practical statistics

without mathematics

J.M.M.J. Vogels

preface

To whom is this article intended? To people in practice. The challenge for the author was: is it possible to write a piece about statistics, without quite a lot of mathematics? Often statistics compels the student to several years of mathematical study. No doubt this is necessary if he wants to get access to scientific literature. For work in research, the mathematics is inevitable.

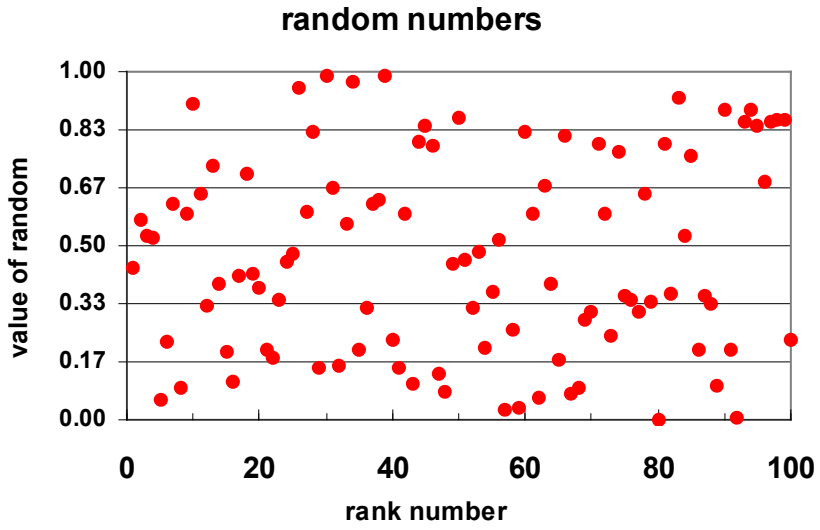
But may someone in practice, like stock management, sales or teaching at schools, learn the main principles in an easier way? He will not be interested in mathematical proofs, in correlation coefficients, least squares adaptations and so on. If he may learn the basics by experimenting on his own computer, he has a quick access to what he needs. Statistics by experiments is the content of this article.

No one wants to buy a pig in a poke however. With some mathematical notes, outside the main text, the content may be checked or proven. Don't worry. The outcome of the experiments can be trusted, as the people from mathematics can follow what we claim.

playing with dice

Dice are common statistical tools. Casting dice shows 1, 2, 3, 4, 5 or 6 dots. Every number has a probability 1/6 . How is probability defined? It is the number of hits, divided by the total number of attempts, when one tries it infinitely many times. When casting 6 times, usually not every outcome will occur once. But casting 10000 times gives nearly equal numbers of hits for each number of dots.

In a spreadsheet a similar tool is present, the procedure RAND, by which one obtains some number between 0 and 1 , with equal probability for each number. The procedure may be seen as a kind of roulette-table. In the diagram we see 100 random numbers. Another run would give a fully other picture. There is no system in the position of the numbers.



At the left we have a division into 6 equal intervals between 0 and 1 . In the little table we give the number of counts in each interval.

dots	1	2	3	4	5	6
counts	17	20	18	16	14	15

experiment 1
This procedure simulates a common die, that is used 100 times. We advise the reader to do the experiment himself. In this booklet the experiments are essential. For a correct idea one should do them. We explain some spreadsheet functions:

IF(A1<1/6,1,0)	gives 1 when A1 < 1/6 , else 0
AND(A1>=3/6,A1<4/6)	is true if 3/6 ≤ A1 < 4/6
IF(AND(A1>=3/6,A1<4/6),1,0)	gives 1 when true, else 0
SUM(A1:A1000)	gives A1 + A2 + … + A1000

Note that the procedure RAND() may give every now and then another random number. One may press “delete” on an empty cell to obtain a new value.

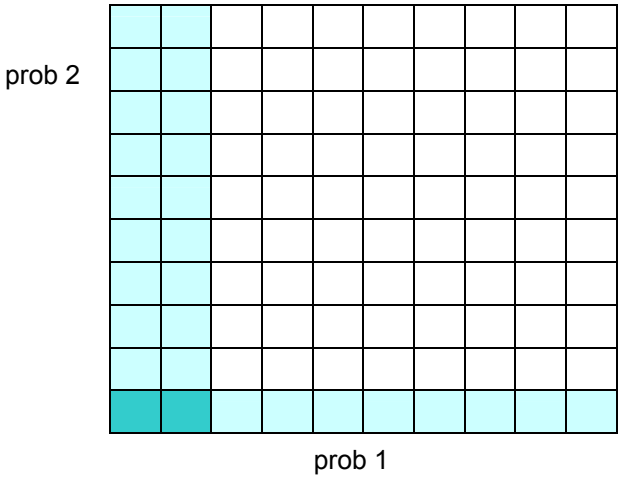
Set in the first column 100 random numbers. In the second column one sets (IF(A1<1/6,1,0), in the third column IF(AND(A1>=1/6,A1<2/6),1,0) , and so on. Below the sorting columns one sets the number of hits in each interval by SUM(B1:B100) .

