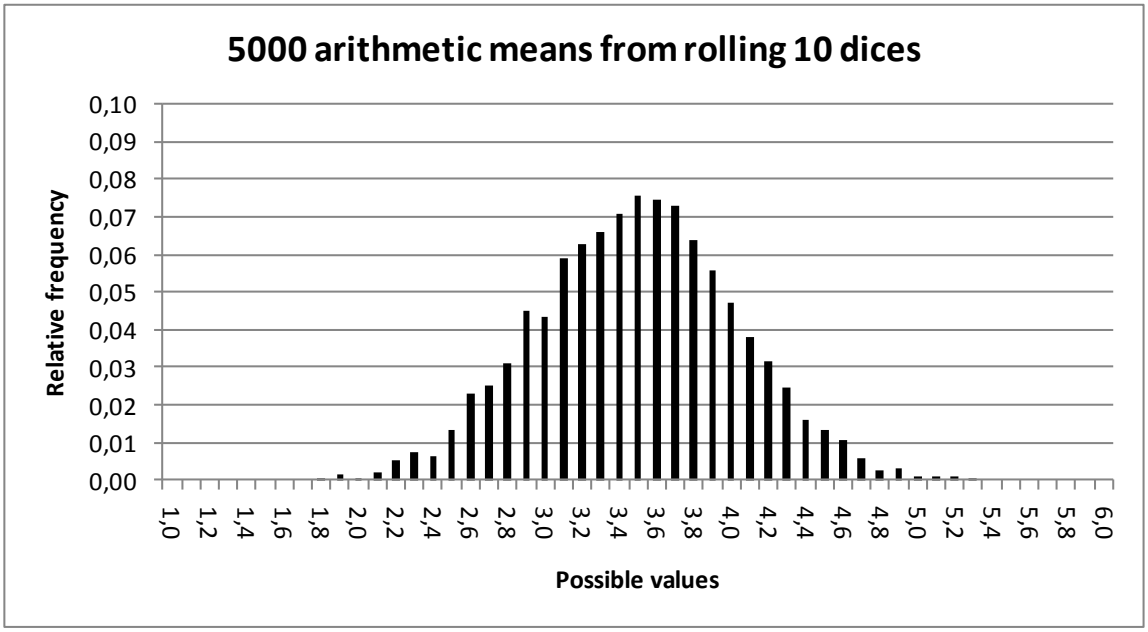

Fig. 6 Bar diagram showing result for 5000 rolls.

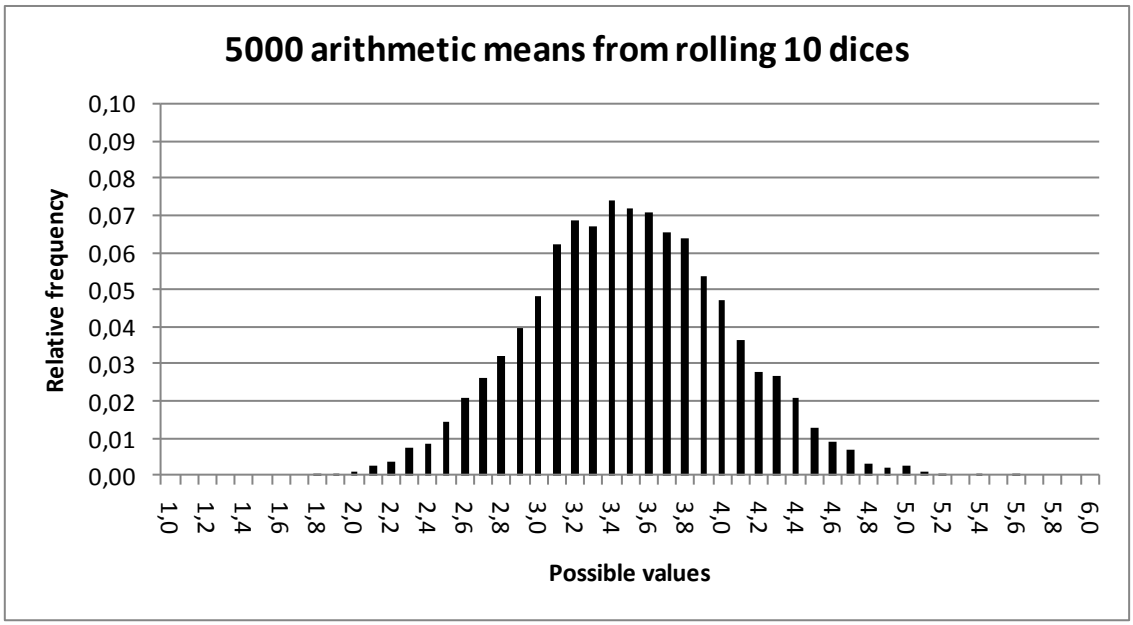

Fig. 7 Pressing F9 rolls the dices 5000 times more.

We now see a pattern that looks more regular in a way. And one gets an idea of what can be expected. Betting a fortune on means near 1 and 6 doesn't seem wise.

