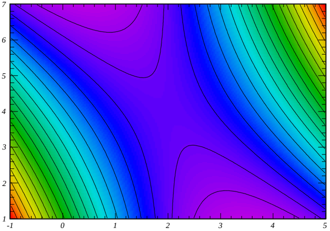
Advanced Methods in Data Analysis

Michael Schmelling – MPI for Nuclear Physics

- *Basics*
- *Monte Carlo Methods*
- *Error propagation*
- *Parameter estimates*
- *Unfolding*
- *Multivariate analysis*
- *Markov Chain Monte Carlo*



Literature

➜ *in alphabetical order. . .*

- R.J. Barlow,*Statistics*, Wiley
- S. Brand, *Data Analysis*, Springer
- G.D. Cowan,*Statistical Data Analysis*, Oxford University Press
- F. James, *Statistical Methods in Experimental Physics*, World Scientific
- H.L. Harney, *Bayesian Inference*, Springer
- D.E. Knuth, *The Art of Computer Programming*, Addison Wesley
- W.T. Press et al., *Numerical Recipes*, Cambridge University Press
- D.S. Sivia, *Data Analysis - A Bayesian Tutorial*, Oxford University Press
- plus many more . . .

→ *statistics everywhere. . .*

(sugar served with espresso)

front side

back side

"statistics sweetens your life"

"during our lives we cover 22150 km on foot"

➜ *the story of the cheating baker*

Once upon a time, in a holiday resort the landlord L. ran a profitable B&B, and every morning bought 30 rolls for breakfast. By law the mass of a single roll was required to be 75 g. One fine day the owner of the bakery changed, and L. suspected that the new baker B. might be cheating. So he decided to check the mass of what he bought, using a kitchen scales with a resolution of 1g.

After one month he had collected a fair amount of data. . .

```
73 79 72 62 67 60 60 67 78 68 66 75 76 73 75 64 70 69 73 59 70 73 64 72 64 69
69 71 69 71 77 69 72 71 67 72 63 66 68 76 71 76 68 71 63 65 65 66 73 73 73 67
70 65 71 69 78 67 65 69 71 71 72 73 72 69 66 66 70 60 72 62 53 65 74 65 68 69
67 75 64 76 72 76 78 67 67 67 69 79 71 67 71 68 71 65 66 65 78 76 71 70 67 65
67 64 73 67 74 79 74 71 73 67 66 76 68 74 76 65 77 67 71 67 71 77 63 66 70 62
68 74 67 67 67 77 65 68 79 72 71 77 68 70 73 67 81 70 74 71 79 62 67 63 68 76
73 81 76 73 68 72 76 61 69 73 71 80 68 70 62 76 58 68 68 64 68 78 69 65 70 70
64 75 73 72 60 86 68 68 64 60 68 71 70 75 70 67 69 67 73 65 66 71 70 70 73 66
72 71 71 64 76 75 72 72 71 72 72 71 75 68 73 70 64 76 72 75 79 70 64 70 67 70
75 70 83 69 61 70 66 69 71 72 70 76 73 62 71 60 73 74 70 68 68 70 78 71 69 71
73 73 75 65 71 67 60 70 77 71 74 64 74 73 60 77 73 70 69 66 70 78 69 75 66 71
75 75 74 69 74 70 75 77 75 66 72 68 72 61 75 65 69 68 65 73 82 67 75 67 80 71
79 72 71 68 73 70 67 75 74 69 63 63 72 70 73 63 70 70 59 78 76 66 72 79 65 71
76 72 69 69 73 70 77 73 83 66 68 67 69 73 76 65 71 70 71 65 78 71 67 70 72 75
67 79 72 64 62 79 68 70 61 65 68 71 73 60 60 68 71 74 75 69 73 70 68 ...
```

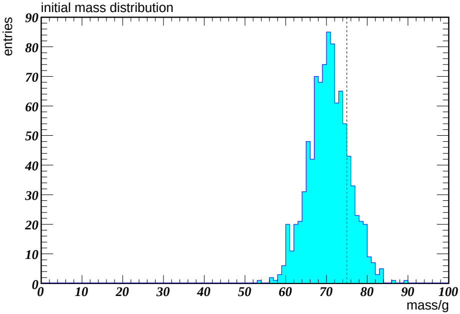
Data reduction

- the raw list of number is not very useful ➜ need some kind of data reduction
- assume that all measurements are equivalent
  - ➜ the sequence of the individual data does not matter (in this example)
  - ➜ all relevant information is contained in the number of counts per reading

```
count[50]=  0   count[60]= 20   count[70]= 85   count[80]=  9
count[51]=  0   count[61]= 11   count[71]= 81   count[81]=  7
count[52]=  0   count[62]= 20   count[72]= 61   count[82]=  3
count[53]=  0   count[63]= 21   count[73]= 65   count[83]=  5
count[54]=  0   count[64]= 31   count[74]= 54   count[84]=  0
count[55]=  0   count[65]= 48   count[75]= 43   count[85]=  0
count[56]=  2   count[66]= 42   count[76]= 33   count[86]=  1
count[57]=  1   count[67]= 70   count[77]= 23   count[87]=  0
count[58]=  3   count[68]= 68   count[78]= 21   count[88]=  0
count[59]=  6   count[69]= 74   count[79]= 20   count[89]=  1
```

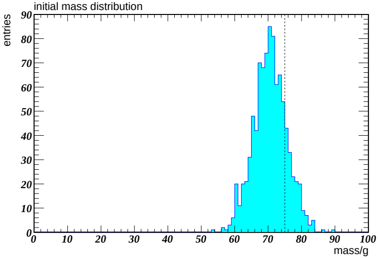
- ➜ much improved presentation of the collected information
- ➜ the above numbers cover the entire data set
- ➜ most of the measurements are lower than the legally required value...

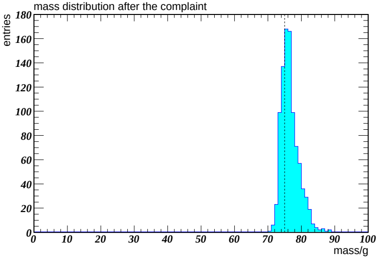
- an even better presentation of the available information: bar-chart
- example for the concept of a histogram
  - ➜ define bins for the possible values of a variable
  - ➜ plot the number of entries in each bin
  - ➜ get an immediate grasp of center and width of the distribution



The rolls produced by baker B. definitely are too light. So L. was right in his suspcion, that B. tried to make some extra profit by cheating...

As a consequence of his findings, L. complained. B. apologized and claimed that the low mass of the rolls was an accident which will be corrected in the future. L., however, continues to monitor the quality delivered by the baker. One month later, B. asked whether now everything was all right. L., for his part, acknowledged that the weight of the rolls now matched his expectations, but also voiced the opinion that B. was still cheating. . .