Now that we have a digital die, we may see what happens if we sum the numbers of dots with two dice together.

6	7	8	9	10	11	12
5	6	7	8	9	10	11
4	5	6	7	8	9	10
3	4	5	6	7	8	9
2	3	4	5	6	7	8
1	2	3	4	5	6	7
	1	**2**	**3**	**4**	**5**	**6**

The sum with two dice. The probability of a sum 7 is 6/36 = 0.167 = 16.7% . The middle values occur more frequently than the extreme ones.

What can we say about two incidents, each with its own probability? We suppose that the two are not correlated. This means that they occur fully independent of each other. What is the probability that they occur together?



Two probabilities, the one 2/10 and the other 1/10 , without any correlation. The two incidents may occur together, with a probability 2/100 = .02 or 2% .

The probability of both incidents together is the product of the separate probabilities.

An example of that principle. A yachtsman on the sea needs a radio for his safety. Suppose that a radio gets defect once in 1000 sailing-trips. With one reserve radio the risk is reduced to once in a million.

how accurate is an average? packages of raisins

A trader sells raisins in packages of 200 g . Every package has an uncertainty in its mass of about 5 g , because nearly all packages have a mass between 197.5 and 202.5 g . But he is obliged to sell an average mass of 200 g . He weighs a large number of packages and calculates the average. How accurate is the average value? We simulate this problem on the computer. The unit of mass will be 5 g , so that the spread in mass has a magnitude of 1 .

With the spreadsheet procedure RAND we obtain random numbers between 0 and 1. If we repeat this many times, the average value for the whole multitude will be nearly 0.5 . We calculate a noise number r by

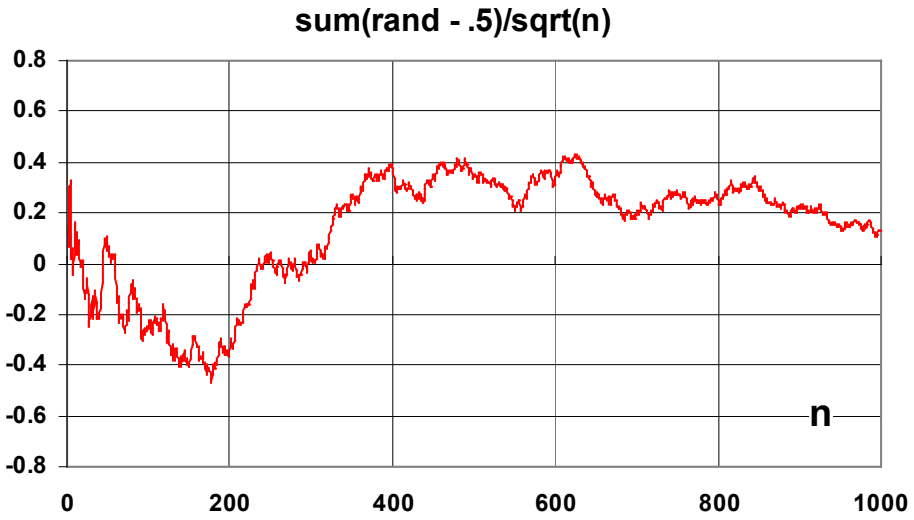
r = RAND - 0.5 .

The average value of the noise in infinitely many tests will be 0 , or

< r > = 0 .

If we calculate r many times, the < r > will be nearly 0 . What is the magnitude of the noise in the average?

experiment 2
In a spreadsheet we set in the first column the rank numbers n = 1, 2, 3, … In the second column we set 1000 different noise numbers. In a third column we set at position n the sum over the first n noise numbers. In a fourth column we set the sum, divided by √n . The square root is given by SQRT(n) . We draw a graph of the fourth column for all 1000 rank numbers. We see that the sum/√n retains a value in or around the interval between -0.5 and +0.5 . The spread is nearly 1 .



Then it is clear that **the average, which is the sum/n , will spread in an interval with magnitude 1/√n** .

This is a very useful rule. There are many examples where the rule may be applied.

