# Random numbers and the central limit theorem


Alberto García

Javier Junquera

# Bibliography

No se puede mostrar la imagen en este momento.

**Cambridge University Press, Cambridge, 2002**
**ISBN 0 521 65314 2**

# Physical processes with probabilistic character

Certain physical processes do have a probabilistic character

**Desintegration of atomic nuclei:**

The dynamic (based on Quantum Mechanics) is strictly probabilistic

**Brownian movement of a particle in a liquid:**

We do not know in detail the dynamical variables of all the particles involved in the problem

We need to base our knowledge in **new laws** that do not rely on dynamical variables with determined values, but with **probabilistic distributions**

**Starting from the probabilistic distribution, it is possible to obtain well defined averages of physical magnitudes, especially if we deal with very large number of particles**

# The stochastic oracle

**Computers are (or at least should be) totally predictible**

**Given some data and a program to operate with them,
the results could be exactly reproduced**

**But imagine, that we couple our code with a special module
that generates randomness**

```
program randomness
real :: x
do
  call random_number(x)
  print "(f10.6)" , x
enddo

end program randomness
```

```
$<your compiler> -o randomness.x randomness.f90
$./randomness
```

**The output of the subroutine randomness is a real number,
uniformly distributed in the interval $[0, 1)$**

**"Anyone who considers arithmetical methods of producing random digits
is, of course, in a state of sin." (John von Neumann)**

# Probability distribution

The subroutine "random number" generates a uniform distribution of real numbers in $[0, 1)$

The probability of occurrence is given by

$$f(x) = \begin{cases} 1 \text{ if } 0 \leq x < 1 \\ 0 \text{ otherwise} \end{cases}$$

This is a continuous probability distribution.
It has to be used as a "density", i.e.,
The probability of having a number in $dx$ is $f(x)dx$

The probability of obtaining a predefined exact number is 0

# Other probability distributions

## Discrete distribution

Some probability for a set of $N$ discrete numbers

The probability of occurrence of a given number is given by

$$p = \frac{1}{N}$$

## Poison distribution

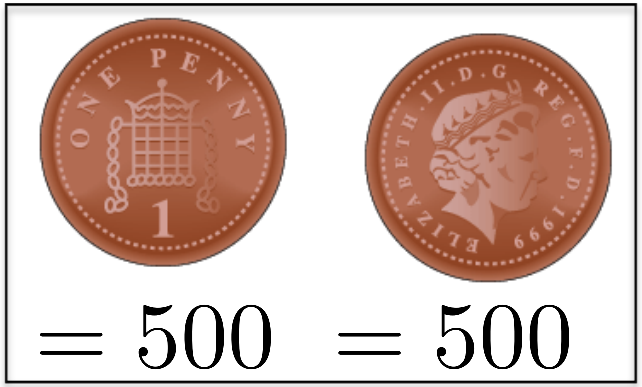
## Gaussian distribution

Important because the central limit theorem

# Jacob Bernoulli (1654-1705): Game Theory

- **<u>Simple idea</u>: lets define the outcome of an experiment $X_i$ , and the result of the experiment is either $X_i$=0 or $X_i$=1. When I repeat the experiment, I sometimes get $X_i$=0, and sometimes I obtain $X_i$=1**


- Let's think of $X_i$ as a coin, which is flipped (head/tail).
- **Experiment: flip the coin <u>1000</u> times**
- **how many "tail", how many "head"?**
- Definition of a Probability:
- **P(head) = N(head) / total** | **P(tail) = N(tail) / total**

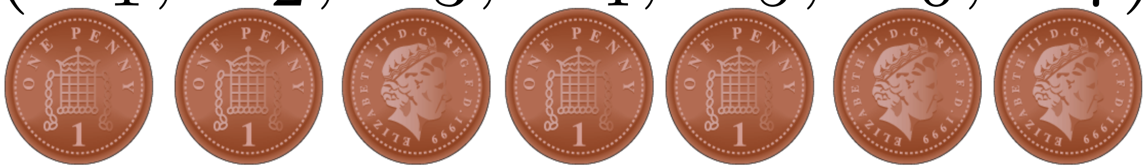



- Reference: *The Life and Times of the Central Limit Theorem*, <u>William J. Adams</u>

# Flipping many coins …

- Now, what if I throw many coins at the same time?
- Each coin is labeled with an index
- We attribute the value 0 to ***tail***, 1 to ***head***

$$\mathbf{X} = (X_1, X_2, X_3, X_4, X_5, X_6, X_7)$$



- 0 0 1 0 0 1 1
- To describe the outcome of this experiment, we define the variable $S_n$ :