➜B. simply selected the heaviest rolls for L.!

Before moving on . . .

➜ *always keep in mind:*

- the name of the game: extract meaning from a stream of numbers
- the tools: “statistical and numerical methods”
  - ➜ need know the relevant methods
  - ➜ need to understand their properties
- basic assumptions
  - ➜ measurements deviate from the respective true values
  - ➜ the deviation is a random variable
  - ➜ statistics builds on probability theory

> A statistical method is neither **“right”** nor **“wrong”**.
> It has **properties**, which have to be known for the interpretation of the result. Possible properties could be, that the output is the most precise estimator, or that the result is robust. The property could also be that the result is wrong, in which case use of this particular method should be discouraged...

Notations

$p(A)$	probability for A
$p(A\|B)$	conditional probability for A if B is given
$x, y, z, t, \ldots$	continuous random variable
$i, j, k, l, m, n \ldots$	discrete random variable (or index)
$\vec{x}$	vector of random variables $\{x_1, \ldots, x_n\}$
p_i, q_i	discrete probabilities
$f(x), g(x)$	probability densities functions (PDFs) of x
$F(x), G(x)$	cumulative distributions of f, g
$f(x, y)$	2-dim probability density in x und y
$f(x\|y)$	conditional PDF for x given y
$a, b, \ldots, \alpha, \beta, \ldots$	parameters
$E[x] = \langle x \rangle = \mu_x$	expectation value von x
$V[x] = \sigma_x^2$	variance von x
$\widehat{a}$	estimate for a
$\overline{x}$	arithmetic average of x
$\sum_{(i)}$	sum over all indices (i)
$\int dx$	integrate over all x

A matrix $A[m,n]$ is an array of numbers with m rows und n columns. Usually the dimensions are not given explicitly. Individual matrix elements are addressed by two indices, A_{ij}, where the first index specifies the row and the second one the column. The following is a summary of the rules for matrix manipulations:

Sum of two matrices:

$$C[m,n] = A[m,n] + B[m,n] \quad \text{or} \quad C_{ij} = A_{ij} + B_{ij}$$

Product of two matrices:

$$C[m,n] = A[m,l] \cdot B[l,n] \quad \text{or} \quad C_{ij} = \sum_{k=1}^{l} A_{ik} B_{kj}$$

Product of three matrices:

$$D[m,n] = A[m,l] \cdot B[l,k] \cdot C[k,n] \quad \text{or} \quad D_{ij} = \sum_{r=1}^{l} \sum_{s=1}^{k} A_{ir} B_{rs} C_{sn}$$

associative law of matrix multiplication:

$$A \cdot (B \cdot C) = (A \cdot B) \cdot C$$

The neutral element with respect to matrix multiplication is the unit matrix

$$\mathbf{1}[n,n] = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \text{using indices} \qquad \mathbf{1}_{ij} = \delta_{ij}$$

$$\text{giving} \qquad A[n,m] \cdot \mathbf{1}[m,m] = \mathbf{1}[n,n] \cdot A[n,m] = A[n,m]$$

Square matrices $A[n,n]$ (of rank n) have a unique inverse matrix A^{-1}:

$$A^{-1} \cdot A = A \cdot A^{-1} = \mathbf{1}$$

For the inverse of a product of square matrices on has:

$$(A_1 \cdot A_2 \cdots A_n)^{-1} = A_n^{-1} \cdots A_2^{-1} \cdot A_1^{-1}$$

Another matrix operation is transposition:

$$A[m,n]^T = B[n,m] \quad \text{or} \quad B_{ij} = A_{ji} \ .$$

For die transpose of a product of matrices one has:

$$(A_1 \cdot A_2 \cdots A_n)^T = A_n^T \cdots A_2^T \cdot A_1^T$$

Linear algebra (iii)

For $n \times n$ matrices there exist n scalar quantities which are invariant under orthogonal transformations of the matrix. The two most important ones are determinant and trace, the product and the sum of the eigenvalues λ_i of the matrix:

$$\det(A[n,n]) = \prod_{i=1}^{n} \lambda_i \quad \text{and} \quad \mathrm{Tr} A[n,n] = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} A_{ii}$$

The trace is given by the sum of the diagonal elements. Expressed as a function of the matrix elements, the determinant of a 2×2 matrix is

$$\det(A[2,2]) = A_{11}A_{22} - A_{12}A_{21}$$

For the determinant of a product of matrices one finds:

$$\det(A_1 \cdot A_2 \cdots A_n) = \det(A_1) \cdot \det(A_2) \cdots \det(A_n)$$

The trace of a product of matrices is invariant under cyclic permutations:

$$\mathrm{Tr}(A_1 \cdot A_2 \cdots A_n) = \mathrm{Tr}(A_2 \cdots A_n \cdot A_1)$$

A special class of matrices are vectors. In the following a letter with an arrow denotes a column vector. Row vectors are obtained by transposition (T) of a column vector.

$$\vec{b} = b[n, 1] \quad \text{column vector}$$

$$\vec{a}^{\,T} = a[1, n] \quad \text{row vector}$$

For two vectors $\vec{a}$ and $\vec{b}$ of dimensions n, $\vec{a}^{\,T} \cdot \vec{b}$ is a scalar and $\vec{a} \cdot \vec{b}^{\,T}$ is a matrix:

$$\vec{a} \cdot \vec{b}^{\,T} = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \dots \\ a_2 b_1 & a_2 b_2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

It follows:

$$\vec{a}^{\,T} \cdot \vec{b} = \mathrm{Tr}(\vec{a}^{\,T} \cdot \vec{b}) = \mathrm{Tr}(\vec{b} \cdot \vec{a}^{\,T})$$

Expectation values of matrices are defined by element:

$$\langle A \rangle_{ij} = \langle A_{ij} \rangle$$

Manipulations of Sums

The product of two sums can be written as a sum over two indices

$$\left(\sum_i x_i\right)\left(\sum_j y_j\right) = \sum_{ij} x_i y_j$$

i.e. interpreting x_i or y_i as elements of a vector $\vec{x}$ or $\vec{y}$, respectively, every element of $\vec{x}$ is multiplied with every element of $\vec{y}$ and the individual terms summed up.

