Simulation of χ^2 - distribution

John Andersen

Intro

One day I rolled 60 dice

Fig. 1 Photo of the 60 dice at my desktop

Rearranging to determine frequencies then gave this

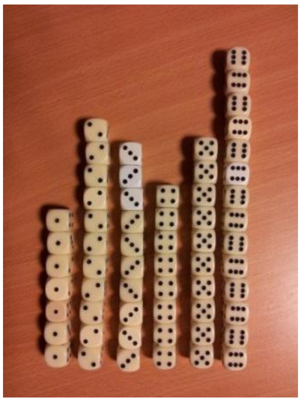


Fig. 2 One way to represent frequencies of 60 dices

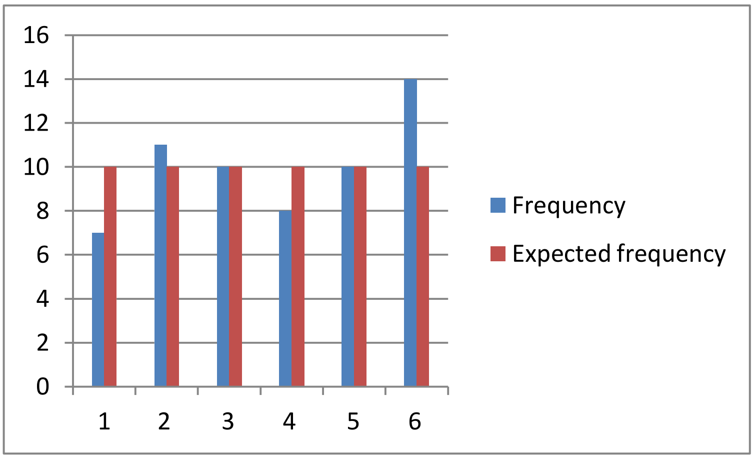


Fig. 3 Bar diagram over dice from Fig. 2

With the help of a spread sheet I simulated the roll of 6000 dice and got this picture of the outcome

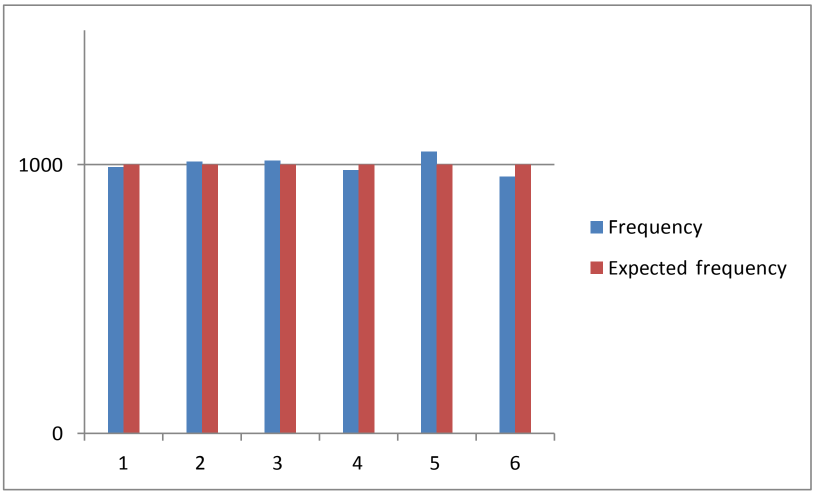


Fig. 4 Bar diagram representing a computersimulated roll of 60 dice

As you see the frequencies from the second experiment is much more in accordance with the expected frequencies when the number of dice is large. This is in accordance with the law of large numbers as described in [2].

But what if you have to decide from the situation in Fig. 1 whether the difference between actual outcome and expected outcome is too large? Since the graphical picture can vary a lot it is difficult to get a simple picture of the differences. People have developed ways to represent the variations by numbers.