This calls for refining the spreadsheet so that many more rolls can be made. Why not a hundred thousand? There is space enough. Number of rows in mine version of Excel 2007 is

$$2^{20} = 1048576$$

and number of columns is

$$2^{14} = 16384$$

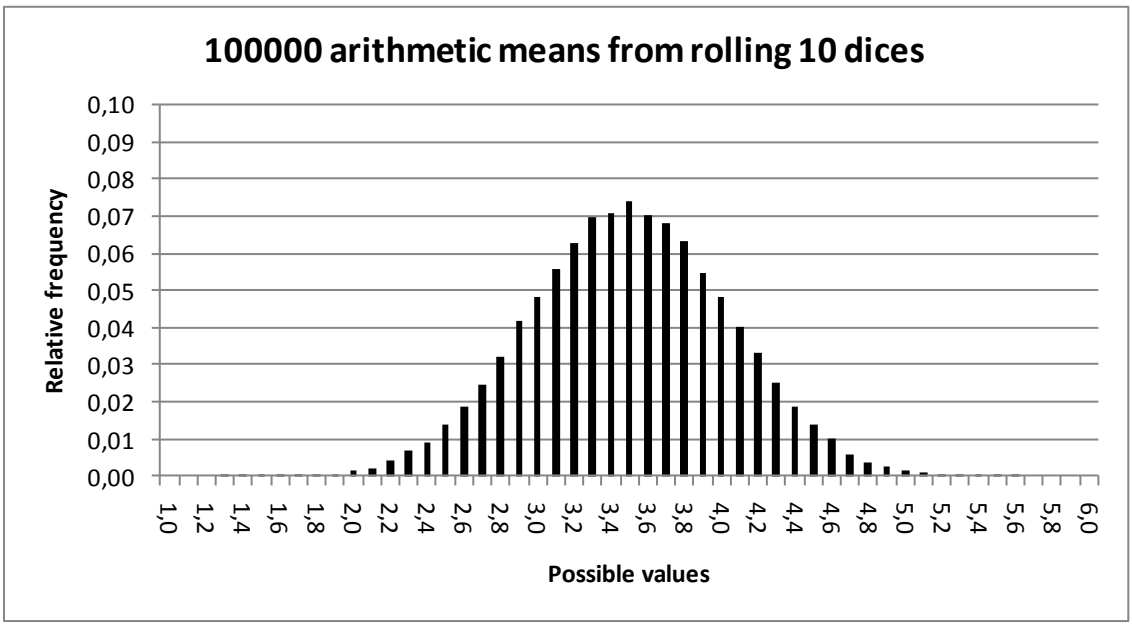

Fig. 8 Beautiful bell shaped pattern of the barchart for 100000 rolls.

Pressing F9 brings you another picture looking almost as the same as Fig. 8. And doing it over and over produces diagrams very much alike.

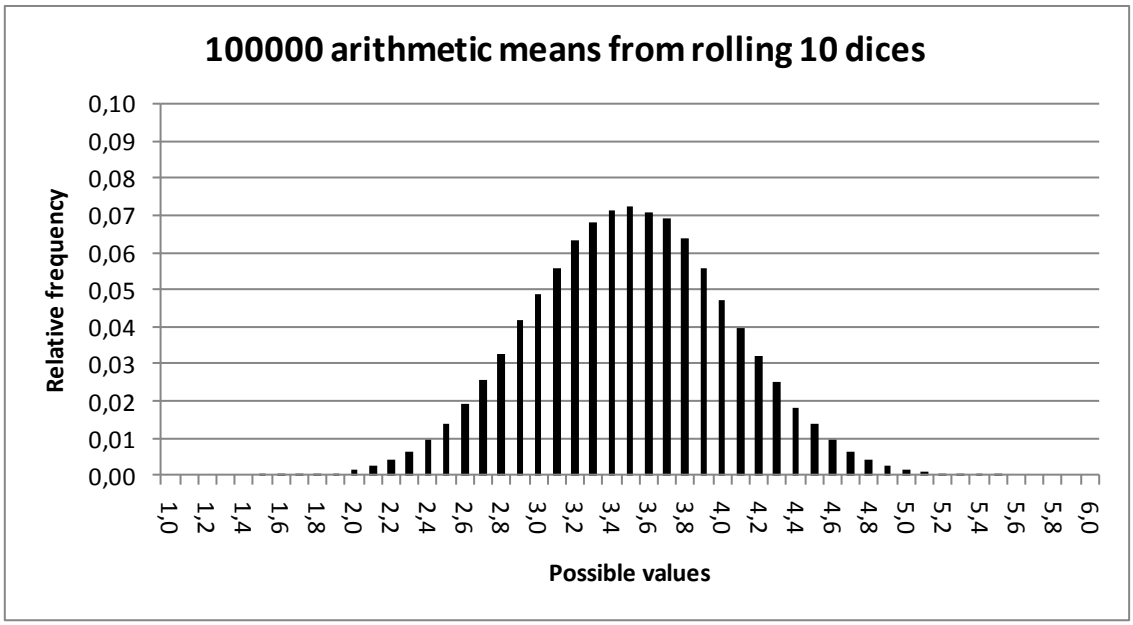

Fig. 9 Another 100000 roll of the 10 dices produce means distributed almost as in Fig. 8.

The only difference in formulas compared to Fig. 3 is the region which is counted by COUNTIF.

K	L	M	N
Number of rolls	**Possible values**	**Frequency**	**Reletive frequency**
=TÆL(A2:J10001)	1	=TÆL.HVIS($A$2:$J$10001;L2)	=M2/$K$2
	=L2+0,1	=TÆL.HVIS($A$2:$J$10001;L3)	=M3/$K$2
	=L3+0,1	=TÆL.HVIS($A$2:$J$10001;L4)	=M4/$K$2
	=L4+0,1	=TÆL.HVIS($A$2:$J$10001;L5)	=M5/$K$2
	=L5+0,1	=TÆL.HVIS($A$2:$J$10001;L6)	=M6/$K$2

Fig. 10 Part of formulas for counting 100000 rolls of 10 dices. TÆL.HVIS is danish for COUNTIF

The regular shape of the bar chart in Fig. 9 suggests that some formula describing it may exist. And it does. Perhaps I wouldn't be writing this article if I didn't know in advance that it is already found many years ago. It is the famous Gaussian or normal distribution.