We consider a second example. A school class makes a test. Each child has an uncertainty in its marks. The teacher calculates the average over a whole group of 30. The uncertainty in the average will be $1/\sqrt{30} \approx 0.2$ times the individual uncertainty.

[see note 1]

a reasonable stock fluctuations in sales figures

A shop sells towels. Once a week a truck arrives with new supply. It is known what the average figure of sales is over one week. How many towels should the shop take in stock? Of course the average sale must be in stock, but also some margin.

We suppose that very many people visit the shop, but that only some of them buy a towel. This assumption of a very large number of opportunities, but a small number of actual hits, constitutes the so called Poisson statistics. The large number of opportunities is essential!

experiment 3
Suppose that there are weekly 1000 visitors. Only 2% of them buy one towel. The shop sells 20 towels a week, on an average. For a first idea we make a simulation in a spreadsheet. In the first column we set 1000 random numbers between 0 and 1 . In the second column we set 0 or 1 : if the random number is smaller than .02 , we set 1 in this column, else 0 . We may expect about 20 hits. In some cell we set the sum over the second column, which is the number of hits. Also a threshold value may be given with an indicator: if the number of hits exceeds a threshold value, we print "*****", else nothing or "". The asterisks give a quick view. Let the spreadsheet run (press delete on an empty cell). The frequency by which the number of hits exceeds the threshold value is easy to see. A threshold number of 25 will frequently be exceeded. A number of 30 hits or more is rare. More than 35 hits will be very rare. The shop takes weekly between 25 and 35 towels in stock, so the margin above the average of 20 will be between 5 and 15.

The outcome of the experiment does not depend strongly on the number of opportunities, here 1000, provided that this number is many times the number of hits. Only the average number must fit exactly.

We would present statistics without mathematics. It is fair to tell however that we present some tables on this stuff, not by many tests, but from the formula for Poisson statistics.

average number	above 95%	above 99%
1	3	4
2	5	6
3	6	8
5	9	11
10	15	18
20	28	31
30	39	43
50	62	67
100	117	124

The first column gives the average number of hits. In the second column we see how large the stock should be, if we want to have enough in 95% of all weeks. The third column gives the required stock to have enough in 99% of all weeks. The stock is rounded to the lowest integer number, that is at least required.

In such a table the margin for the stock is easily seen. But there is another way to look at the margins. We denote the average number of hits as a. We take 3 values of the margin. Then in the green area we denote for how many percent of the weeks the stock will be sufficient.

average a	margin √a	margin 2√a	margin 3√a
1	91	98	99.6
2	94	98	99.8
3	91	98	99.8
5	93	98	99.7
10	91	98	99.8
20	88	97	99.8
30	88	97	99.8
50	88	98	99.8
100	85	97	99.8

The percentage of all weeks that the margin will be sufficient.

margins rounded up

percentages rounded down

Also in this statistical problem an uncertainty or margin is involved that is related to a square root. We investigate this relation nearer in a third table. There we consider only an average a = 10 . But in the second table we saw already that an expression for a safe margin does not depend critically on the average a .

average a=10	percentage safe
margin 0.0	58
0.5 √a	79
1.0 √a	91
1.5 √a	95
2.0 √a	98
2.5 √a	99
3.0 √a	99.8
3.5 √a	99.97
4.0 √a	99.99

For a = 10 we accept in the left column several margins. In the right column we see the percentage of all weeks in which the margin is sufficient.

margins rounded up

percentages rounded down

For each level of safety a reasonable margin can be seen in the tables.

The same kind of statistics will be valid for other problems, where a limited number of hits is present, at a very large number of opportunities. How many cars are offered daily in a garage for repair? How many people visit an office every day? How many cars pass by through a street in one hour? If the possible number is very large but the actual number is moderate, all these problems behave according to Poisson statistics.

extrapolation to large numbers

Let us estimate a safe margin when a = 1000. We take a sum over 10 cases with a = 100. In the problem of raisin packages we saw that the noise interval must increase by a factor $\sqrt{10}$. When a=100 a 95% safety requires a margin of 17 according to the first table. We expect a comparable safety at a=1000 with a margin $17 * \sqrt{10} \approx 54$. More generally: let a noise interval around an average have a typical width of $\pm \sqrt{a}$, then a noise interval around an average k*a should have a typical width of $\pm \sqrt{(k*a)}$. In the second table we see furthermore that the formula for a safe margin is nearly equal over a large range of a.