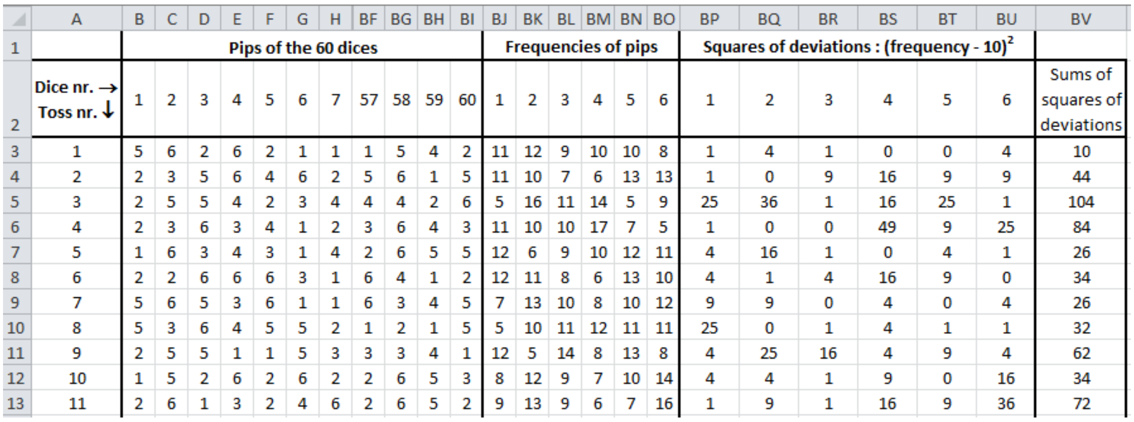
A „home made“ checking method

At this place one can imagine that someone e.g. inspired from least square.
On one hand if this is zero all 6 frequencies are equal to 10 which may be too good to be true (maybe manipulation with data?) and on the other hand if the number is too large the difference between actual frequencies and expected are too big so either you have a rare outcome or something is wrong with the expected frequencies. Some loading of the dice may have been performed.

It is not at all easy to see what can be expected of the number. To get a grip on the big picture we can use Excel to simulate a lot of tosses of 60 dice.

First of all: How to simulate this situation. How do I manage to make these data with Excel? The spread sheet is shown in Fig. 5 (resulting numbers) and Figs. 6-8 (formulas).

- Column A gives the number of each roll of the 60 dice
- Columns B:BI (a lot are hidden) gives the outcomes of the individual dice. Row 2 gives the number of the dice.
- Columns BJ:BO gives the frequencies of the 6 possible number of pips. (Possibilities shown in row 2).
- Columns BP:BU calculates the squared deviations.
- Column BV gives the sums of squared deviations.



	A	B	C	D	E	F	G	H	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV
1		Pips of the 60 dices											Frequencies of pips						Squares of deviations : (frequency - 10)^2						
2	Dice nr. → Toss nr. ↓	1	2	3	4	5	6	7	57	58	59	60	1	2	3	4	5	6	1	2	3	4	5	6	Sums of squares of deviations
3	1	5	6	2	6	2	1	1	1	5	4	2	11	12	9	10	10	8	1	4	1	0	0	4	10
4	2	2	3	5	6	4	6	2	5	6	1	5	11	10	7	6	13	13	1	0	9	16	9	9	44
5	3	2	5	5	4	2	3	4	4	4	2	6	5	16	11	14	5	9	25	36	1	16	25	1	104
6	4	2	3	6	3	4	1	2	3	6	4	3	11	10	10	17	7	5	1	0	0	49	9	25	84
7	5	1	6	3	4	3	1	4	2	6	5	5	12	6	9	10	12	11	4	16	1	0	4	1	26
8	6	2	2	6	6	6	3	1	6	4	1	2	12	11	8	6	13	10	4	1	4	16	9	0	34
9	7	5	6	5	3	6	1	1	6	3	4	5	7	13	10	8	10	12	9	9	0	4	0	4	26
10	8	5	3	6	4	5	5	2	1	2	1	5	5	10	11	12	11	11	25	0	1	4	1	1	32
11	9	2	5	5	1	1	5	3	3	3	4	1	12	5	14	8	13	8	4	25	16	4	9	4	62
12	10	1	5	2	6	2	6	2	2	6	5	3	8	12	9	7	10	14	4	4	1	9	0	16	34
13	11	2	6	1	3	2	4	6	2	6	5	2	9	13	9	6	7	16	1	9	1	16	9	36	72