The inspiration to the article came from a wish to concretise that distribution by giving a concrete example of the knowledge held by the so called Central Limit Theorem from probability.

I will not go into details since it is not a proof but rather an exemplification rooted in concrete actions I go for – as long as you can call computer simulation concrete actions – and so I advocate. For details see [1] and [2].

Checking limiting distribution of rolling 10 dices

What to look for? Central limit theorem suggest that the arithmetic mean will be approximately normal distributed with mean m and variance $s^2/10$, where m and s^2 are mean and variance of throwing one dice.

Instead of sticking to formulas I calculate these numbers by using Excel. It will make their definitions more concrete to us. So take a look of Figs. 11 and 12.

	A	B	C	D	E
1	Number of eyes	Proba-bility	Calculating mean	Calculating variance	
2	k	P(D = k)	k·P(D = k)	$(k-m)^2 \cdot P(D = k)$	
3	1	1/6	0,1667	1,0417	
4	2	1/6	0,3333	0,3750	
5	3	1/6	0,5000	0,0417	
6	4	1/6	0,6667	0,0417	
7	5	1/6	0,8333	0,3750	
8	6	1/6	1,0000	1,0417	
9			Mean	Variance	Standard deviation
10			$\Sigma k \cdot P(D = k)$	$\Sigma(k-m)^2 \cdot P(D = k)$	
11			m	s^2	s
12			3,5000	2,9167	1,7078

Fig. 11 Calculating mean, variance and standard deviation for one dice.

	A	B	C	D	E
1	Number of eyes	Probability	Calculating mean	Calculating variance	
2	k	P(D = k)	k·P(D = k)	$(k-m)^2 \cdot P(D = k)$	
3	1	=1/6	=A3*B3	=(A3-$C$12)^2*B3	
4	2	=1/6	=A4*B4	=(A4-$C$12)^2*B4	
5	3	=1/6	=A5*B5	=(A5-$C$12)^2*B5	
6	4	=1/6	=A6*B6	=(A6-$C$12)^2*B6	
7	5	=1/6	=A7*B7	=(A7-$C$12)^2*B7	
8	6	=1/6	=A8*B8	=(A8-$C$12)^2*B8	
9			Mean	Variance	Standard deviation
10			$\Sigma k \cdot P(D = k)$	$\Sigma(k-m)^2 \cdot P(D = k)$	
11			m	s^2	s
12			=SUM(C3:C8)	=SUM(D3:D8)	=KVROD(D12)

Fig. 12 Formulas for the calculations in Fig. 11. KVROD is Danish for SQRT

Since normal distribution is a continuous distribution we cannot directly compare the bar chart from Fig. 9 with the probability density of a normal distribution.

If we consider the 100000 means as being instances of continuous variables we will group the numbers in intervals. Counting the number of means in every subinterval gives the interval frequency. Dividing these by 100000 gives relative interval frequencies and they are represented by rectangles over each interval with an area equal to the relative interval frequency.

Dividing into subintervals of length 0,5 we get the pictures in Figs. 13 and 14

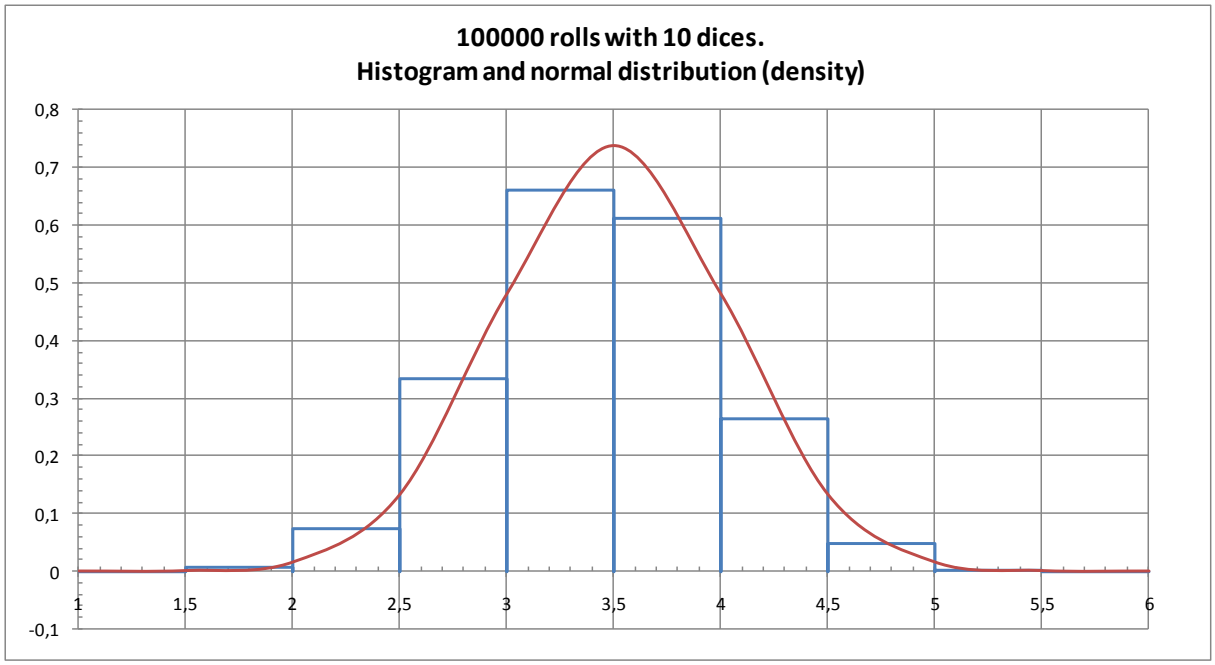

Fig. 13 Histogram compared with normal distribution with mean 3,5 and standarddeviation 0,54