[see note 2]

experiments with squares

Suppose that we have a noise number r between -.5 and +.5 . Then for many cases together we expect an average <r> = 0. But what may be the average of r^2 ?

experiment 4
In a column on a spreadsheet we set 10000 noise numbers r (= RAND - .5). We sum them and divide the sum by 10000. This average contains only noise, with a typical magnitude below .01 .
Then we set the squares of the noise numbers in a second column. We sum them and divide the sum by 10000. We multiply the average of r^2 by 12. We see that $< r^2 > = 1/12$. The average of the squares of noise may have a value that does not go to 0 as the number of cases goes to infinity.

[see note 3]

experiment 5
In a spreadsheet we set noise numbers r_1 in the first column and r_2 in the second column. In the third column we set $r_1 + r_2$. In the fourth column we set the squares of (r_1+r_2) . The average over the fourth column has a value of 1/6 .

Note that $(r_1+r_2)^2 = r_1^2 + r_2^2 + 2\, r_1\, r_2$. The mixed term $2\, r_1\, r_2$ has the character of noise, with an average of 0 for infinitely many cases. So for the sum of squares of noise we find the property

$< (r_1+r_2)^2 > = < r_1^2 > + < r_2^2 >$.

experiment 6
In a spreadsheet cell we sum 12 noise numbers (RAND-.5). All values in the first column are the sum of 12 noise numbers. In the second column we denote the squares of the sums of the first column. We extend the columns to 10000 cases. The second column is rather noisy. Finally we calculate the average over the second column. The average is 1, apart from some remaining noise with a magnitude of typically .01 .

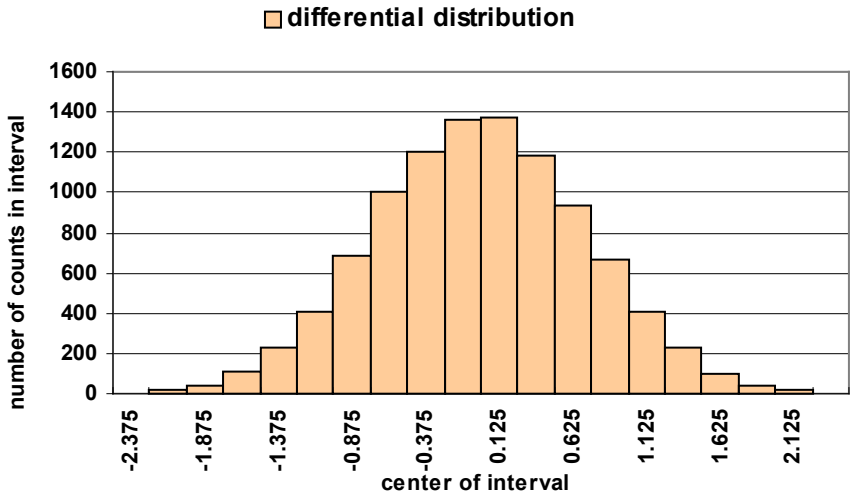
When the sum of 12 noise numbers is squared, there are 66 mixed terms. The mixed terms will have magnitudes below 1/2 and the sum will have a magnitude of typically $(1/2)*\sqrt{66} \approx 4$. Averaging noise numbers with these magnitudes over 10000 cases leads to a resulting noise of lower than .04 . In fact this estimate is still too large, as a resulting noise will seldom exceed the value of .01 .

At last we construct with 12 noise numbers a series of values x with $< x^2 > = ½$. This is easily done if we divide the sum by $\sqrt{2}$. So first A = rand() + rand() + …… [12 terms] - 6 and next $x = A / \sqrt{2}$. Then automatically $< x^2 > = ½$. A scale transformation is easy this way. But the number of noise contributions used remains 12.