Special case: $\vec{y} = \vec{x}$

$$\left(\sum_i x_i\right)\left(\sum_j x_j\right) = \left(\sum_i x_i\right)^2 = \sum_{ij} x_i x_j = \sum_{i=j} x_i^2 + \sum_{i\neq j} x_i x_j$$

Since the expectation value (formally defined later) is a linear operator sums and expectation values commute:

$$\left\langle \sum_i x_i \right\rangle = \sum_i \langle x_i \rangle$$

Lagrange multipliers (i)

➜ *general problem: minimization subject to constraints*

Consider the general constrained minimization problem in 2 dimensions:

$$C(x,y) \stackrel{!}{=} \min \quad \text{with} \quad g(x,y) = 0$$

➜ *default approach:*

Use $g(x,y) = 0$ to solve for $y = G(x)$, substitute

$$\frac{\partial}{\partial x} C(x, G(x)) = 0 \quad \text{with} \quad g(x, G(x)) = 0$$

and determine $x_{\min}$ and $y_{\min} = G(x_{\min})$.

- ➜ conceptually straightforward ansatz
- ➜ minimization problem with reduced number of dimensions
- ➜ breaks the symmetry between the variables
- ➜ often impossible to do in practice

try to come up with something better...

➜ *the Lagrange multiplier approach*

Example: The Milkmaid's Problem
A milkmaid is sent to a field close to the river in order to milk a cow. Entering the field at point M, the milkmaid spots the cow at C. Normally she would go directly to the cow, – but then realizes that her bucket first needs cleaning in the river. The problem is to find the shortest path connecting M and C via the bank of the river.

❖ mathematical formulation:

cost function:

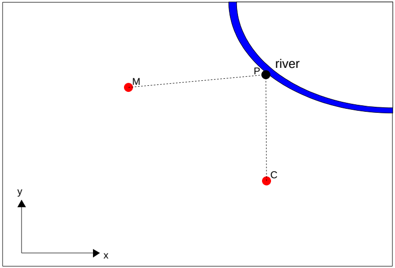
$$C(P_x, P_y) = |\vec{M} - \vec{P}| + |\vec{P} - \vec{C}|$$

description of the distance to the river:

$$g(x, y) = c$$

constraint:

$$g(P_x, P_y) = 0$$



Lagrange multipliers (iii)

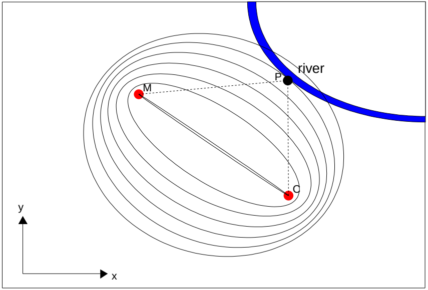
The points where the sum of the distances to two "focal" points is constant are located on an ellipse. Contours of equal cost thus are given by ellipses around C and M. The best solution is the smallest ellipse touching the river. At this point the contour lines C =const and g =const have to be parallel.

- ➜ contour lines are orthogonal to function gradients
- ➜ parallel contour lines implies parallel gradients

❖ condition for the best point P:

$$\nabla C(x, y) \propto \nabla g(x, y)$$

Exploit this to find an elegant way for solving constrained optimization problems. . .



➜ *insight by Lagrange*

The stationary point of a linear combination of cost function C and constraint function g is the solution of a constrained minimization. Introducing

$$F(x, y) = C(x, y) + \lambda \cdot g(x, y)$$

one finds

$$\nabla F(x, y) = 0 = \nabla C(x, u) + \lambda \cdot \nabla g(x, y) \quad \text{i.e.} \quad \nabla C(x, u) \propto \nabla g(x, y)\,.$$

➜ *discussion*

- minimization of F is usually much easier than the “default approach”
- fully symmetric in all variables
- the result is a function of λ, i.e $x(\lambda)$,$y(\lambda)$
- λ can be determined from the condition $g(x(\lambda), y(\lambda)) = 0$
- in many cases the explicit value of λ is not needed

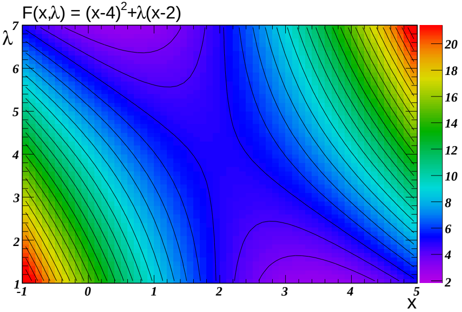
➜ additional remarks

Introduction of λ increases the dimension of the minimization problem and a stationary point is determined in a higher dimensional space. Since the extended cost function $F(x,y)$ is linear in λ the stationary point will be saddle point.

Example: $C(x) = (x-4)^2$ and $g(x) = x-2$

$$F(x,\lambda) = (x-4)^2 + \lambda \cdot (x-2)$$

- local minimum in x for every λ
- no global minimum
- the saddle-point has minimum cost for constraint $g(x) = 0$
  - ➜ $x_{\text{min}} = 2$
  - ➜ $\lambda_{\text{min}} = 4$



Combinatorics

An important aspect of many statistical analyses is to count the number of possible results. For discrete states the solution is found by combinatorics. Some of the most important results are collected below:

➜ *words with m-characters from an alphabet with n letters:*

$$N = n^m$$

➜ *Permutations of n objects:*

$$N = n \times (n-1) \times (n-2) \times \ldots 2 \times 1 = n!$$

➜ *Possibilities to select k objects from a total of n (without putting back)*

$$\frac{n(n-1)...(n-2)(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \left(\begin{array}{c} n \\ k \end{array} \right)$$

the “lottery-problem”

➜ *Kolmogorov's axioms on probability*

Starting from set theory, probability theory can be built on a mapping from sets E to real numbers $p(E) \in [0, 1]$. Define

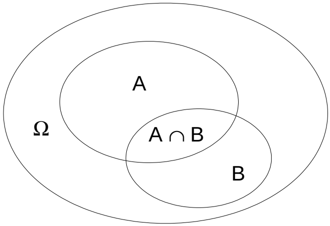
$$\begin{aligned} \Omega &: \text{the entire set} \\ E &: \text{partial set of } \Omega \\ p(E) &: \text{probability of } E \end{aligned}$$

and postulate the following axioms:

1. $0 \leq p(E) \leq 1$
2. $p(\Omega) = 1$
3. $p(E_1 \cup E_2) = p(E_1) + p(E_2)$ if $E_1 \cap E_2 = 0$

Based on these axioms, calculations involving probabilities are unambiguously defined. Interpretation is left completely open . . .

Rules for calculus of probabilities derived from Kolmogorov's axioms can easily be visualized using diagrams from set theory. For example:



$$p(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Consider $P(B|A)$, the probability for B if A is given

- ➜ the diagram suggests $P(B|A) \propto P(A \cap B)$
- ➜ for $A \in B$ one must have $P(B|A) = 1$
- ➜ $A \in B$ implies $P(A \cap B) = P(A)$ and thus

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$ "conditional probability"

For "independepent events", which implies $P(B|A) = P(B)$, one obtains:

$$P(A \cap B) = P(A) \cdot P(B)$$

Bayes´ theorem

Consider a set of disjoint events A_i, $i = 1, \ldots, n$. It follows

$$p(A_i \cap B) = p(B|A_i)\, p(A_i) = p(A_i|B)\, p(B)$$

$$\Longrightarrow \quad p(A_i|B) = \frac{p(B|A_i)\, p(A_i)}{p(B)}$$

Bayes’ theorem

The prior $p(A_i)$ for A_i is updated by the occurrence of B to become $p(A_i|B)$.