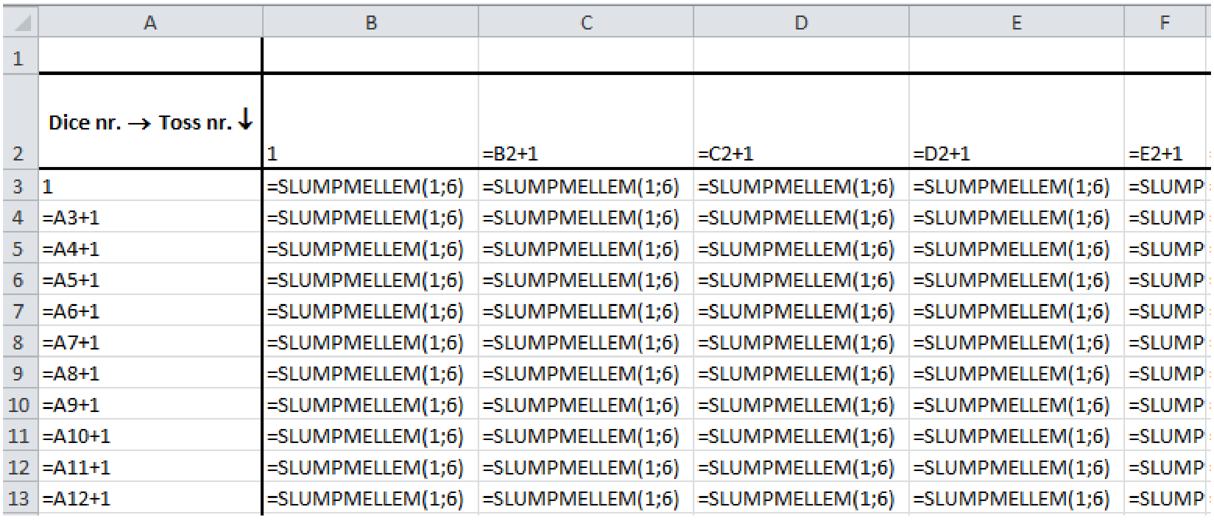
Fig. 5 Spread sheet for simulating a lot of rolls with 60 dice

Keep in mind that once you have set up formulas for one row it is a question of copying rows downwards to increase the number of simulated rolls. In the actual case I copied downwards until I got 20000 rolls.

There are 2^{20} rows to do with in my version of Excel. The working memory of my computer runs out long before I have filled in all possible rows. So if you want really big numbers of rolls you'll have to switch to more advanced programming tools. But one of my points in this connection is that I want to explore how far you can get by using rather elementary Excel skills.

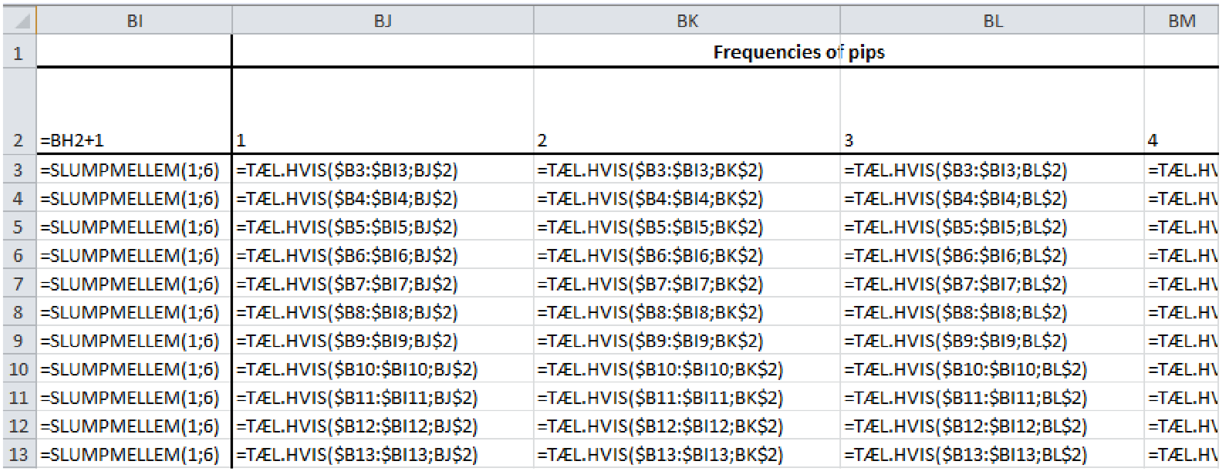
On the next 3 figures are shown the formulas behind the Fig 5.

I will not go into details since a careful study of the formulas and relations between them and the cell references may be the best way to build up an understanding. Please zoom in if the characters show up too small on your screen.