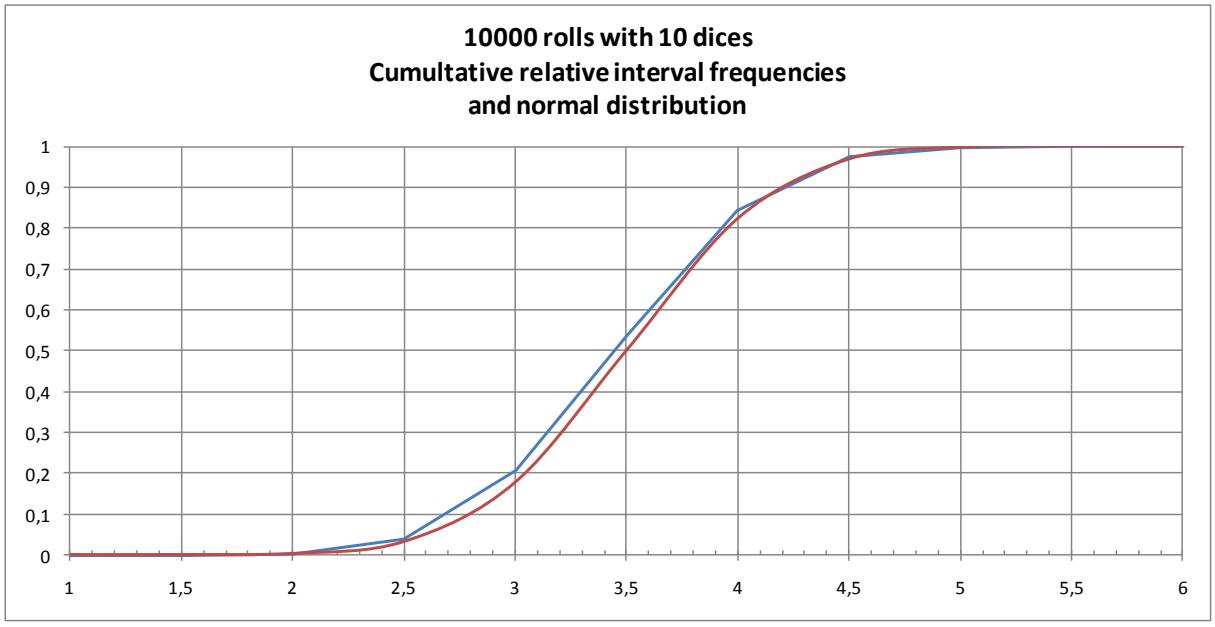

Fig. 14 Cumultative normal distribution with same parameters as in Fig. 13 compared to cumultative relative intervalfrequencies.

Since there is 100000 numbers in the data set we can refine the subdivision. In Fig 14 and 15 intervals are chosen as 1.05<1.25<1.45<.....<5.85<6.05 to have all observations in the interior of intervals.

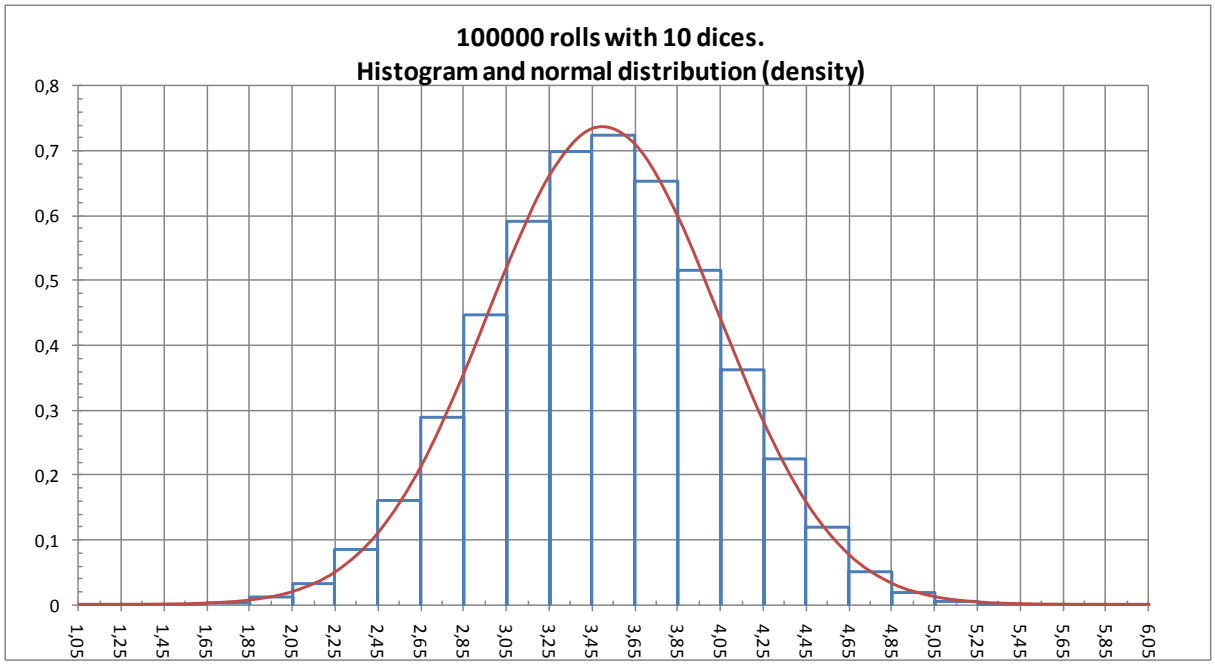

Fig. 15 Now with smaller subdivisions as explained in the text.