Bayes’ theorem is at the heart of statistical inference based on empirical input. If the A_i are exhaustive, i.e. if one of them is realized with unit probability independently of B, then one has

$$p(B) = \sum_i p(B|A_i) p(A_i)$$

and thus

$$p(A_k|B) = \frac{p(B|A_k) p(A_k)}{\sum_i p(B|A_i) p(A_i)}$$

➜ applications . . .

Bayes' theorem - example 1

A new test for the common cold hits the market, designed to detect an infection in the early stages where an efficient cure is available. The probability to test positive in case of an infection is $p(+|I) = 0.98$, the probability for a negative result on a healthy subject is $p(-|H) = 0.97$. Series tests are performed in summer, where the a priori probability for infection is $p(I) = 0.001$.

What's the probability that a person tested positive has actually contracted a cold?

the probabilities are:

$$\begin{array}{lclclcl} p(I) & = & 0.001 & p(H) & = & 0.999 \\ p(+|I) & = & 0.980 & p(-|I) & = & 0.020 \\ p(+|H) & = & 0.030 & p(-|H) & = & 0.970 \end{array}$$

where the rows sum up to unity. Application of Bayes' theorem then yields

$$p(I|+) = \frac{p(+|I)p(I)}{p(+|I)p(I) + p(+|H)p(H)} \approx 0.032$$

Simply administering sweets to all patients that diagnosed "infected" already will yield a "healing rate" around 97%.

Bayes´ theorem - example 2

Three boxes contain each two rings made of either gold (G) or silver (S). The boxes contain (GG), (SS) and (GS). The content of a specific box is unknown. A person is allowed two draws of a single ring from any of the boxes. The first draw yields gold.

Which box for the second draw maximizes the number of gold rings?

Calculate the probability that the box of the first draw contains (GG). A priori the probabilities are $p(GG) = p(GS) = p(SS) = 1/3$. The probabilities to get (G) in the first draw become

$$p(G|GG) = 1 \quad , \quad p(G|GS) = \frac{1}{2} \text{ and } p(G|SS) = 0 \; .$$

Bayes' theorem then yields the probability that the selected box is (GG):

$$p(GG|G) = \frac{p(G|GG)p(GG)}{p(G|GG)p(GG) + p(G|GS)p(GS) + p(G|SS)p(SS)} = \frac{2}{3}$$

The second draw should be taken from the same box.

Two old friends A and B who have gotten out of touch accidentally meet in a pub and decide to celebrate the occasion. A suggests to flip a coin in order to determines who will pay the next round. B agrees and then pays all the drinks.

What is the probability that A is cheating each time he throws the coin?

Consider the hypotheses h and c that A is an honest guy or that he is a cheater. The probability for A to win n times in a row is

$$p(n|h) = 2^{-n} \quad \text{and} \quad p(n|c) = 1$$

With the prior probabilities $p(h)$ and $p(c) = 1 - p(h)$, Bayes' theorem allows to determine the probability that A, after having won n times, is a cheater:

$$p(c|n) = \frac{p(n|c)p(c)}{p(n|c)p(c) + p(n|h)p(h)} = \frac{p(c)}{p(c) + 2^{-n}p(h)}$$

the result depends on $p(b)$:

$$\begin{array}{lllllll} p(c) & = & 0.00 & \Longrightarrow & p(c|n) & = & 0 \\ p(c) & = & 0.05 & \Longrightarrow & p(c|1) & \approx & 0.095 \\ & & & & p(c|6) & \approx & 0.771 \\ & & & & p(c|\infty) & = & 1 \end{array}$$

"bayesian" update of knowledge

➜ *definition of a probability density function (PDF)*

A function $f(x)$ can be interpreted as a PDF if

$$f(x) \geq 0 \ \ \forall \, x \quad \text{and} \quad \int_{-\infty}^{+\infty} dx\, f(x) = 1 \; .$$

➜ *interpretation:*

The probability to observe an event in the infinitesimal interval $[x, x + dx]$ is:

$$p(x, x + dx) = f(x)\, dx \; .$$

➜ *relation to discrete probabilities:*

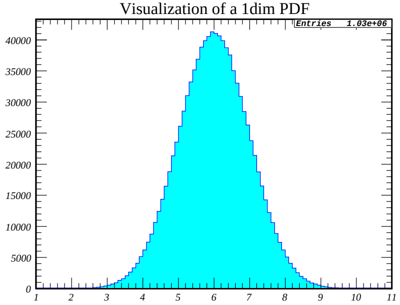
discrete probabilities p_i, i.e. finite probabilities for discrete values, can be written as a PDF using Dirac's delta-function:

$$f(x) = \sum_{i=1}^{n} p_i \, \delta(x - i) \quad \text{where} \quad \int dx\, f(x) = \sum_i p_i = 1$$

- graphical representation of the density
- problem in practical applications
  - ➜ density function not known
  - ➜ only a random sample of size N
- obvious solution: mark the values
- better solution: histogram
  - ➜ divide the range into bins
  - ➜ count entries inside each bin
  - ➜ regarding bin limits:
    - ✔ too many bins: large fluctuations
    - ✔ too few bins: loss of information
    - ✔ use “reasonable” binning

➜ *to illustrate the point...*

for a range $-1 \le x \le 1$ avoid histograms with 25 bins on the interval $[-1.1, +1.1]$. Use 20 bins between -1 and 1.



(demo 01)

- variations:
  - ➜ density plots for small N
  - ➜ variable bin widths
  - ➜ logarithmic axes
  - ➜ ...

Relation: histogram $\Longleftrightarrow$ PDF

➜ *given*

N : total number of entries in the histogram

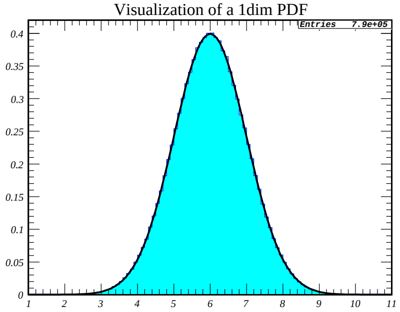
h : bin width

n_k : number of entries in bin k $[x_k - h/2, x_k + h/2]$

➜ *it follows*

$$n_k = N \times p(x - h/2, x + h/2)$$
$$= N \int_{x_k - h/2}^{x_k + h/2} dx\, f(x)$$
$$\approx N\, f(x_k)\, h$$

result: $$f(x_k) \approx \frac{n_k}{h \cdot N}$$



(demo)

➜ *goal:*

Summarize the properties of a PDF by (a few) numbers, so-called moments:

➜ *moments are "expectation values", defined by*

$$\int_{-\infty}^{\infty} dx\, f(x)\, w_k(x) = \langle w_k \rangle$$

i.e. as a mapping $f(x) \mapsto C$ of a PDF $f(x)$ onto a (complex) number via integral transform with a (family of) weight function(s) $w_k(x)$.