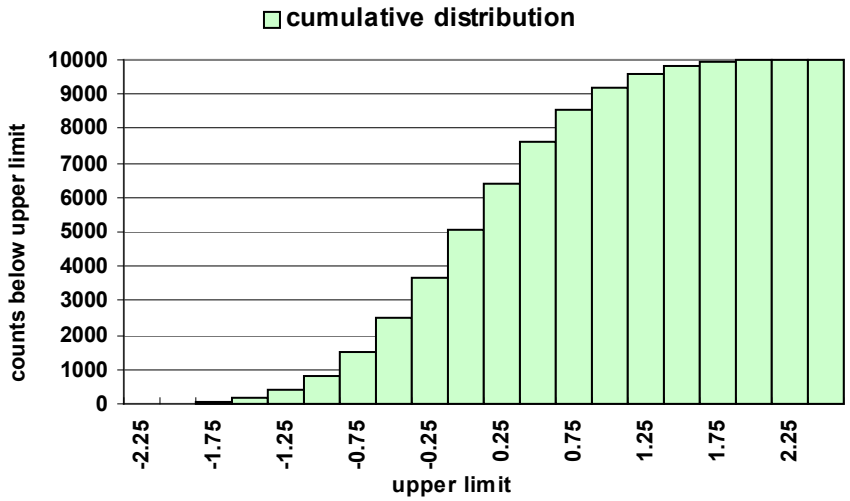
bell shaped distributions

Now that we have a method to construct numbers x with $< x^2 > = ½$, it is a question how these numbers are distributed. In the table with the sum over two dice we saw that a sum in the middle is more likely than an extreme value. When we sum 12 noise numbers, we may expect a similar result. We test it.

experiment 7
In the first cell of a spreadsheet column we set the (sum of 12 RAND terms) - 6 . The same is done in the cells downward. In the second column we set the values of the first column, divided by √2 . In the third and following columns we sort the results of the second column in intervals, just as we did in experiment 1 . We sort in a range between -2.5 and +2.5 with intervals of 0.25 . The first interval gets a statement IF(AND (x>=-2.5,x<-2.25),1,0) . So in this cell there appears 1 if the number x is in the right range, else 0 . A similar procedure is followed for the other intervals. We calculate 10000 values in each column. Finally we calculate the sum in each sorting column on a line below rank number 10000. With these sums we make a diagram. On the y scale we see the number of hits in each interval. The diagram is the differential distribution.



In a second diagram we give all counts below some threshold value. It is called a cumulative distribution.



One may experiment with the scale factor on the x-axis. Then one replaces in the second column of the spreadsheet the √2 by some other value.

We conclude that indeed the most results are concentrated around 0. The differential distribution is also called a bell shaped distribution or a Gaussian distribution, referring to the mathematician C. F. Gauss (1777-1855).

The cumulative distribution follows an error function, as we will see.

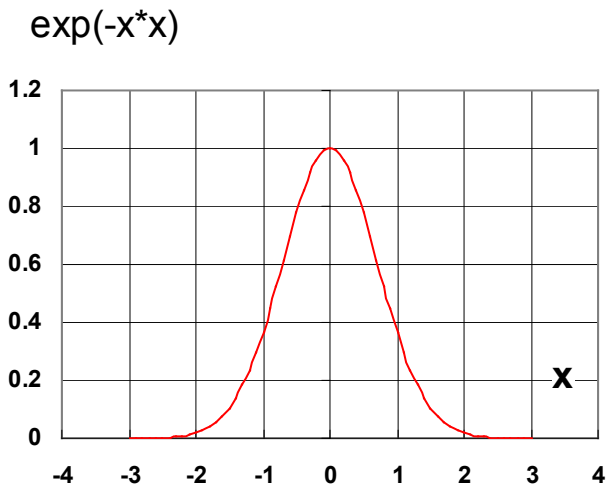
In the graphs one may see a general principle. Let some quantity have a noise that is the sum over a dozen of noise causes. Then the resulting noise tends to a Gaussian distribution. This is an often observed situation in statistics.

[see note 4]

practical use of the bell curve a schooltest

A Gaussian distribution is frequently seen in statistics. How to handle it in practice?

[see note 5]