$$S_n = X_1 + X_2 + ... + X_n$$

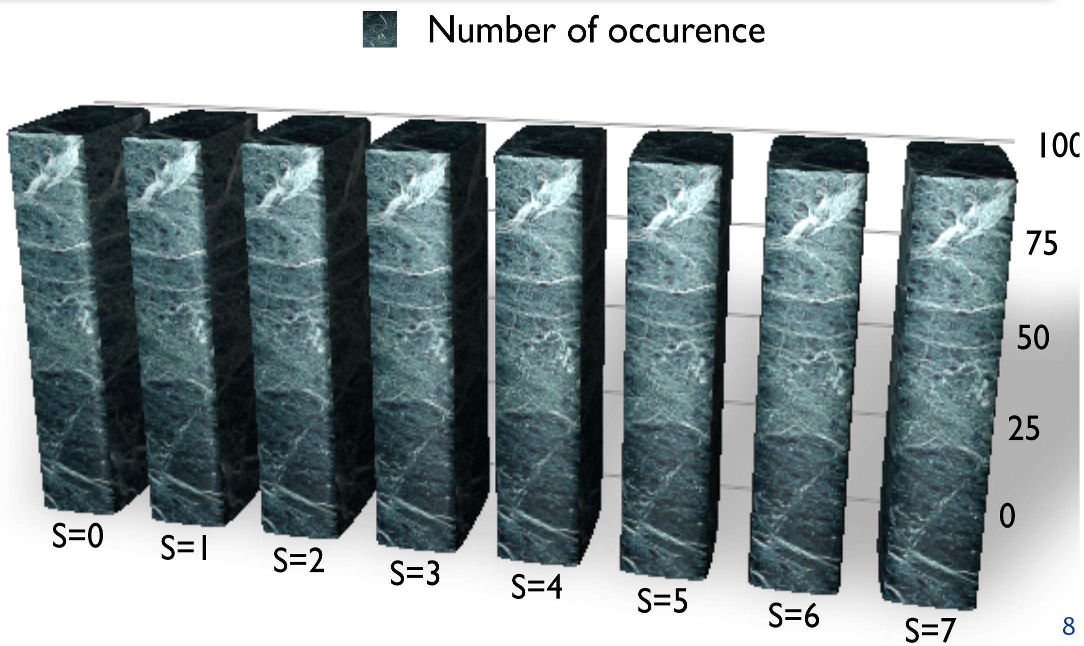
- In the experiment above I get : **$S_n$ = 3**
- Now, I throw the coins again …

- We get (fill the blank) : **$S_n$= […] ?** 1 1 1 0 0 1 0

# Central limit theorem

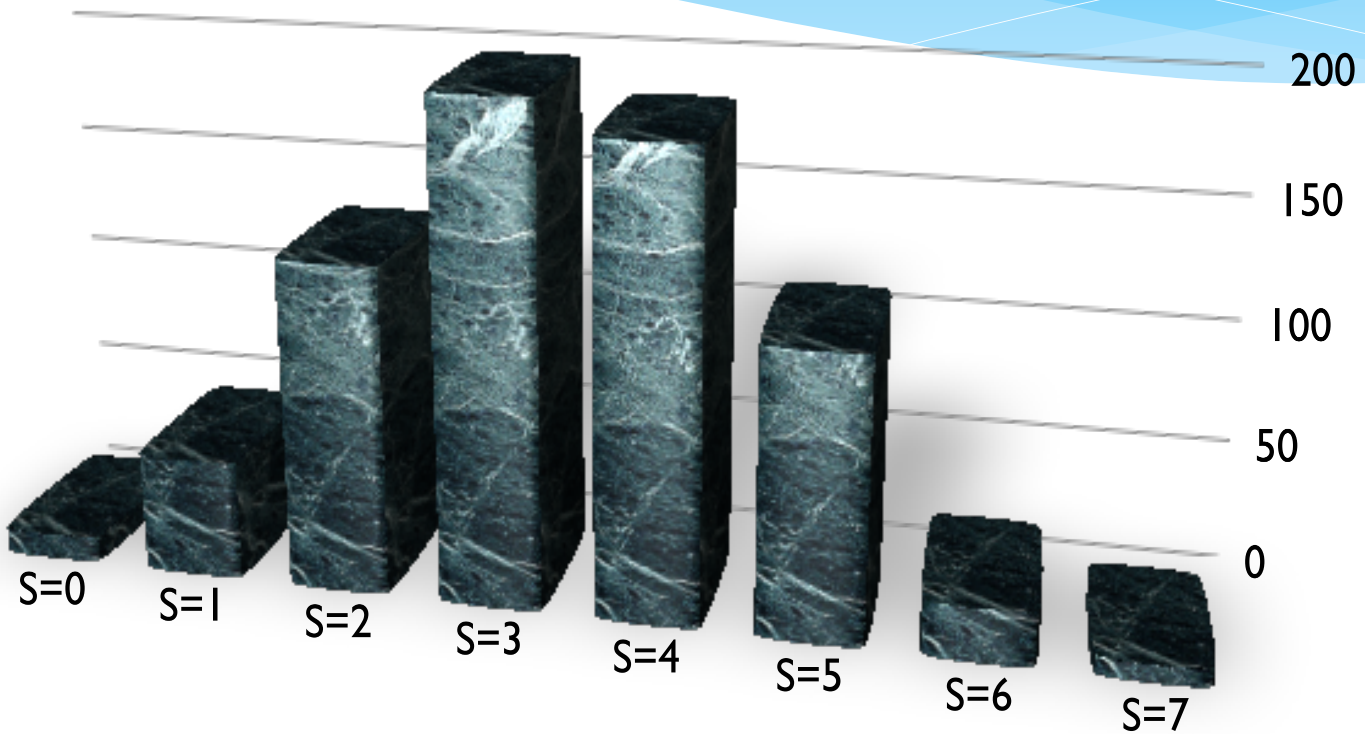
- I repeat the experiments many times … I throw 7 coins, count how many heads I obtain, I do it again
- I obtain the sequence :
  - S=3, S=4, S=1, S=7, S=4, S=3, S=4, S=5, S=4, S=0, S=2 ……………..
- I count how many times I obtain S=0, how many times I obtain S=1, … , and how many times I obtain S=7

**Question**: if I throw the coins 700 times, how many times will you obtain the result S=0? As many times as you obtain S=1, S=2, S=3, S=4, S=5, S=6 and S=7 right?