Example: cumulative distribution

$$w_X(x) = \Theta(X - x)$$

$$\langle w_X \rangle = \int_{-\infty}^{\infty} dx\, f(x)\, \Theta(X - x) = \int_{-\infty}^{X} dx\, f(x) = F(X)$$

$F(x)$ is the primitive of $f(x)$: $F(-\infty) = 0$, $F(\infty) = 1$

➜further examples . . .

A possible measure for the scatter s of x with PDF $f(x)$ around a point a is

$$s^2 = \int dx\, (x-a)^2 f(x)$$

To use s for characterizing $f(x)$, the point a should be chosen such that s becomes minimal. Minimization of s^2 yields:

$$\frac{\partial s^2}{\partial a} = -2 \int dx\, (x-a) f(x) \stackrel{!}{=} 0 \quad \text{i.e..} \quad a_{\text{min}} = \int dx\, x\, f(x) = \langle x \rangle$$

It follows that the mean value (or “expectation value”) $\langle x \rangle$ is a way to characterize the center of a PDF. For symmetric PDFs it is also the symmetry point:

$$\langle x \rangle = \int dx\, x\, f(x) = \int dx\, (x-a)\, f(x) + a \int dx\, f(x) = 0 + a \times 1 = a$$

The scatter σ around the mean value $\langle x \rangle$ is also referred to as “standard deviation” oder “rms”-scatter, its square as “variance”. The following relation holds:

$$\sigma^2 = \int dx\, (x - \langle x \rangle)^2 f(x) = \int dx\, (x^2 - 2x \langle x \rangle + \langle x \rangle^2)\, f(x) = \langle x^2 \rangle - \langle x \rangle^2$$

→ *median*

The center of a distribution can also be taken as the median m, defined by

$$\int_{-\infty}^{m} dx\, f(x) = \int_{m}^{\infty} dx\, f(x)$$

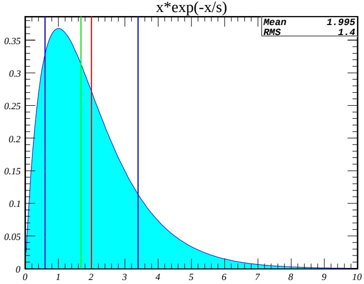
i.e. same probability on both sides. For symmetric distributions one has $\langle x \rangle = m$.

→ *quantiles*

Quantiles are locations x_α on a PDF up to which with the probability content is $\alpha\%$. A possible measure for the width of a PDF is $x_{84} - x_{16}$.

→ *discussion:*

- mean value and standard deviation
  - \+ linear functions of the PDF, i.e. easy to use in theoretical calculations
  - − sensitive to outliers and tails in the PDF
- median and quantiles
  - \+ insensitive against outliers and tails
  - − non-linear functions of the PDF, difficult to handle analytically



probability content of [av-sd,av+sd] = 0.732

(demo 02)

➜ *for different PDFs*

- mean value $\langle x \rangle$
- standard deviation σ
- probability content of the interval $[\langle x \rangle - \sigma, \langle x \rangle + \sigma]$
- median

➜ *conclusion:*

- there are many possibilities to characterize a PDF
- other options, not discussed in detail are:
  - ➜ take as center the maximum
  - ➜ take as width the minimum interval with a given probability
- still most important: algebraic moments and derived quantities

➜ *algebraic moments:*

$$M_k \equiv \int dx\, f(x)\, x^k$$

- $M_0 = 1$: normalization of $f(x)$
- $M_1 = \mu$: mean value $f(x)$

➜ *central moments:*

$$Z_k \equiv \int dx\, f(x)\, (x - \mu)^k$$

- $Z_0 = 1$: normalization of $f(x)$
- $Z_1 = 0$
- $Z_2 = \sigma^2$: variance of $f(x)$

➜ *other commonly used moments:*

$$S = \frac{Z_3}{\sigma^3} \quad \text{“skewness”} \quad \text{and} \quad K = \frac{Z_4}{\sigma^4} - 3 \quad \text{“kurtosis”}$$

Normalization by σ makes S and K to quantities which depend only on the shape. For symmetric distributions one has $S = 0$. K measures how quickly the PDF drops to zero. For gaussian distributions one has $K = 0$.

➜ *probability content in the tails*

Given any PDF $f(x)$ und eine Funktion $w(x) \geq 0$, there is a relation between $\langle w \rangle$ and the probability $p(w(x) \geq C)$, to observe x in a region with $w(x) \geq C$:

$$\langle w \rangle = \int dx\, f(x)\, w(x) \geq \int_{w(x)\geq C} dx\, f(x) w(x) \geq C \int_{w(x)\geq C} dx\, f(x) = C\; p(w(x) \geq C)$$

and thus $$p(w(x) \geq C) \;\leq\; \frac{\langle w \rangle}{C}$$

The special choice $w(x) = (x-\mu)^2$ and $C = k^2\sigma^2$ then yields the result:

$$p_k \equiv p\left((x-\mu)^2 > k^2\sigma^2\right) \leq \frac{1}{k^2}$$

The probability content beyond $\pm k\,\sigma$ around the mean value μ is at most $1/k^2$.

- upper limit for probability in the tails of a PDF
- actual probability contents for most PDFs are much lower
  - ➜ e.g. gaussian: $\{p_1, p_2, p_3, p_4\} \approx \{0.317, 0.0555, 0.0027, 0.000063\}$

➜ *convolution of two distributions*

Given two PDFs $f_1(x_1)$ und $f_2(x_2)$, determine the PDF $g(y)$ of $y = h(x_1, x_2)$, when x_1 and x_1 are distributed according to $f_1(x_1)$ and $f_2(x_2)$, respectively. For the cumulative distribution $G(Y)$ one has:

$$G(Y) = \int_{-\infty}^{Y} dy\; g(y) = \int dx_1\, dx_2 f_1(x_1) f_2(x_2)\; \Theta(Y - h(x_1, x_2))$$

Here the products of all probabilities $dp_1 = dx_1 f_1(x_1)$ and $dp_2 = dx_2 f_2(x_2)$ are summed which satisfy the constraint $h(x_1, x_2) < Y$. Differentiation with respect to the upper limit Y then yields the solution:

$$g(y) = \left.\frac{d}{dY} G(Y)\right|_{Y=y} = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \delta(y - h(x_1, x_2))$$