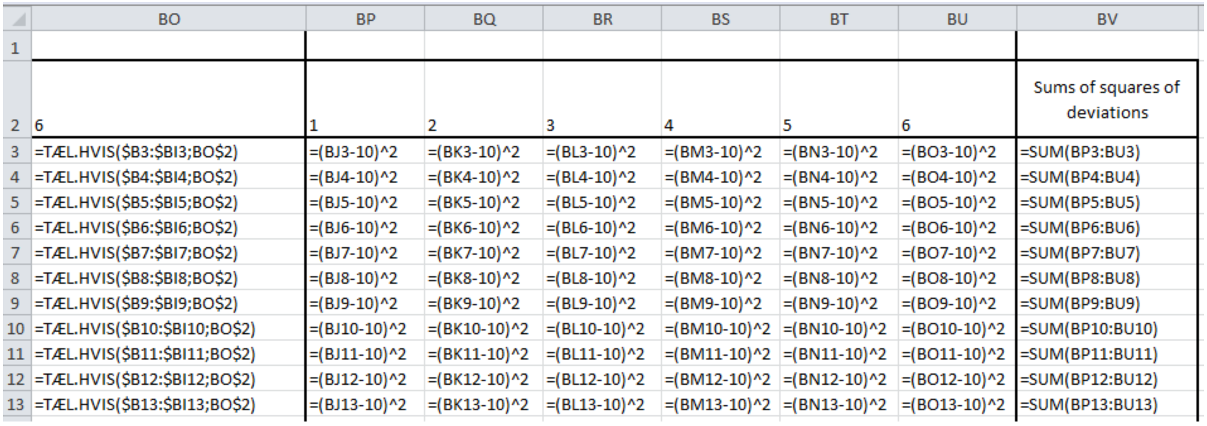
	A	B	C	D	E	F
1						
2	Dice nr. → Toss nr. ↓	1	=B2+1	=C2+1	=D2+1	=E2+1
3	1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
4	=A3+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
5	=A4+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
6	=A5+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
7	=A6+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
8	=A7+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
9	=A8+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
10	=A9+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
11	=A10+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
12	=A11+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP
13	=A12+1	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMPMELLEM(1;6)	=SLUMP

Fig. 6 Formulas for Fig 5 - part 1.



	BI	BJ	BK	BL	BM
1			Frequencies of pips		
2	=BH2+1	1	2	3	4
3	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B3:$BI3;BJ$2)	=TÆL.HVIS($B3:$BI3;BK$2)	=TÆL.HVIS($B3:$BI3;BL$2)	=TÆL.HV
4	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B4:$BI4;BJ$2)	=TÆL.HVIS($B4:$BI4;BK$2)	=TÆL.HVIS($B4:$BI4;BL$2)	=TÆL.HV
5	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B5:$BI5;BJ$2)	=TÆL.HVIS($B5:$BI5;BK$2)	=TÆL.HVIS($B5:$BI5;BL$2)	=TÆL.HV
6	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B6:$BI6;BJ$2)	=TÆL.HVIS($B6:$BI6;BK$2)	=TÆL.HVIS($B6:$BI6;BL$2)	=TÆL.HV
7	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B7:$BI7;BJ$2)	=TÆL.HVIS($B7:$BI7;BK$2)	=TÆL.HVIS($B7:$BI7;BL$2)	=TÆL.HV
8	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B8:$BI8;BJ$2)	=TÆL.HVIS($B8:$BI8;BK$2)	=TÆL.HVIS($B8:$BI8;BL$2)	=TÆL.HV
9	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B9:$BI9;BJ$2)	=TÆL.HVIS($B9:$BI9;BK$2)	=TÆL.HVIS($B9:$BI9;BL$2)	=TÆL.HV
10	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B10:$BI10;BJ$2)	=TÆL.HVIS($B10:$BI10;BK$2)	=TÆL.HVIS($B10:$BI10;BL$2)	=TÆL.HV
11	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B11:$BI11;BJ$2)	=TÆL.HVIS($B11:$BI11;BK$2)	=TÆL.HVIS($B11:$BI11;BL$2)	=TÆL.HV
12	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B12:$BI12;BJ$2)	=TÆL.HVIS($B12:$BI12;BK$2)	=TÆL.HVIS($B12:$BI12;BL$2)	=TÆL.HV
13	=SLUMPMELLEM(1;6)	=TÆL.HVIS($B13:$BI13;BJ$2)	=TÆL.HVIS($B13:$BI13;BK$2)	=TÆL.HVIS($B13:$BI13;BL$2)	=TÆL.HV

Fig. 7 Formulas for Fig 5 - part 2.