# Simulating this experiment with a computer

* This program will be written in the practice session
* Result : The number of $S_1$,$S_2$,$S_3$… $S_7$ obtained




* This histogram gives the probability P(S) to obtain S, if we divide the number of occurrence of each $S_i$ by the total number of experiments (700 in this case)
* Who can spot something wrong in the histogram ?

# Seriously?

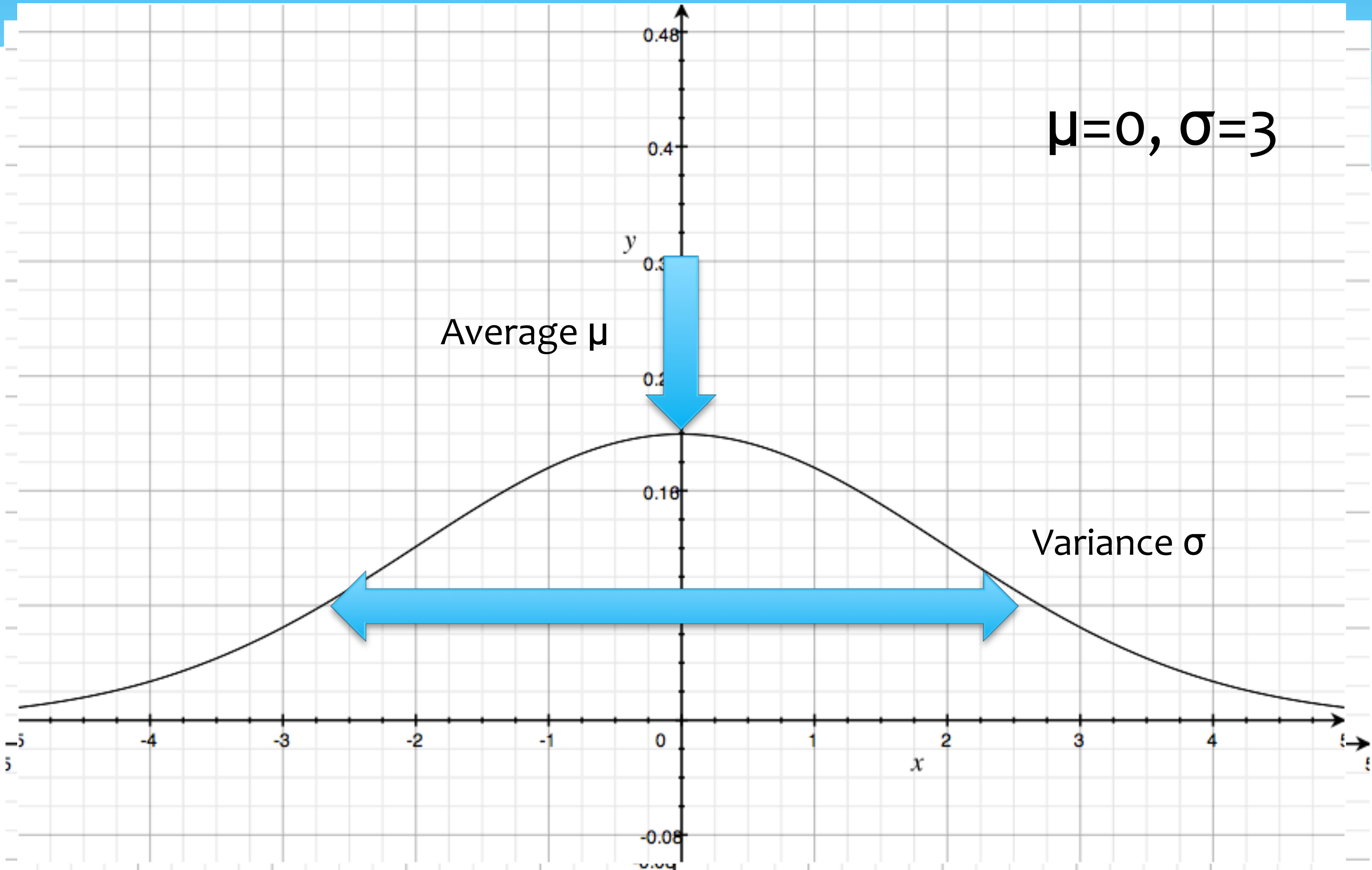
* Nope, I am not cheating you…
* Central limit theorem (CLT) :
  * the sum of a large collection of random variables ($S_n = X_1 + \ldots + X_n$) is distributed as a binomial law (or gaussian)

$$P(S) = e^{-(S-\mu)^2/(2\sigma^2)}$$

  * Where μ is called the average and σ the variance
  * First discussed by ***Abraham De Moivre*** (1667-1754)