"general convolution integral"

For the special case $h(x_1, x_2) = x_1 + x_2$ follows the known result

$$g(y) = \int dx_1 f_1(x_1) f_2(y - x_1)$$

➜ consider moments. . .

$$M_k(y) = \int dy\, y^k g(y) = \int dy\, y^k \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \delta(y - x_1 - x_2)$$
$$= \int dx_1\, dx_2 f_1(x_1) f_2(x_2) \int dy\, y^k \delta(y - (x_1 + x_2))$$
$$= \int dx_1\, dx_2 f_1(x_1) f_2(x_2) (x_1 + x_2)^k$$

Leading order moments:

$$\langle y^0 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2) = 1$$

$$\langle y^1 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2)(x_1 + x_2) = \langle x_1 \rangle + \langle x_2 \rangle$$

$$\langle y^2 \rangle = \int dx_1\, dx_2 f_1(x_1) f_2(x_2)(x_1 + x_2)^2 = \langle x_1^2 \rangle + 2 \langle x_1 \rangle \langle x_2 \rangle + \langle x_2^2 \rangle$$

and thus $\quad \langle y^2 \rangle - \langle y \rangle^2 = \left[\langle x_1^2 \rangle - \langle x_1 \rangle^2 \right] + \left[\langle x_2^2 \rangle - \langle x_2 \rangle^2 \right]$

➜ *convolutions are normalized, mean value and variance add up for any PDFs!*

The central limit theorem

➜ *conditions:*

- n PDFs $f_i(x_i)$ with mean values μ_i and variances $\sigma^2(x_i)$
- all algebraic moments are finite, i.e. the PDFs $f_i(x_i)$
  - ➜ drop for $|x_i| \to \infty$ faster than any power of x_i
  - ➜ or only within a finite interval one has $f_i(x_i) \neq 0$
- consider the derived variable y:

$$y = \sum_{y=1}^{n} y_i = \sum_{i=1}^{n} \frac{x_i - \mu_i}{\sigma} = h(x_1, \dots x_n) \quad \text{with} \quad \sigma^2 = \sum_{i=1}^{n} \sigma^2(x_i)$$

  - ➜ y is a convolution of n PDFs with mean value $\mu = 0$
  - ➜ y is dimensionless
  - ➜ y is constructed such that the variances is $\sigma^2(y) = 1$

➜ *central limit theorem:*

For $n \to \infty$ the PDF of y converges towards a normal distribution $N(0,1)$:

$$g(y) = \lim_{n\to\infty} \int \prod_{i=1}^{n} dx_i\, f_i(x_i)\, \delta\left(y - h(x_1, \dots x_n)\right) = \frac{1}{\sqrt{2\pi}}\; e^{-y^2/2}$$

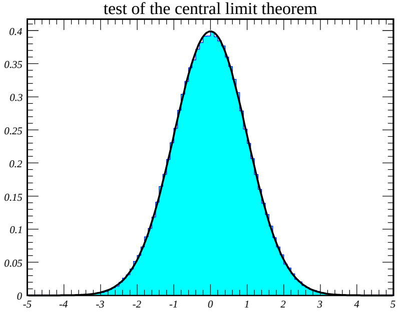
Illustration of the central limit theorem

➜ *convergence toward a normal (gaussian) distribution*

- generate n random numbers x_i according to two PDF
  - ➜ uniform distribution with $\sigma = 1/\sqrt{12}$
  - ➜ exponential distribution with $\sigma = 1$
- calculate the function $y = h(x_1, \ldots, x_n)$
  - ➜ $h = \sqrt{12/n}\sum_i x_i$ foruniform random numbers
  - ➜ $h = \sqrt{1/n}\sum_i x_i$ for exponential random numbers
- histogram y
- study convergence

A simple example how to do convolutions numerically

(demo 03)



➜ *generalization of 1-dim PDFs*

- non-negative, normalizable functions in n dimensions
- discuss the most important concepts with 2-dim PDFs

❖ 2-dim PDF:

$$f(x,y) \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\, f(x,y) = 1$$

❖ interpretation:

Probability for (x,y) in the (infinitesimal) rectangle $[x, x+dx] \times [y, y+dy]$

$$p(x, x+dx; y, y+dy) = f(x,y)\, dx\, dy$$

❖ independence of variables:

Two variables x and y are independent if the PDF factorizes

$$f(x,y) = f_y(x) \cdot f_x(y) = \left(\int dy\, f(x,y) \right) \cdot \left(\int dx\, f(x,y) \right)$$

The covariance between two variables

→ *look for a moment sensitive two dependencies between two variables*

normalisation	$\langle 1 \rangle$
first moments	$\langle x \rangle$, $\langle y \rangle$
second moments	$\langle x^2 \rangle$, $\langle xy \rangle$, $\langle y^2 \rangle$
third moments	$\langle x^3 \rangle$, $\langle x^2 y \rangle$, $\langle xy^2 \rangle$, $\langle y^3 \rangle$ etc.

The lowest order term sensitive to possible dependencies between x and y is $\langle xy \rangle$. For independent variables with $f(x, y) = g_1(x)\, g_2(y)$ one finds

$$\langle xy \rangle = \int dx \int dy\, x \cdot y \cdot g_1(x) \cdot g_2(y) = \left(\int dx\, x \cdot g_1(x) \right) \left(\int dy\, y \cdot g_2(y) \right)$$

and thus $\langle xy \rangle = \langle x \rangle \langle y \rangle$

❖ obvious candidat for a measure of correlation:

$C_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle$ "covariance" of x and y

The correlation coefficient

❖ consider the special case $y = ax + b$

$$\begin{aligned}
\langle y \rangle &= \langle ax + b \rangle &&= a\langle x \rangle + b \\
\langle xy \rangle &= \langle ax^2 + bx \rangle &&= a\langle x^2 \rangle + b\langle x \rangle \\
\text{and thus} \quad C_{xy} &= \langle xy \rangle - \langle x \rangle \langle y \rangle &&= a(\langle x^2 \rangle - \langle x \rangle^2) = aC_{xx}
\end{aligned}$$

Here the covariance is proportional to the slope between x and y, i.e. it measures linear correlation. The dimensionless correlation coefficient ρ derived from C_{xy} is a normalized measure for the correlation strength.

➜ (linear) correlation coefficient: $$\rho = \frac{C_{xy}}{\sigma_x \sigma_y} = \frac{C_{xy}}{\sqrt{C_{xx} C_{yy}}}$$

For $y = ax + b$ one has $C_{xy} = aC_{xx}$ and $C_{yy} = a^2 C_{xx}$ and thus:

$$y = ax + b \quad \rightarrow \quad \rho = \text{sign}(a) = \pm 1$$

The correlation is 100%. If the linear relation only holds between x and $\langle y \rangle$, i.e. $\langle y \rangle = a\,x + b$, then one has $|\rho| < 1$.