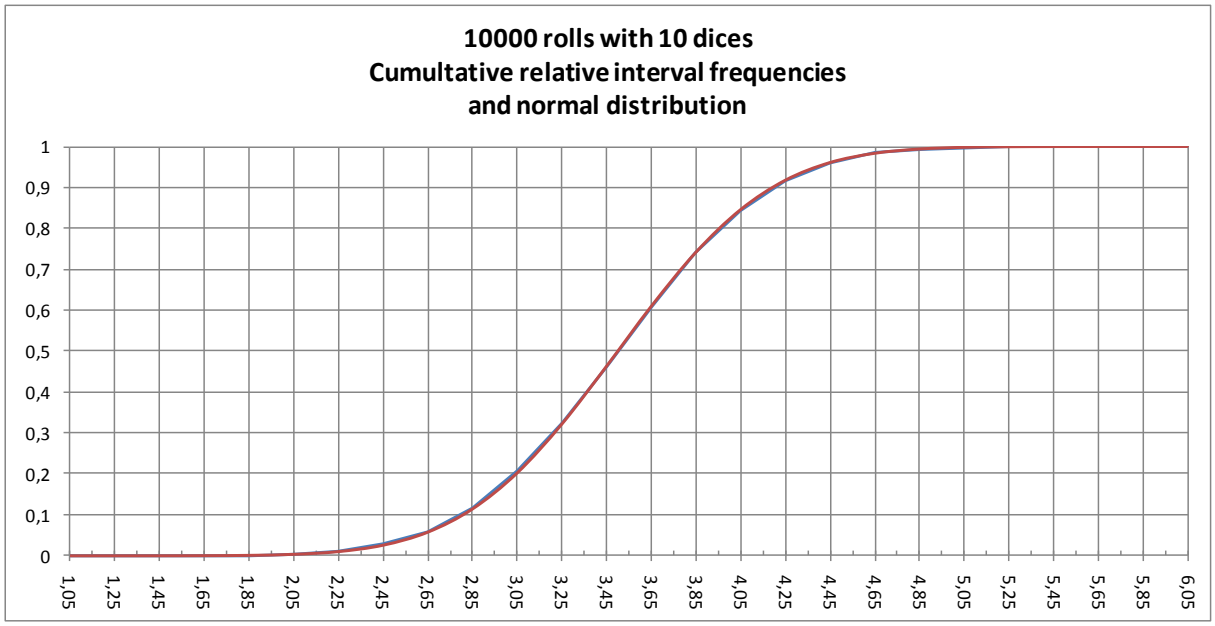

Fig. 16 Now with smaller subdivisions as explained in the text.

Next we try rolling 100 dices and calculating average. I use almost the same spreadsheet as shown i Fig. 3.

To key in the function calculating the average of the 100 dices copy-paste from a Word document is used to create the formula shown in Fig 17 which is then copied to the first cell in the spreadsheet. Rest is copying in the spreadsheet as usual.

```
=(SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;
6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1
;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(
1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM
(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLE
M(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELL
EM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMEL
LEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPME
LLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPM
ELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMP
MELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUM
PMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLU
MPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SL
UMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+S
LUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+
SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)
+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6
)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;
6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1
;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(1;6)+SLUMPMELLEM(
1;6)+SLUMPMELLEM(1;6))/100
```

Fig. 17 Large formula for calculating average of 100 dices in one cell. But even a formula that big can be treated as much smaller formulas.

This time there are 501 possibly values for the average (1.00<1.01<....<5.99<6.00). If one would create a bar graph as e.g. in Fig. 4 there would be too many bars to be handy. Instead of using a bar graph a xy-point plot is used. Only dots are marked and you can think of them as marking the upper bound of bars.

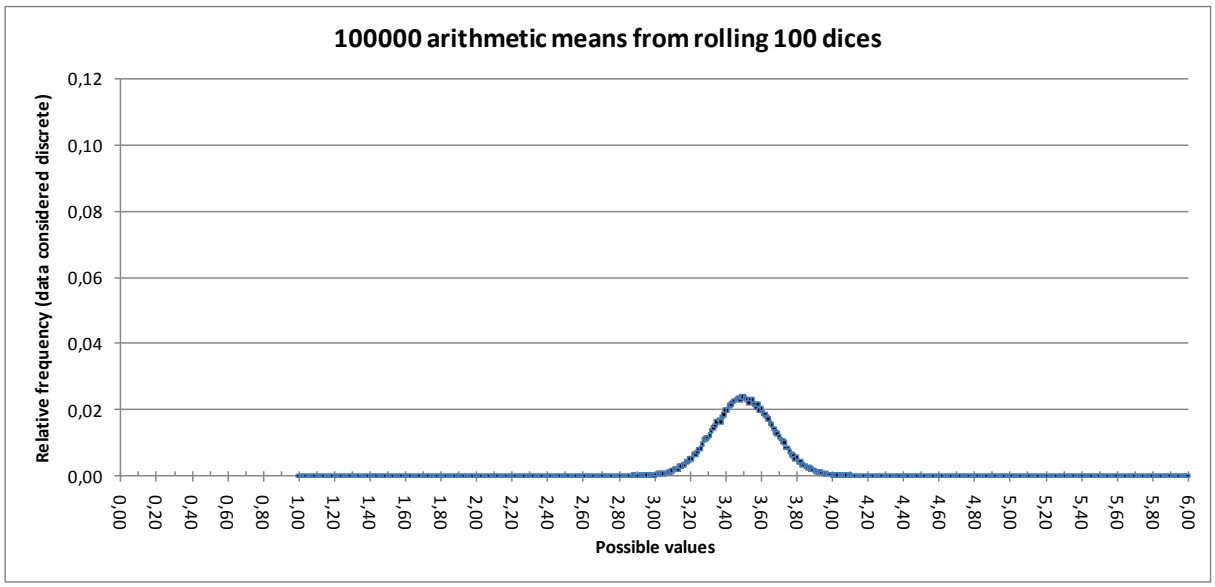

Fig. 18 xy-plot used to illustrate bar graph.