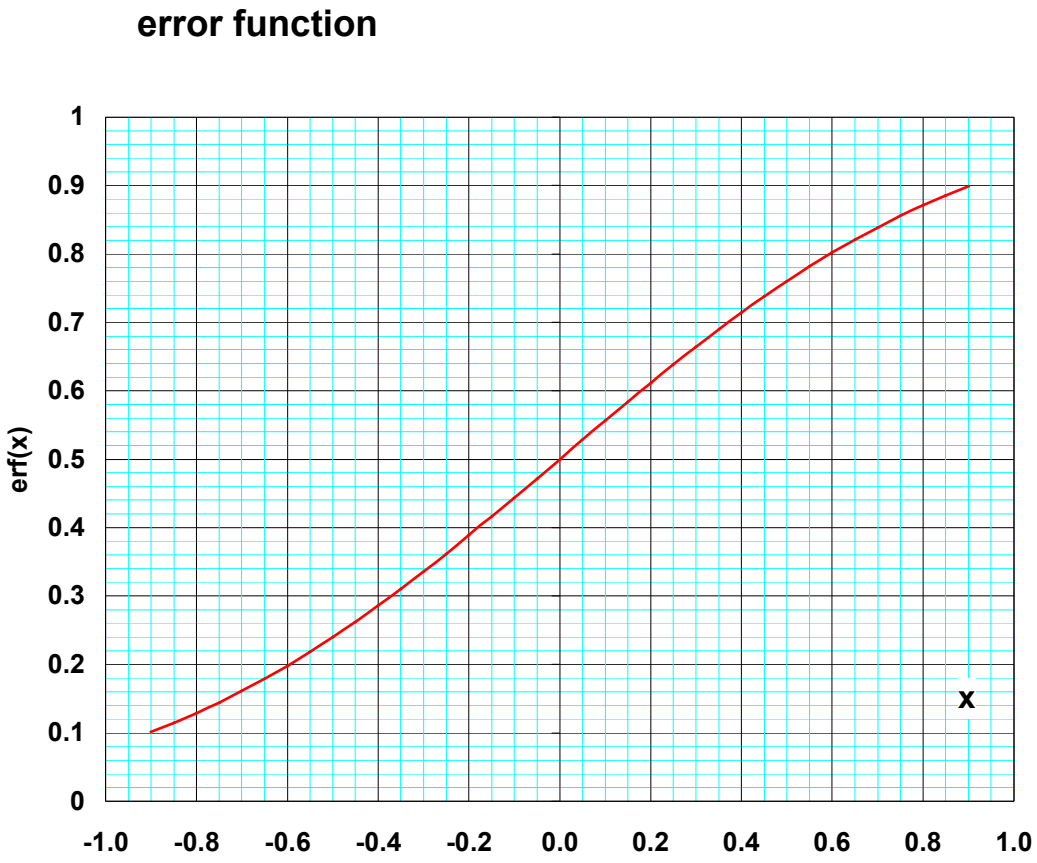


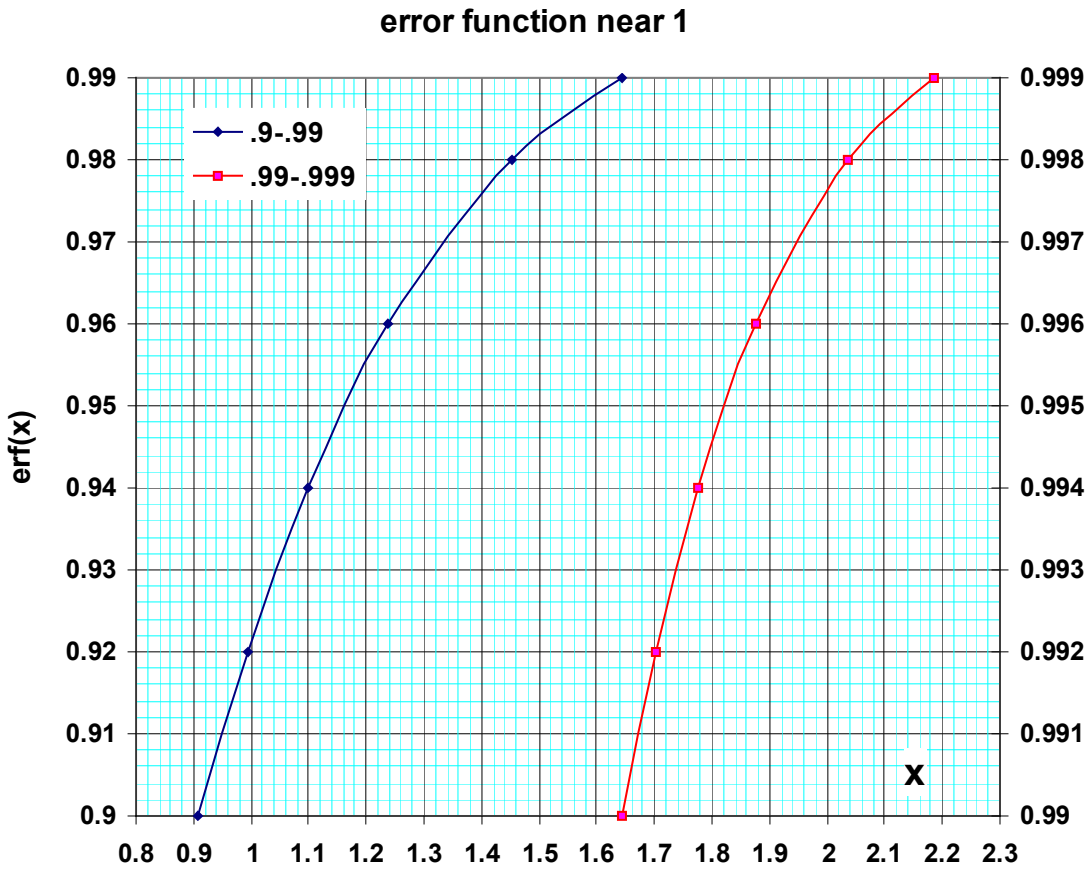
Gaussian curve

With this definition one obtains $< x^2 > = ½$.