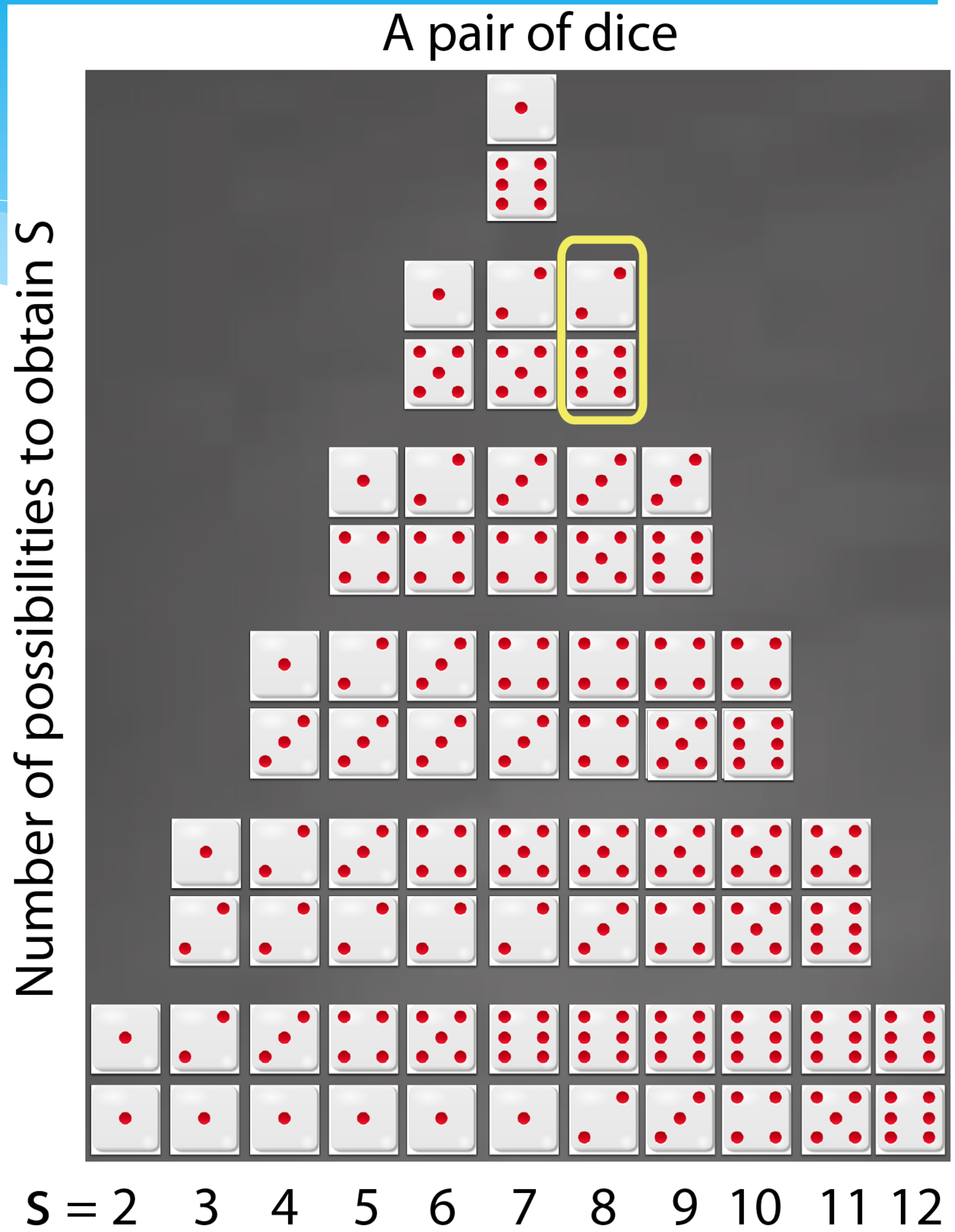
* **Physical experiment in a laboratory**: There is a large number of unknown parameters which are uncontrolled and contribute to your physical measurement (measure velocity with a timer, … )
* **CLT** : by repeating the same measurement, you will obtain a distribution of results (data), this distribution will a gaussian centered around the average value (**μ=your final measurement**) and with a given width (**σ=your error bars**)