	BO	BP	BQ	BR	BS	BT	BU	BV
1								
2	6	1	2	3	4	5	6	Sums of squares of deviations
3	=TÆL.HVIS($B3:$BI3;BO$2)	=(BJ3-10)^2	=(BK3-10)^2	=(BL3-10)^2	=(BM3-10)^2	=(BN3-10)^2	=(BO3-10)^2	=SUM(BP3:BU3)
4	=TÆL.HVIS($B4:$BI4;BO$2)	=(BJ4-10)^2	=(BK4-10)^2	=(BL4-10)^2	=(BM4-10)^2	=(BN4-10)^2	=(BO4-10)^2	=SUM(BP4:BU4)
5	=TÆL.HVIS($B5:$BI5;BO$2)	=(BJ5-10)^2	=(BK5-10)^2	=(BL5-10)^2	=(BM5-10)^2	=(BN5-10)^2	=(BO5-10)^2	=SUM(BP5:BU5)
6	=TÆL.HVIS($B6:$BI6;BO$2)	=(BJ6-10)^2	=(BK6-10)^2	=(BL6-10)^2	=(BM6-10)^2	=(BN6-10)^2	=(BO6-10)^2	=SUM(BP6:BU6)
7	=TÆL.HVIS($B7:$BI7;BO$2)	=(BJ7-10)^2	=(BK7-10)^2	=(BL7-10)^2	=(BM7-10)^2	=(BN7-10)^2	=(BO7-10)^2	=SUM(BP7:BU7)
8	=TÆL.HVIS($B8:$BI8;BO$2)	=(BJ8-10)^2	=(BK8-10)^2	=(BL8-10)^2	=(BM8-10)^2	=(BN8-10)^2	=(BO8-10)^2	=SUM(BP8:BU8)
9	=TÆL.HVIS($B9:$BI9;BO$2)	=(BJ9-10)^2	=(BK9-10)^2	=(BL9-10)^2	=(BM9-10)^2	=(BN9-10)^2	=(BO9-10)^2	=SUM(BP9:BU9)
10	=TÆL.HVIS($B10:$BI10;BO$2)	=(BJ10-10)^2	=(BK10-10)^2	=(BL10-10)^2	=(BM10-10)^2	=(BN10-10)^2	=(BO10-10)^2	=SUM(BP10:BU10)
11	=TÆL.HVIS($B11:$BI11;BO$2)	=(BJ11-10)^2	=(BK11-10)^2	=(BL11-10)^2	=(BM11-10)^2	=(BN11-10)^2	=(BO11-10)^2	=SUM(BP11:BU11)
12	=TÆL.HVIS($B12:$BI12;BO$2)	=(BJ12-10)^2	=(BK12-10)^2	=(BL12-10)^2	=(BM12-10)^2	=(BN12-10)^2	=(BO12-10)^2	=SUM(BP12:BU12)
13	=TÆL.HVIS($B13:$BI13;BO$2)	=(BJ13-10)^2	=(BK13-10)^2	=(BL13-10)^2	=(BM13-10)^2	=(BN13-10)^2	=(BO13-10)^2	=SUM(BP13:BU13)

Fig. 8 Formulas for Fig 5 - part 3.

The 20000 sums of squares of deviations from column BV (Se Fig 5) are now the data I want to get an overview of. Therefore by traditional statistical methods (and Excel) I draw histograms and curves over accumulated relative interval frequencies. Resulting diagrams are shown in Figs. 9-10.