As an example we consider a schooltest for all children of the country. The level of each child has several noise contributions: some genetic qualities, education, situation at home, friends, qualities of teachers and so on. The series of noise contributions will naturally cause a Gaussian distribution of the score q.

Corresponding with this bell shaped distribution there is the so called errorfunction erf(x). That describes for a value of x how large the fraction [percentage / 100] of all cases is, that has a value below x . We give two graphs with errorfunctions, the one in the middle range, the other in the extremes near 1 .


error function
erf(x)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
x



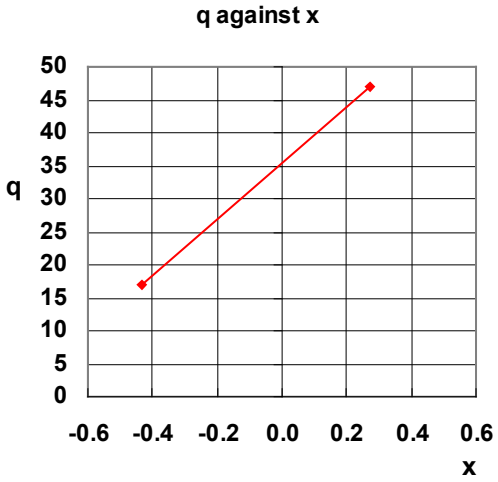
Note that the errorfunction is symmetric around the point (0, .5) , so that the very small values near 0 can simply be derived from those near 1. Furthermore we remark that an errorfunction sometimes is defined with a distribution with $\langle x^2 \rangle = 1$. We made a choice for $\langle x^2 \rangle = \frac{1}{2}$.

It is known that the schooltest yields a Gaussian distribution and that 27% of the children has a score below q=17 . Also 65% gains below q=47 . What is the centre of the distribution? In the graph of erf(x) we find for each fraction the corresponding x . We set the values of x in the table.

fraction	below q =	below x =
.27	17	-.43
.65	47	.27

Now we make a little graph of q against x . The relation between x and q is linear, a straight line in a graph.

[see note 6]



The centre of the distribution is at x = 0 and so q = 35 . We can also take some q = 44 at x = 0.2 . There the errorfunction has the value .61 or 61% . With the linear relation between q and x one may swith from the one scale to the other. In a perfectly Gaussian distribution, the relation between q and x is always linear.

miscellaneous subjects

previous selection

A previous selection of data sometimes leads to wrong conclusions.

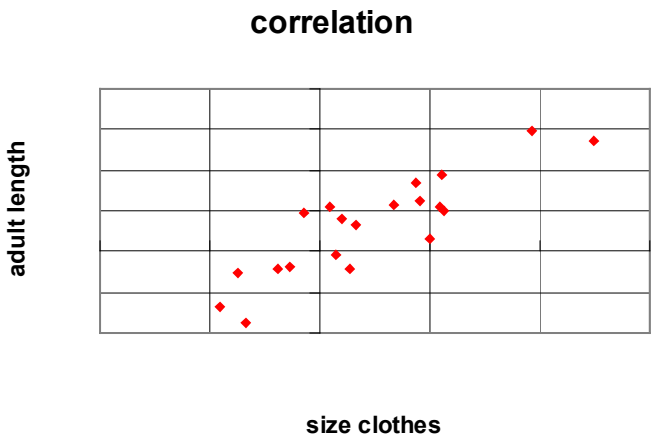
Suppose that in a large hospital one selects the 1 $^{o}/_{oo}$ most strange cases of death. A combination of diseases or something else may lead to a very strange picture. The resulting list is given to the police, which is perplex by such strange deaths in the hospital. The physicians involved are arrested.

What happens here? In every hospital there occur unexpected cases. Physicians are sometimes surprised by strange coincidences. Such things happen. But when one makes a selection of these cases, one obtains alarming statistics.