Each dot representing a bar above a 2-decimal number on the abscissa

You can compare with Fig. 19 which is a xy-plot version of bar graphs such as shown in Figs. 8 and 9.

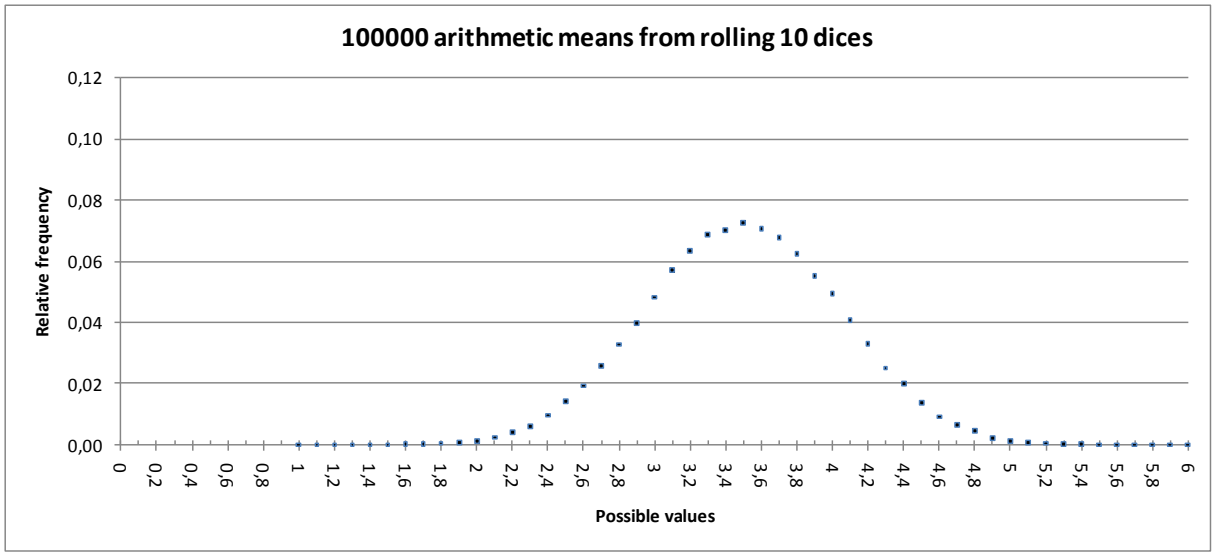

Fig. 19 xy-plot substituting bar graphs as in Fig. 8

Here each bar is represented by its upper end point.

To compare with normal distribution you must remember that relative frequencies have to be associated with area of rectangles. Here an interval such as 1.5 to 1.6 of lengthe 0.1 is used so we have to divide the relative interval frequency by 0.1 (or multiply by 10).

Then you can compare with normal distribution with mean 3.5 and standard deviation 1.7078/√10 as is seen in Fig. 20.

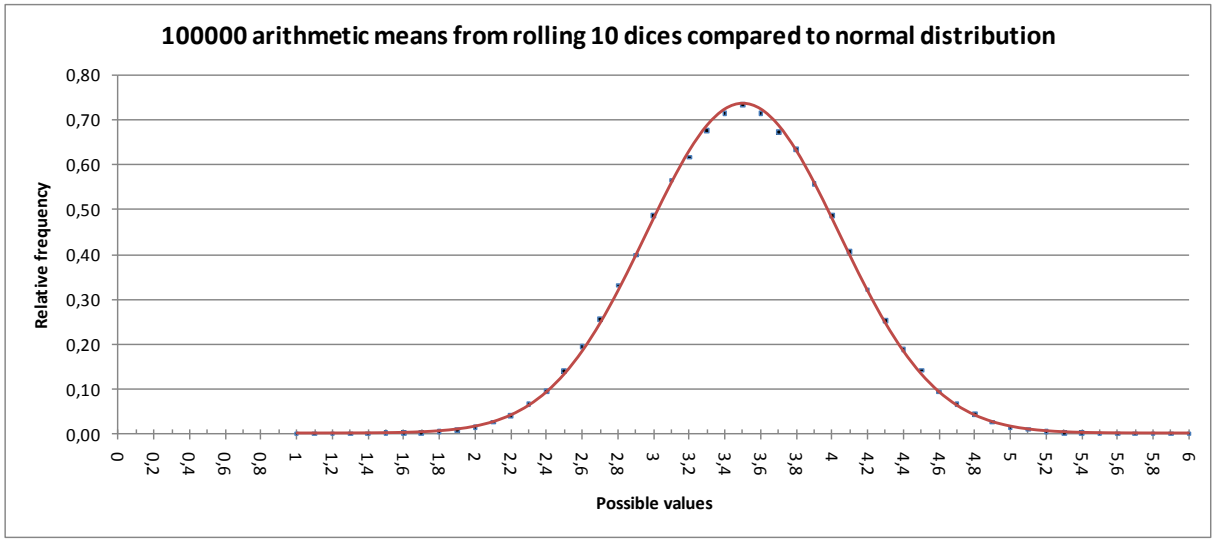

Fig. 20 "Dot histogram" compared with probability density function for normal distribution.