The covariance matrix

➜ *matrix of covariances between all pairs of variables in an n-dim PDF:*

$$C_{ij} = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

Expressed through standard deviations and correlation coefficients it is

$$C_{ij} = \rho_{ij} \cdot \sigma_i \sigma_j \quad \text{with} \quad \rho_{ii} = 1 \, .$$

➜ *note:*

- the diagonal terms C_{ii} are the variances of the individual variables
- off-diagonal terms are covariances
- the covariance matrix is symmetric and positive definite
- it can be diagonalized by rotation in the space of the variables
- C also is referred to as “error matrix”

The covariance matrix C_{ij} is the matrix of all 2nd order moments of an n-dimensional PDF $f(x_1, x_2, \ldots, x_n)$. Mean values $\langle x_i \rangle$ and C_{ij} describe the location, extension and orientation of the PDF.

➜ manipulations of sums. . .

Consider a transformation $y_k = \sum_i A_{ki} x_i$. Given the covariance matrix $C_{ij}(x)$, the covariance matrix $C_{kl}(y)$ of the transformed quantities shall be determined:

$$
\begin{aligned}
C_{kl}(y) &= \langle y_k y_l \rangle - \langle y_k \rangle \langle y_l \rangle \\
&= \left\langle \sum_i (A_{ki} x_i) \sum_j (A_{lj} x_j) \right\rangle - \left\langle \sum_i A_{ki} x_i \right\rangle \left\langle \sum_j A_{lj} x_j \right\rangle \\
&= \sum_{ij} A_{ki} A_{lj} (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) = \sum_{ij} A_{ki} A_{lj} C_{ij}(x)
\end{aligned}
$$

Matrix notation yields the compact expressions

$$\vec{y} = A \cdot \vec{x} \quad \text{and} \quad C(y) = A \cdot C(x) \cdot A^T \, .$$

- ➜ if $C(x)$ is positive definite, so is $C(y)$
- ➜ A need not be a square matrix - the number of rows is arbitrary

The n-dimensional gaussian

→ *functional form:*

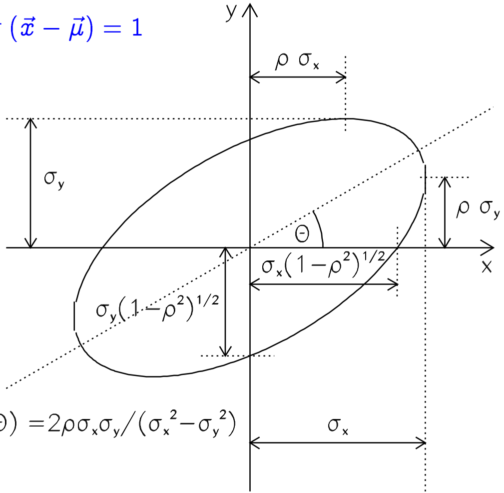
$$f(\vec{x};\vec{\mu},C)=\frac{1}{\sqrt{(2\pi)^n\det C}}\exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T\,C^{-1}\,(\vec{x}-\vec{\mu})\right]$$

- exponential of a general n-dimensional parabola
- pre-factor guarantees proper normalization
- vector of expectation values $\vec{\mu}$
- covariance matrix C
- orientation and extension of the PDF described by C
- (hyper)plannes of constant probability density are ellipsoids
- complete description in n dimensionens:
  - → n expectation values
  - → n variances (diagonal elements of C)
  - → $n(n-1)/2$ covariances
  - → in total $n(n+3)/2$ Parameter

→ 2-dim case

→ *line of constant probability density:*

$$(\vec{x}-\vec{\mu})^T\, C^{-1}\, (\vec{x}-\vec{\mu}) = 1$$



$$\tan(2\Theta) = 2\rho\sigma_x\sigma_y/(\sigma_x^2-\sigma_y^2)$$

Hauptachsen: $2\sigma^2=\sigma_x^2+\sigma_y^2\pm[(\sigma_x^2-\sigma_y^2)^2+4\rho^2\sigma_x^2\sigma_y^2]^{1/2}$

→ *basic idea:*

Study inherent statistical processes by direct simulation or map deterministic problems to statistical ones, which then are solved by simulation. The latter exploits that expectation values are defined via integrals.

→ *needed:*

Random numbers which are distributed according to well defined PDFs.

- start with random numbers with uniform distribution in the intervall $[0, 1]$.
- the derive other distributions from those

→ *technical realization: "pseudo random numbers"*

- generation via numerical algorithms
- not random, but hopefully indistinguishable from true random numbers
- reproducible sequence – important for debugging

Pseudo random number generators (i)

➜ *D.E. Knuths's 10-decimal-digits X "Super-random" number generator*

1. $Y = X/10^9$ iterate the next steps Y times
2. $Z = X/10^8$ mod 10 jump to step $Z + 3$
3. $\texttt{if}\ (X < 5 \cdot 10^9)\ \{X += 5 \cdot 10^9\}$
4. X = midsquare(X)
5. $X = (X \cdot 1001001001)$ mod 10^{10}
6. $\texttt{if}(X < 100000000)\ \{X += 9814055677\}\ \texttt{else}\ \{X = 10^{10} - X\}$
7. swap upper and lower 5-digit blocks
8. $X = (X \cdot 1001001001)$ mod 10^{10}
9. reduce every digit > 0 by 1
10. $\texttt{if}\ (X < 10^5)\ \{X = X^2 + 99999\}\ \texttt{else}\ \{X -= 99999\}$
11. $\texttt{while}\ (X < 10^9)\ \{X *= 10\}$
12. replace X by the central 10 digits of $X(X-1)$

➜ extremely complex - sequence of steps is randomized internally

➜ properties of the generator are not discernible

➜ the generator is useless: 6065038420 is a fixed point of the algorithm

lesson learned: use only generators with known properties!