# Gaussian function

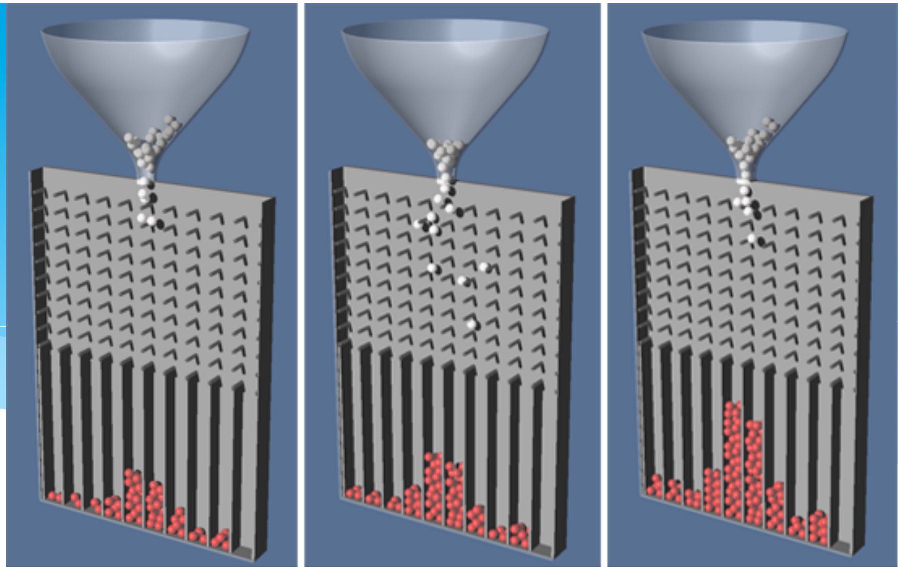


# Example 1 : gambling (Dices)

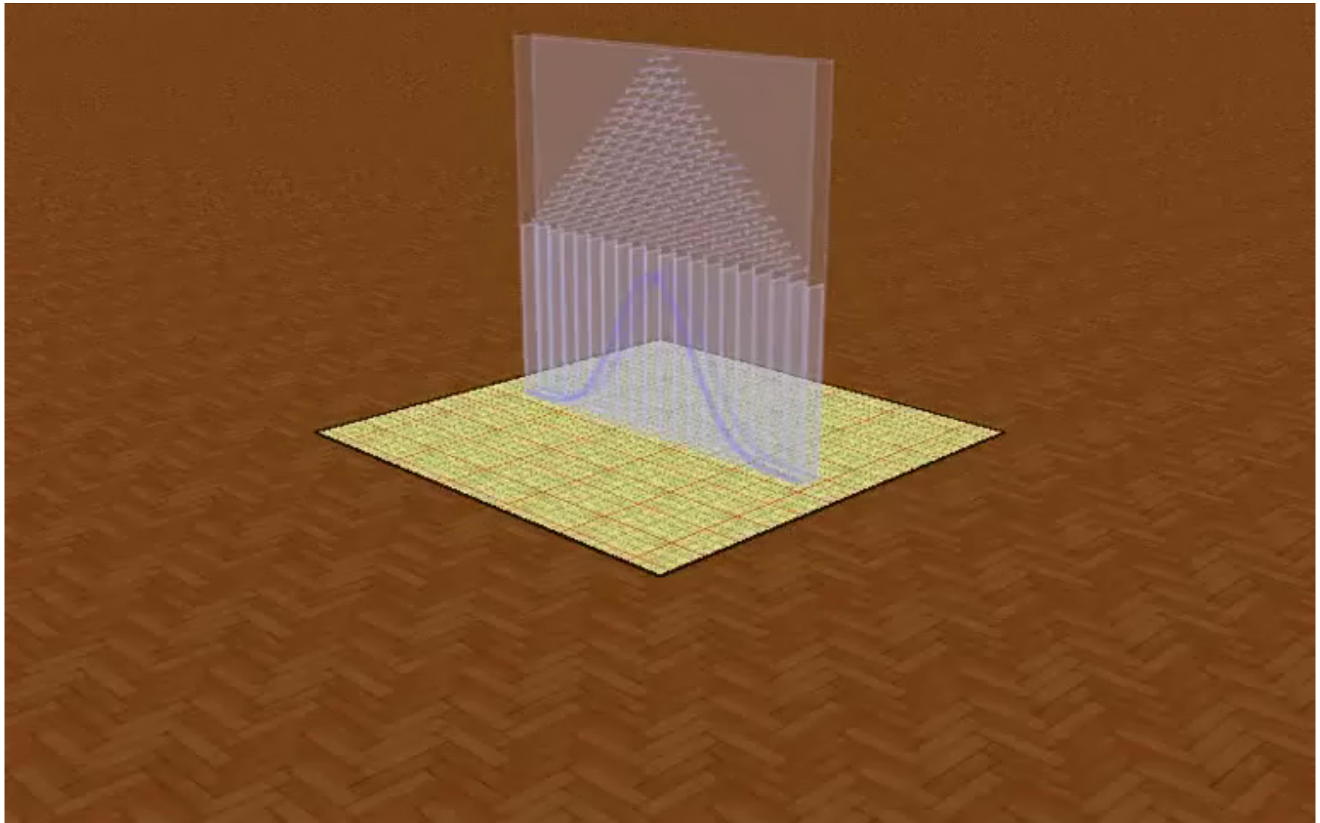
* Problem to be solved in the practice session
* Throwing to the casino a pair of dice at the same time
* Sum of a pair of dice takes values from 2 to 12
* Where do you put your bet?
  * S=2 ?
* Rolling many dices : Distribution is again gaussian



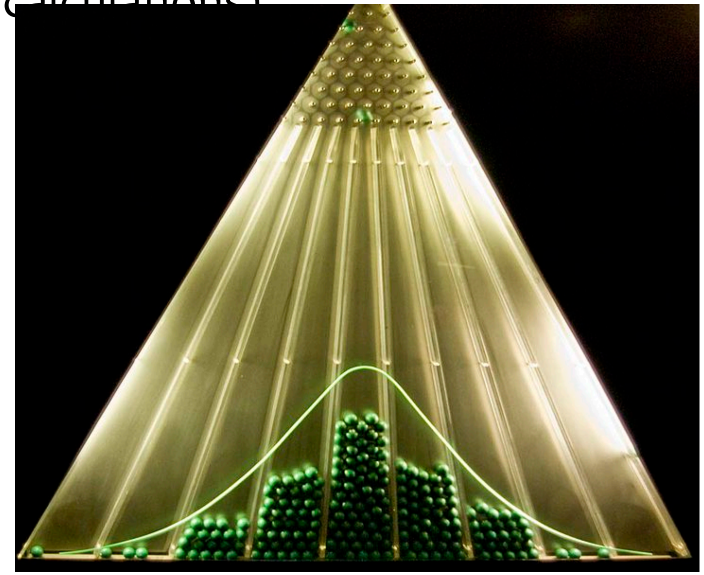
# Example 2: Galton



- ❑ Balls are bumping against many ticks during their fall
- ❑ Balls are collected at the end of the free fall
- ❑ Gaussian distribution