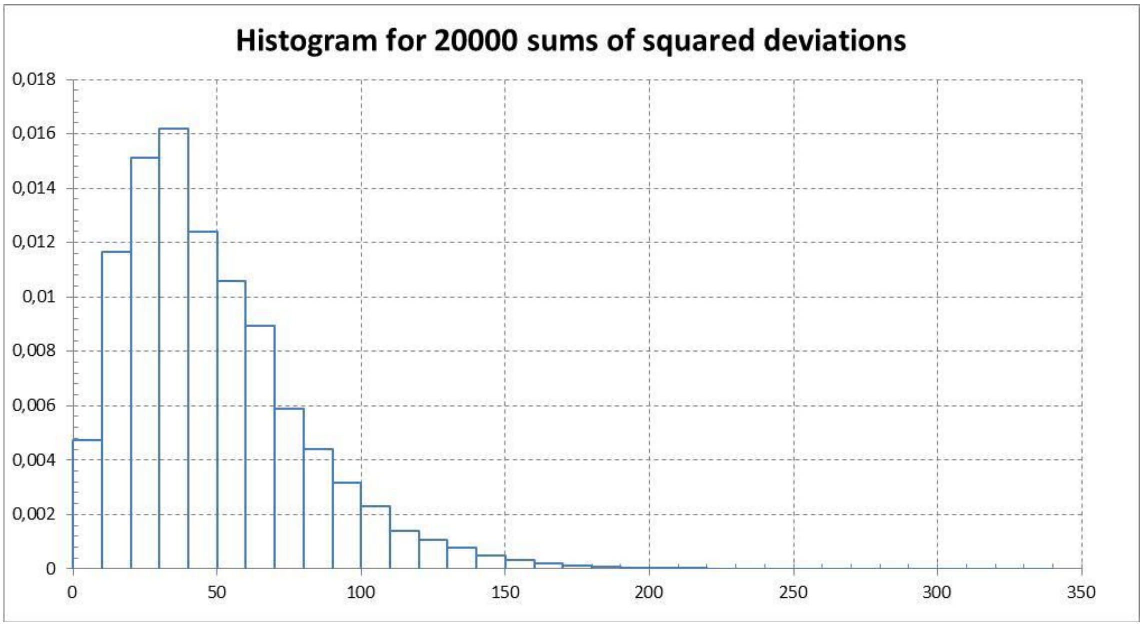


Fig. 9

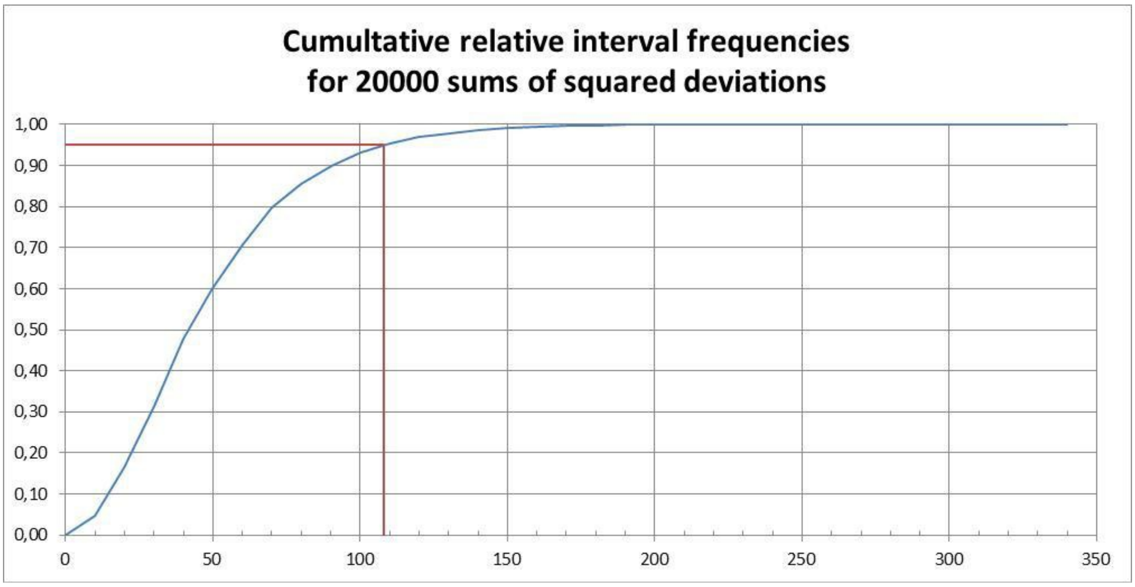


Fig. 10

One sees that approximately 95% of the sums are less than 110 (108 actually seen by trial and error in the spread sheet). So only in 5% of cases one will expect sums greater than 108.

What about the dice in Fig. 1? Calculating the sum of squared deviations gives the number 30 which seems to be a very typical value as may be seen from Fig. 9. So nothing is alarming about that roll.

Notice that in all the preceding the only statistical theory used is grouping observations and making histogram and cumulative relative frequencies diagram combined with “brute force” e.g. managing 20000 simulations of tossing 60 dice. You need to watch your steps in doing so but it is more a question of endurance than of sophisticated mathematical theories.

A standard χ^2 - test

Problems of this kind have been studied long before the birth of computers and theories have been developed. They are usually handled by using so called χ^2 - test

One calculates the sum (division by 10 is due to some sort of normalization) which can be shown to be approximately distributed as a χ^2 - distribution with 5 degrees of freedom.

Looking up in a table for this distribution

Cumulative probability	Chi square
0,99	15,09
0,98	13,39
0,97	12,37
0,96	11,64
0,95	11,07
0,94	10,60
0,93	10,19
0,92	9,84

Table 1

you see that 95 % of such a distribution should not exceed 11,07.