➜ linear congruential generators

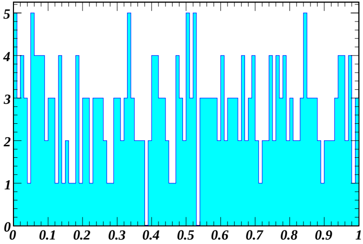
$$x_{n+1} = (a \cdot x_n + b) \bmod m$$

- ➜ multiplication with a scrambles the digits of x_n
- ➜ the constant term b prevents trivial fixed points
- ➜ mod m takes care the x stays in the range $[0, m-1]$ (x/m in $[0,1]$)
- ➜ properties/quality is determined by the parameters a, b and m

❖ study the properties of the generator for $a = 1601$, $b = 3456$ und $m = 10000$

```
seed=1601 - x=0.1601
seed=6657 - x=0.6657
seed=1313 - x=0.1313
seed=5569 - x=0.5569
seed=9425 - x=0.9425
seed=2881 - x=0.2881
seed=5937 - x=0.5937
seed=8593 - x=0.8593
seed=0849 - x=0.0849
seed=2705 - x=0.2705
etc.
```

➜ distribution of function values



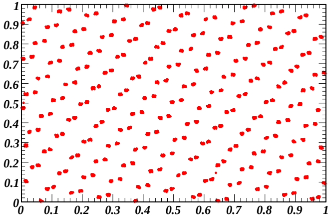
(demo 04)

➜ *result:*

- linear congruential random numbers are located on (hyper)planes
- the number of hyperplanes is a function of the plot-dimension d
- wanted: number of hyperplanes as large and period as long as possible
- both characteristics grow with the number of bits per integer
- the choice of parameters a, b, m is important, too

❖ maximum number of hyperplanes with t-bit integers: $p = (d!\,2^t)^{1/d}$

bits	$d=3$	$d=4$	$d=6$	$d=10$
t=16	73	35	19	13
t=32	2953	566	120	41
t=36	7442	1133	191	54
t=48	119086	9065	766	126
t=60	1905376	72520	3064	290



➜ need to increase the number of bits being used . . .

➜ *example for an state-of-the-art generator:* `RANLUX`

- based on the Marsaglia-Zaman algorithm
  - ➜ mathematically equivalent to a linear-congruential generator
  - ➜ completely different implementation
  - ➜ for details consult Martin Lüscher, hep-lat/9309020

implementation:

$$z_n = (a \cdot z_{n-1}) \text{ mod } m$$

with

$$m = 2^{576} - 2^{240} + 1 \quad \text{(prime)} \quad \text{and} \quad a = 2^{576} - 2^{552} - 2^{240} + 2^{216} + 1$$

❖ Discussion:

Effectively `RANLUX` uses 576-bit integer variables. The period is $\approx 5.2 \cdot 10^{171}$, and the number of hyperplanes in $d = 100$ dimensiones ist $h \approx 2000$. However, since the multiplier a has only very few bits set, subsequent 576-bit states are still correlated. As in deterministic chaos, the correlation decays exponentially with the distance of two numbers, and `RANLUX`/luxury-level 4, discards 8760 bits, before the next 576 bits are accepted.

Non-uniform random numbers (i)

➜ *generate non-uniform random numbers from uniform ones*

❖ Hit-or-Miss method

➜ algorithm to generate $\rho(x) < M$ in the intervall $[a, b]$:

(1) generate x_1 uniform in $[0, 1]$

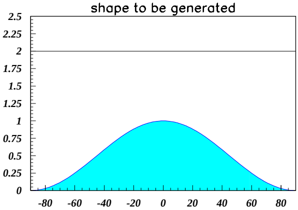
(2) scale $x = a + (b - a)x_1$

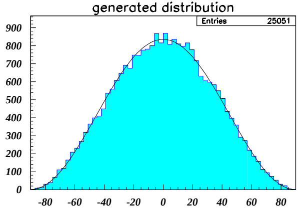
(3) generate y_1 uniform in $[0, 1]$

(4) scale $y = My_1$

(5) accept x if $y < \rho(x)$

(6) goto (1)





- properties
  - ➜ simple concept
  - ➜ normalization of ρ not necessary
  - ➜ small efficiency if $M \gg \langle\rho\rangle$
  - ➜ recycling of y_1 as next x_1 is possible

example: $\rho(x) \sim \cos^2\left(x\frac{\pi}{180}\right)$

for $-90 \leq x \leq 90$; generation with $M = 2$

➜ *The transformation method*

If x is uniformly distributed in $[0,1]$, then $y = h(x)$ is a random variable with a different distribution $g(y)$. With proper choice of $h(x)$ it should be possible to realize any distribution $g(y)$. From the section about convolutions we know:

$$g(y) = \int dx\, f(x)\delta(y - h(x)) = \int_0^1 dx\, \delta(y - h(x)) = \left.\frac{1}{h'(x)}\right|_{x=h^{-1}(y)}$$

On the other hand we have

$$h(h^{-1}(y)) = y \quad \text{and, differentiating w.r.t. } y \quad h'(h^{-1}(y)) \cdot (h^{-1}(y))' = 1$$

giving

$$g(y) = \frac{1}{h'(h^{-1}(y))} = (h^{-1}(y))' \quad \text{and finally} \quad h^{-1}(Y) = \int_{y_{\min}}^{Y} dy\, g(y)$$

i.e. the transformation h is the inverse of the integral of the target distribution g. Note also that the transformation method outlined above can be generalized to the case that h is a function of several variables.

➜ *some transformation laws for uniformly-distributed inputs $x_1, x_2, \ldots$*

$y = h(x_1, \ldots, x_n)$	➜	$g(y)$
$\sqrt{x_1}$	➜	$2y$
$-a \ln(x_1)$	➜	$\frac{1}{a} e^{-y/a}$
$\sqrt{-a \ln(x_1)}$	➜	$\frac{2}{a} y e^{-y^2/a}$
$-a \ln(x_1 x_2)$	➜	$\frac{1}{a^2} y e^{-y/a}$
$-\ln(x_1 x_2 \ldots x_n)$	➜	$y^{n-1} e^{-y}$
$\sqrt{-2\ln(x_1)} \left\{ \begin{array}{l} \cos(2\pi x_2) \\ \sin(2\pi x_2) \end{array} \right.$	➜	$\frac{1}{\sqrt{2\pi}} e^{-y^2/2}$

Verification of these relations:

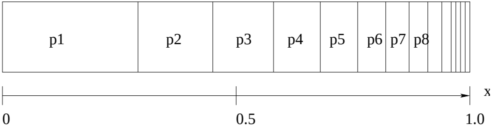
$$g(y) = \int dx_1 \cdot dx_2 \cdots x_n \, \delta(y - h(x_1, x_2, \ldots, x_n))$$

➜ Generation of Discrete Probability Distributions

Starting from the fact that the sum of discrete probabilities p_i is normalized,

$$\sum_i p_i = 1 \; ,$$

the individual probabilities can be arranged along $0 \le x \le 1$. Drawing then a uniform random number from $[0, 1]$, the interval containing the generated value x determines the discrete state to be returned, i.e. n if the interval taken by p_n is hit.



Iterative algorithm to find the hit interval:

start with $S_0 = 0$ and iterate $S_n = S_{n-1} + p_n$ until $S_n > x$.

Note that this algorithm is most efficient if the p_n are ordered in decreasing size.