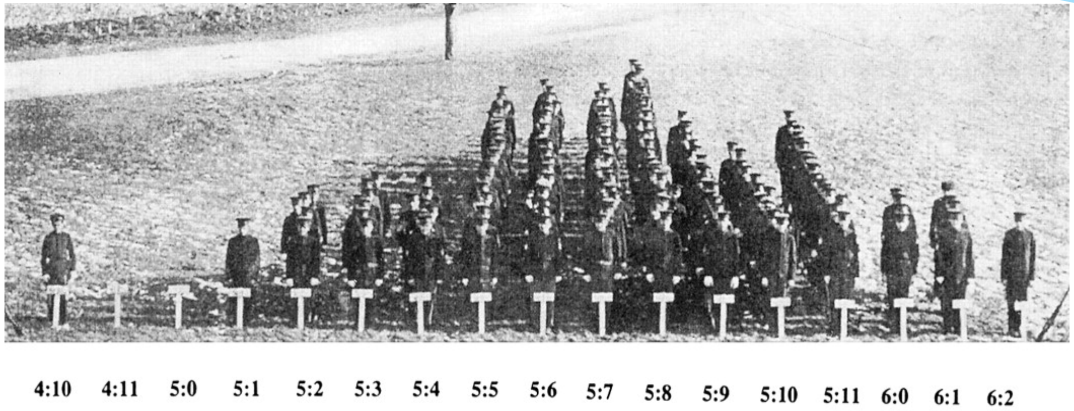


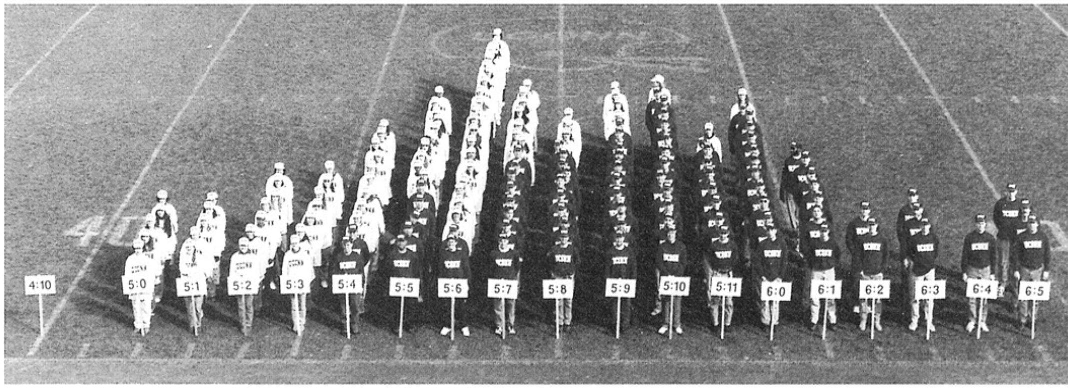
http://www.elica.net (software to perform simple model calculations)

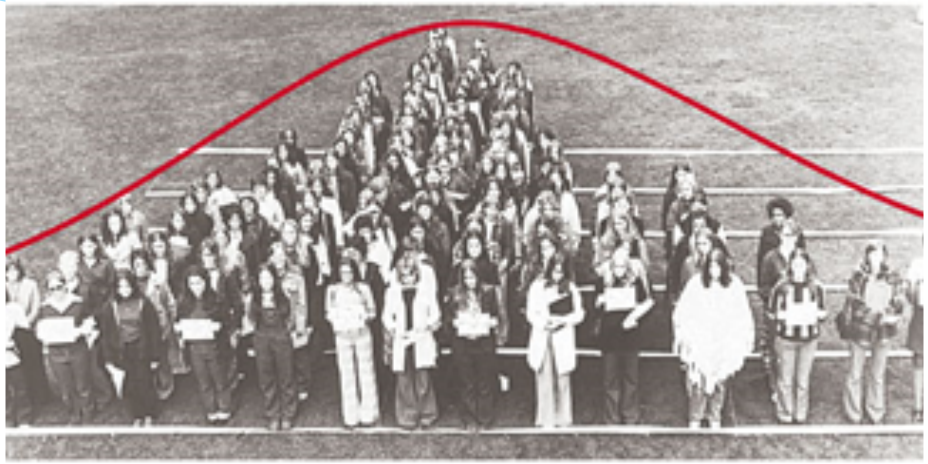


# Example 3: population height

* Distribution of people according to their height is a gaussian as well (top: college male students, bottom: men in black and women in white)