When one is alarmed by statistical data, one should always ask whether some selection has been made. Statistics with hidden or concealed selection may lead to the most terrifying confusion.

correlations and causes

One of the main pitfalls in statistics is a correlation. Young people may wear large or small clothes. One may note the size. After 20 years one notes their length. In a graph one denotes horizontally the size of the clothes and vertically the length in later years. It is very likely that one sees a correlation. Someone concludes then that young people should not wear small clothes, because they cause a smaller growth in later years. A correlation is easily confused with a causation. This error is often seen in statistics.



test of a hypothesis

We return to the towels. Suppose that the truck with new supply has visited the shop with the towels. The new stock is 30 towels. The next day all of them have disappeared. The assistant of the store is suspected of theft, but he declares that the whole stock has been sold. The manager of the shop takes this explanation as an hypothesis. If the weekly sale of 20 towels is obtained in 5 days, the daily sale is 4 towels on an average. A typical margin must be several times $\sqrt{4} = 2$. The average number is exceeded by 26 towels. How many times the typical margin is contained in 26 ? Let

$26 = k \sqrt{4}$

or $k = 13$. In the tables of the Poisson statistics the manager sees that a margin of $13 \sqrt{a}$ would give an astronomically small probability. He rejects the hypothesis.

Eindhoven, november 2014.

mathematical notes

note 1
The sum over n noise numbers is called a random walk. Many random walks together behave according to a diffusion equation $\partial f / \partial t = D \partial^2 f / \partial x^2$ (t = time, x = position in R_1). Purely for dimension reasons one expects a typical distance $x = \sqrt{(D*t)}$. The time t corresponds with the number of steps.

note 2
What is the probability P of n hits, if the average value of n is a ? The Poisson statistics behave like

$P = a^n \exp(-a) / n!$

A cumulative distribution in a spreadsheet is easily obtained by the sum over all probabilities below some threshold.

note 3
Noise numbers r on an interval between -.5 and +.5 have an average $< r^2 >$ according to

$< r^2 > = \int r^2 \, dr = 1/12$.

note 4
The Gaussian distribution may be obtained as follows. We sort our random numbers x in a rank of equal urns, with labels 1, 2, 3, …., K . In urn 1 there are n_1 hits, in urn 2 there are n_2 and so on.

As a first condition we know the total number $n_1 + n_2 + \ldots = N$.

We know as a second condition that $n_1 x_1^2 + n_2 x_2^2 + \ldots = N < x^2 >$.

Further the probability P of a configuration (n_1, n_2, …) is given by a multinomial distribution

$P = A / n_1! \, n_2! \ldots$, where A is some constant.

For large numbers n we may use the Stirling formula

$\ln n! \approx n \ln (n) - n$,

so that for ln(P) we get an entropy expression S

$S = A' - n_1 \ln (n_1) - n_2 \ln (n_2) \ldots\ldots \approx \ln (P)$.

We maximize S in n_k space with the Lagrange multiplier method, under the two conditions:

$\partial S / \partial n_k = \alpha \, \partial N / \partial n_k + \beta \, \partial (N < x^2 >) / \partial n_k$

so that

$\ln n_k = \alpha' - \beta x_k^2$

or

$n_k = \alpha'' \exp (- \beta x_k^2)$

which leads, with the right parameters α" and β , to a Gaussian distribution. The Gaussian distribution gives the maximum of entropy when $< x^2 >$ is given.

Another view on a Gaussian distribution is the similarity with a diffusion process. If we sum 12 random numbers, this may be seen as a random walk in 12 noisy steps. With very many of such walks one obtains a diffusion process. In note 1 we saw already the diffusion equation. If there is a pointlike start in x = 0 at t = 0 , the function f obeys $f = (1/\sqrt{t}) \exp(-x^2 / 4 D t)$. At a given time t there is a Gaussian distribution.

note 5

$(1/\sqrt{\pi}) \int \exp(-x^2) dx = 1$ [integral between -∞ and +∞]

$(1/\sqrt{\pi}) \int x^2 \exp(-x^2) dx = ½$

so $< x^2 > = ½$

define the error function:

$\mathrm{erf}(x) = (1/\sqrt{\pi}) \int \exp(-x^2) dx$ [integral between -∞ and x]

note 6
Is the relation between q and x linear? We have a differential distribution like

$\exp (- (p-p_0)^2 / \lambda)$,

where p_0 is the centre and λ gives the width of the distribution. This expression may be matched to

$\exp (-x^2)$

by a linear relation

$x = (p-p_0) / \sqrt{\lambda}$.

If one knows that $< x^2 > = ½$ and if also $< (p-p_0)^2 >$ is known, one may calculate λ by

$\lambda = 2 < (p-p_0)^2 >$.

This may be a short way to adapt the scales of x and p .