On the next figure 21 you see the cumulative probability and relative interval frequencies corresponding to Fig. 20.

Examining the pairs Fig. 20 & 22 and Fig. 21 & 23 shows how deviation for 100 dice averages are much smaller than for 10 dice. In fact they differ by a factor √10 = 3.16.

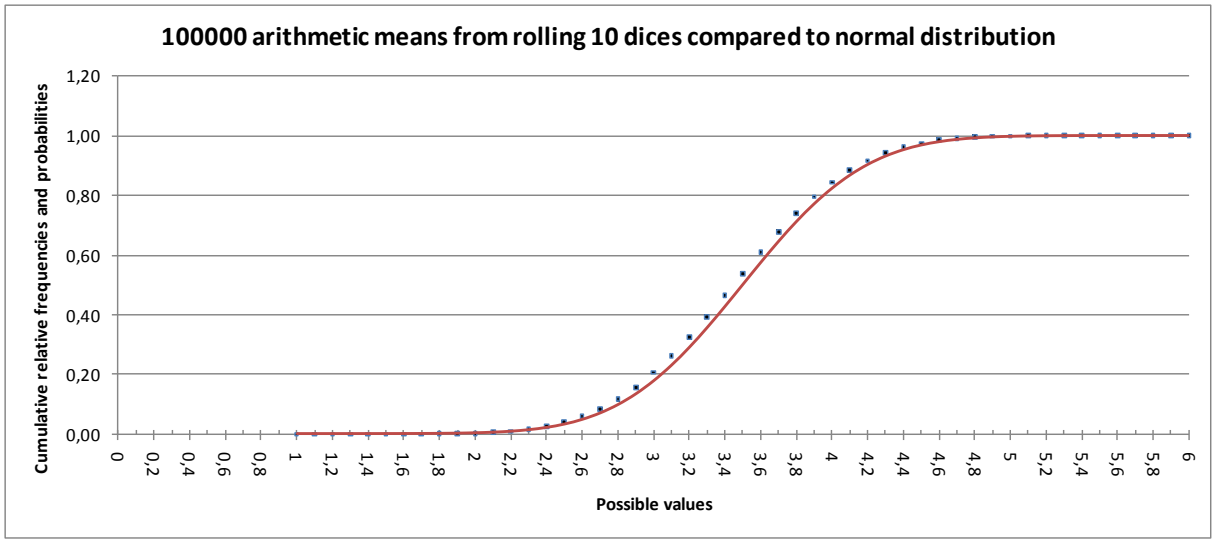

Fig. 21 Cumulative normal probabilities and cumulative relative interval frequencies.

Now comes the graphics for the rolls of 100 dices.

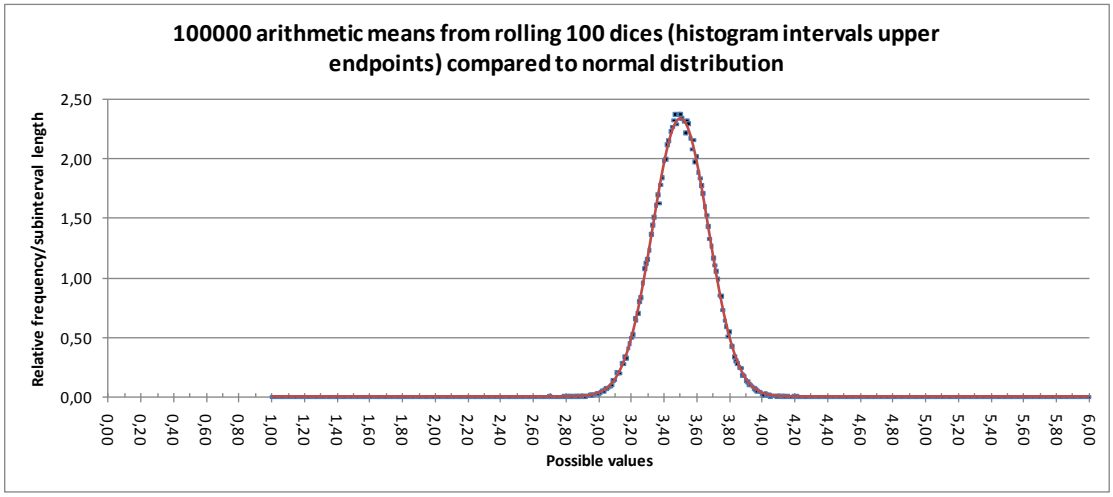

Fig. 22 "Dot histogram" corresponding to subdivision 1,00<1.01<1.02<....5.98<5.99<6.00 compared to normal distribution density with mean 3,5 and standarddeviation 1.7078/√100 = 0.17078