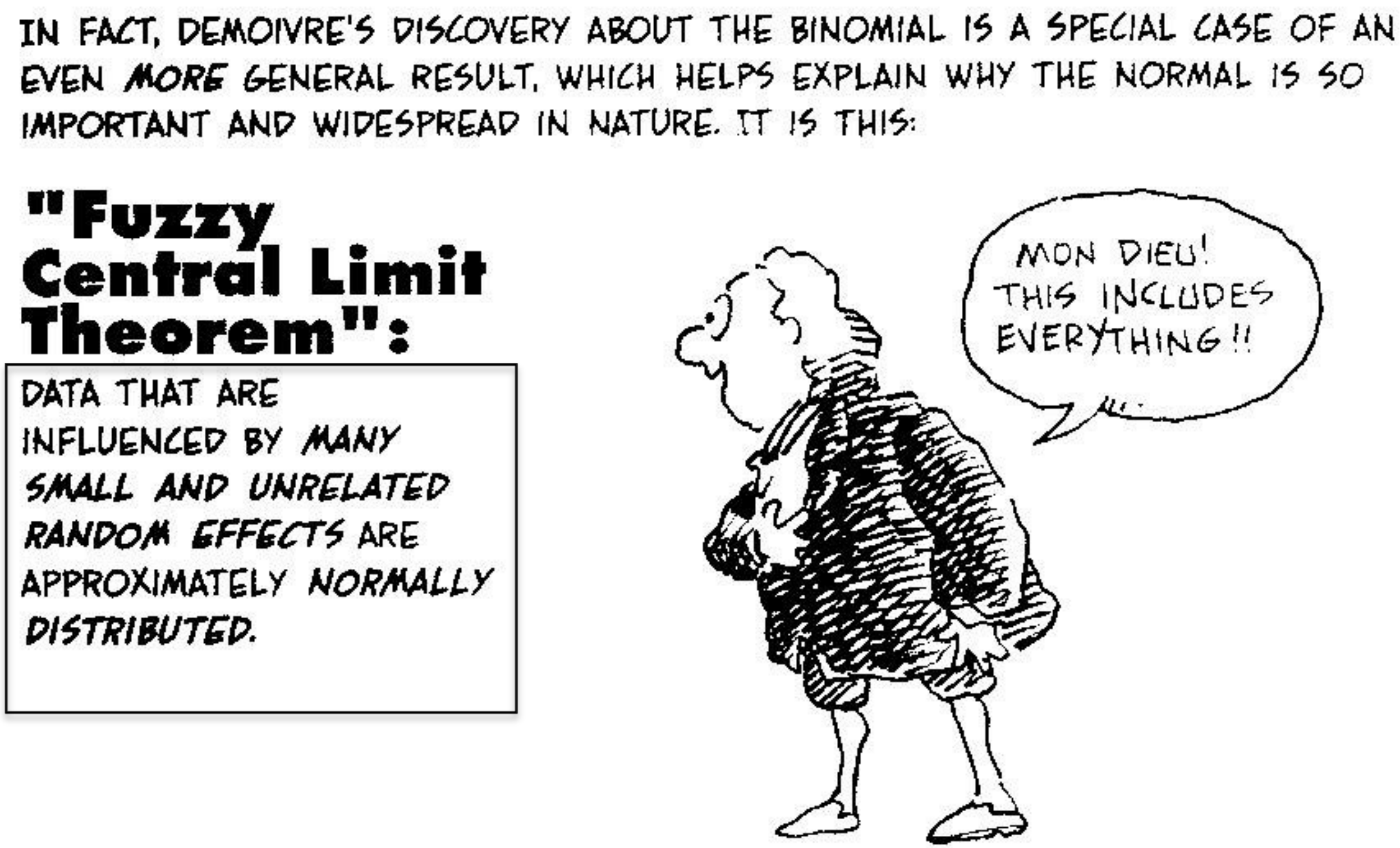


Pr Joiner's students, class of 1975, Penn state

Joiner, B. L. (1975), "Living Histograms," International Statistical Review, 3, 339–340.

Connecticut State College (*J. Heredity* 5:511–518, 1914).

# Fuzzy CLT …

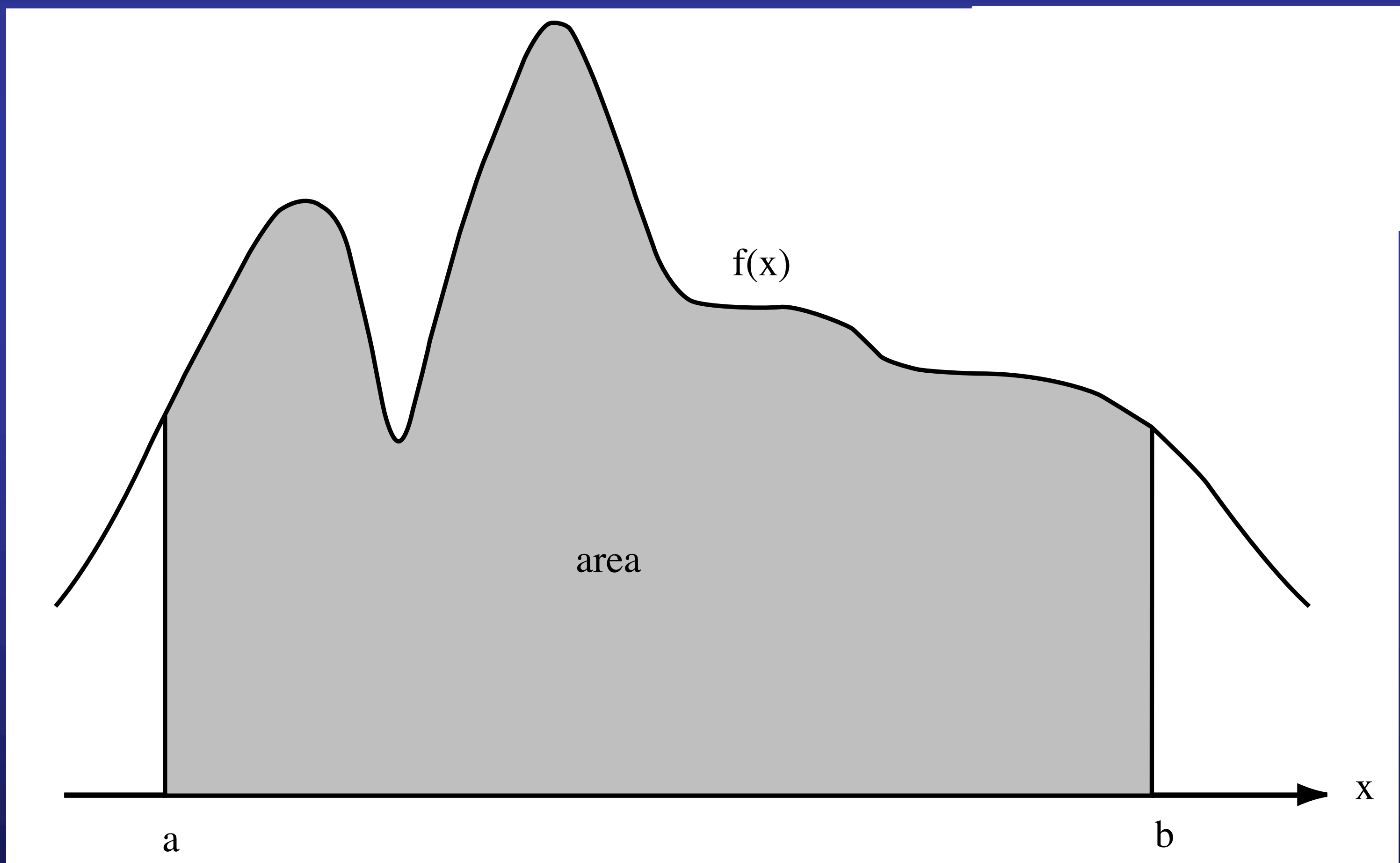



* Larry Gonick, *The Cartoon Guide to Statistics*, New York, NY : Collins Reference / HarperPerennial 1993, p. 83.