So the result 108 obtained from the simulation is OK (remember that a division by 10 has to be performed so it is 10,8 ≈ 11) that has to be compared with 11,07.

On the next Figs. 11-12 I compare simulated data with χ^2 - distribution. The spread sheet for simulating the data for this is obtained from the one in Fig. 5 by dividing the formulas in column BV with 10 so I will not repeat screen shots of the spread sheet.

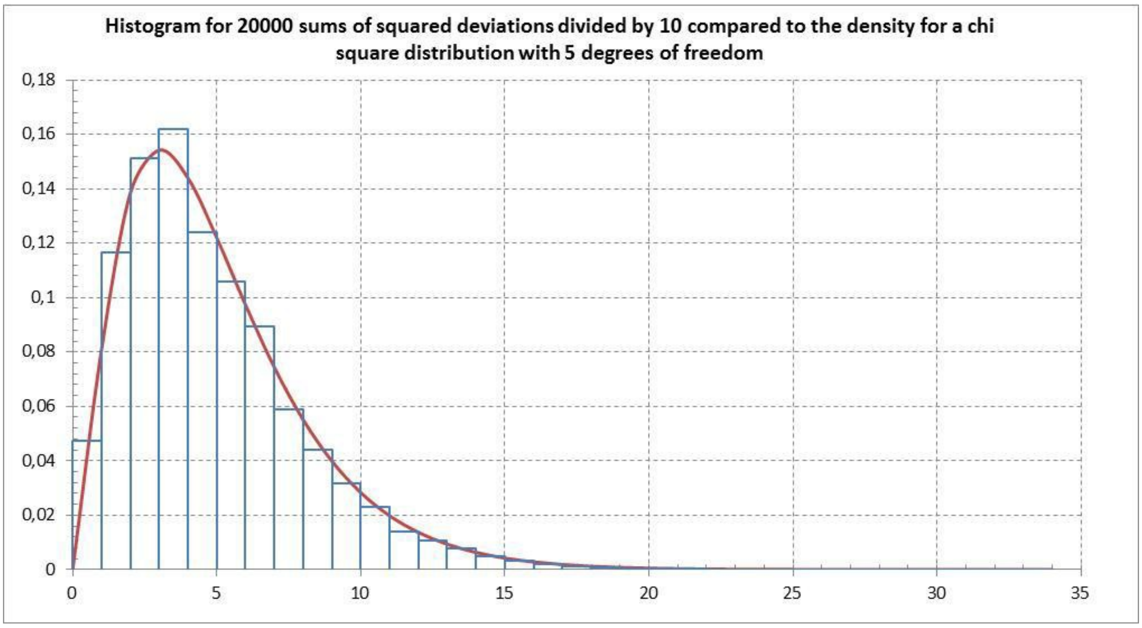


Fig. 11

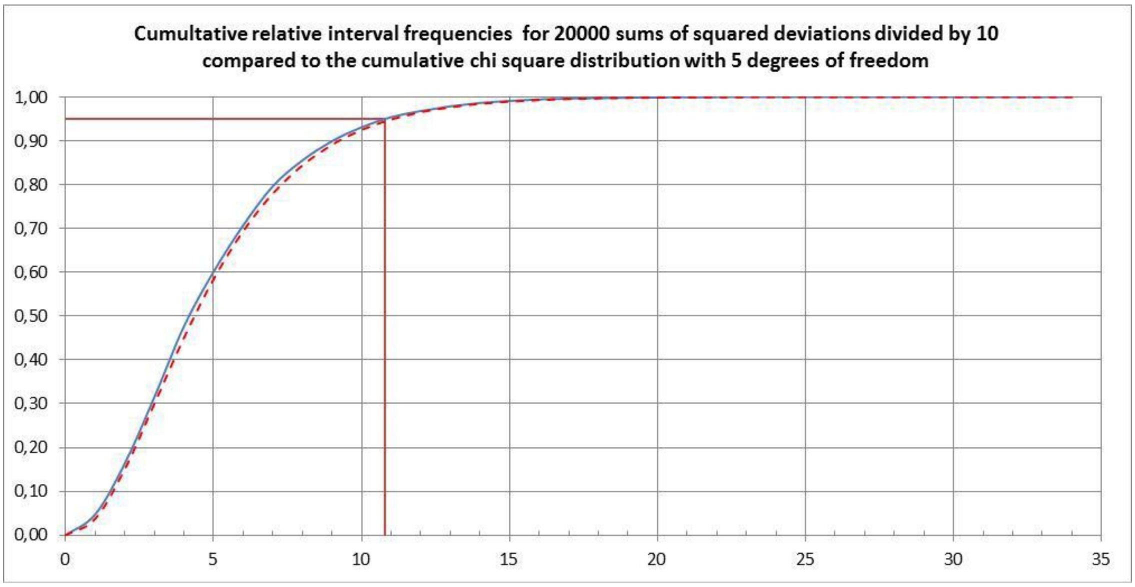


Fig. 12

Checking independence of to categories

Let’s examine a more complicated situation by the same means.