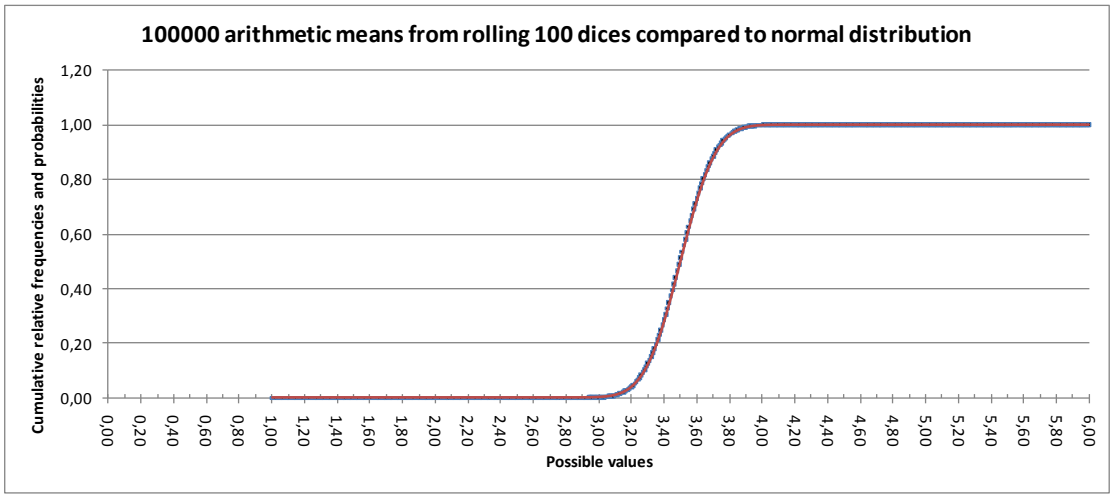

Fig. 23 Cumulative distributions corresponding to Fig. 22

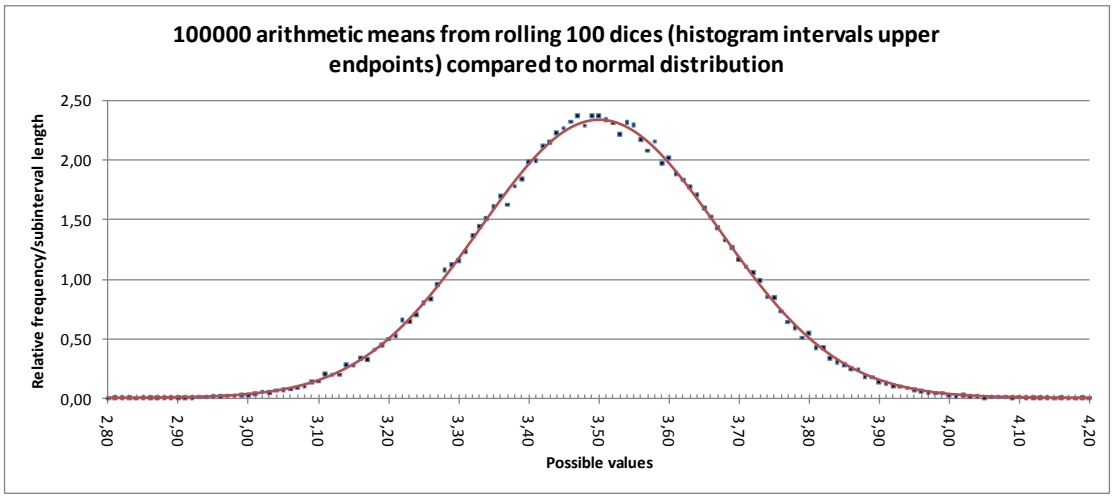

Fig. 24 Stretching part of Fig. 22 to study details.

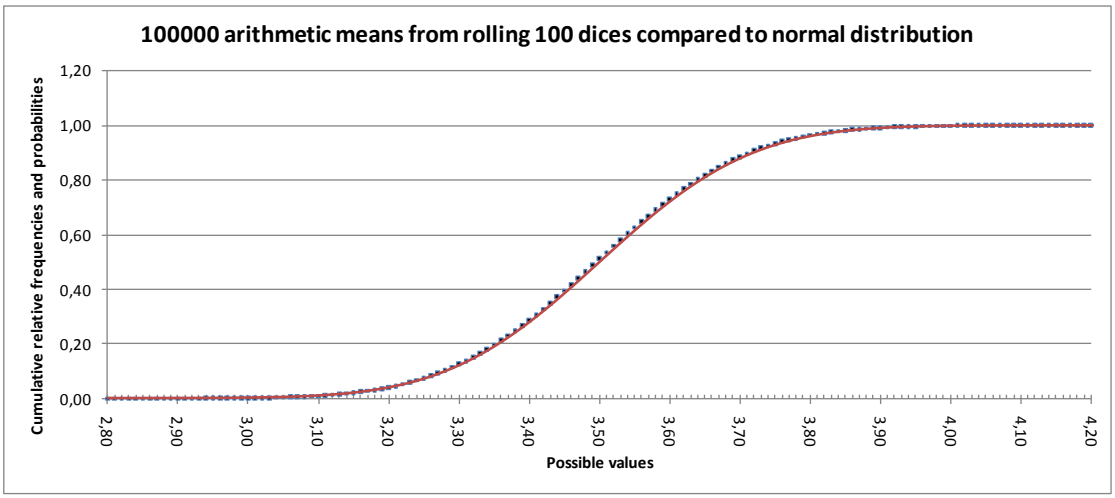

Fig. 25 Stretching part of Fig. 23 to study details.

Concluding remarks

In the previous sections are shown different parameters to vary when doing experiments of this kind.

- Varying the number of dices rolled.
- Varying the number of rolls of a fixed number of dices.
- Varying length of subdivisions for histogram.
- Translating subdivisions to examine effect of values coinciding with endpoints of subdivision intervals.

It can be rather cumbersome to write down a static description of the spreadsheets used. Some details are presented above – others are omitted. On the project webpage you will find example spreadsheets and videos with guidance to handling these problems.

References

[1] http://en.wikipedia.org/wiki/Central_limit_theorem (August 31, 2011)

[2] http://mathworld.wolfram.com/CentralLimitTheorem.html (August 31, 2011)