Suppose you have a country with four political parties P, Q, R and S. From last election you know they have 40, 30, 20 and 10 % of votes. You want to check whether people’s incomes influence which party they give their vote or not. So you categorises peoples income in three categories IC1, IC2 and IC3. Standard tests have been developed for this situation. Here the problem is approached by simulating a lot of interview series to get an idea of how the χ^2 - test statistic is distributed. You find the final graphics in Figs. 23-24.

If you have interviewed 600 voters you can order the result in a table shown below

	P	**Q**	**R**	**S**
IC1	110	96	53	23
IC2	70	60	41	22
IC3	42	41	18	24

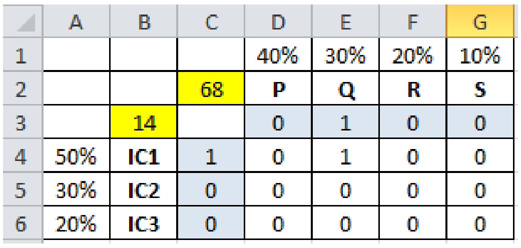
Table 1

This situation resembles the situation in Fig 3 where you have to decide from one set of data something is the case. In the situation with the 60 dice it was if an expected distribution is acceptable as a model, in the situation with voters whether the two criteria can be considered independent or not.

The idea now is similar to the simulation of the 20000 rolls with 60 dice to get an idea of how tables may come out if the criteria really are independent. So the task is to create a lot of simulations of tables like the one above in a way that assures independence of the two criteria. This we can assure by using formulas.

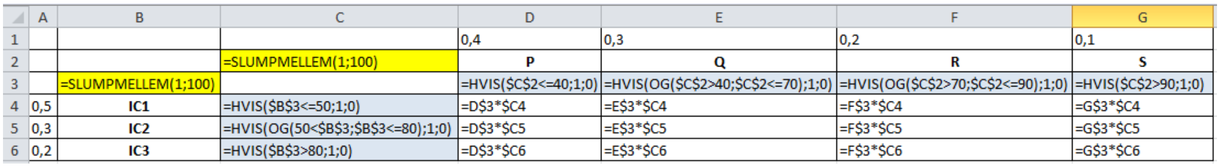
The job is a somewhat more complicated than the dice situation to handle but nevertheless it can be done as is shown in the following screen shots from Excel. Video demonstrations of the actual steps will be uploaded to the project homepage.

First step is to create a spread sheet that picks where to put each new observation one at a time.



	A	B	C	D	E	F	G
1				40%	30%	20%	10%
2			68	**P**	**Q**	**R**	**S**
3		14		0	1	0	0
4	50%	**IC1**	1	0	1	0	0
5	30%	**IC2**	0	0	0	0	0
6	20%	**IC3**	0	0	0	0	0

Fig. 13 Placing one voter in the scheme with independence between criterias.



	A	B	C	D	E	F	G
1				0,4	0,3	0,2	0,1
2			=SLUMPMELLEM(1;100)	**P**	**Q**	**R**	**S**
3		=SLUMPMELLEM(1;100)		=HVIS(C2<=40;1;0)	=HVIS(OG(C2>40;C2<=70);1;0)	=HVIS(OG(C2>70;C2<=90);1;0)	=HVIS(C2>90;1;0)
4	0,5	**IC1**	=HVIS(B3<=50;1;0)	=D$3*$C4	=E$3*$C4	=F$3*$C4	=G$3*$C4
5	0,3	**IC2**	=HVIS(OG(50<B3;B3<=80);1;0)	=D$3*$C5	=E$3*$C5	=F$3*$C5	=G$3*$C5
6	0,2	**IC3**	=HVIS(B3>80;1;0)	=D$3*$C6	=E$3*$C6	=F$3*$C6	=G$3*$C6

Fig. 14 Formulas for Fig. 13

Now we have to make it possible to do a lot of “interviews” and sum up the results..