# Integration of functions

**Standard example of the use of stocastic methods in Applied Mathematics**

**Problem: compute** $\int_a^b f(x)dx$



# Method 1 : Discretization

* Discretize the horizontal axis, approximate the surface enclosed by the function f(x) by a set of rectangles

* Example: **f(x) = cos(x)**

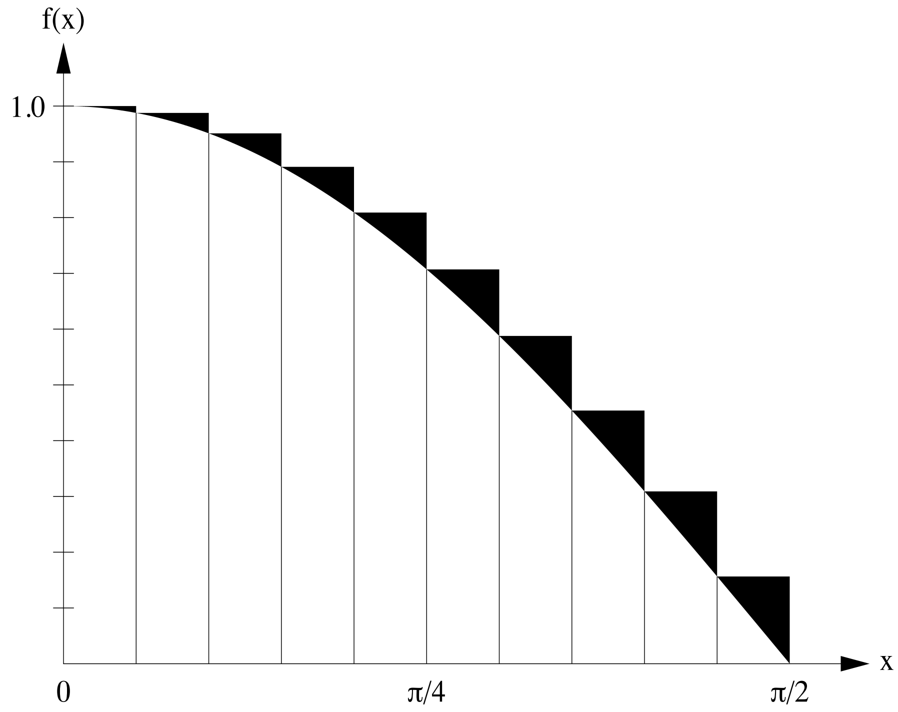
* width of a rectangle:

* Discretization x: $$\Delta x = \frac{b-a}{n}$$

* Area=sum of rectangles:

$$x_n = x_0 + n\Delta x$$

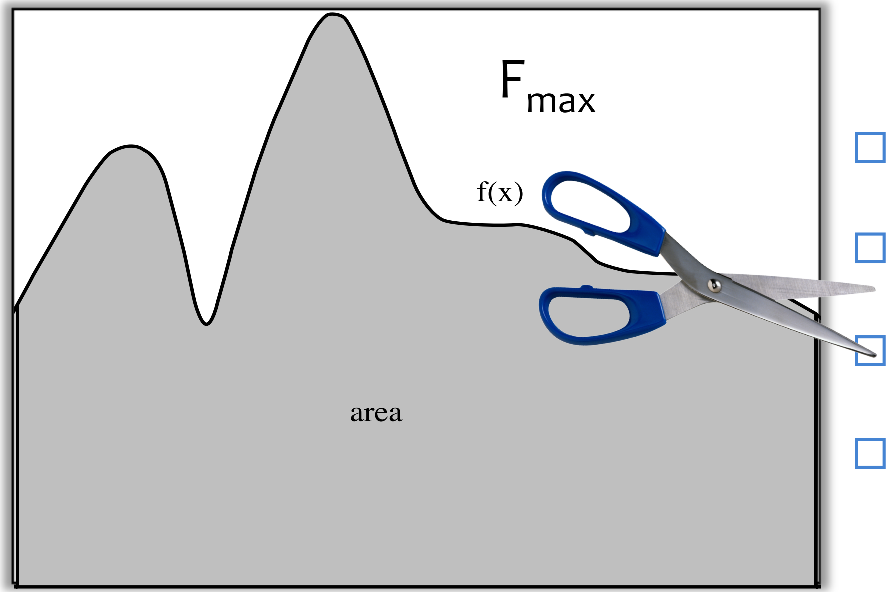
$$F_n = \sum_{i=0}^{n-1} f(x_i)\Delta x$$



# Integration of functions

## Method 2 : cutting and weighting

- Plot the function f(x) on a piece of paper, from **a** to **b**, and from **0** to $\mathbf{F_{max}}$
- Weight the full sheet of paper, we define this as **Ws**
- Plot the function on a sheet of paper
- Cut the sheet in paper along the function
- Weight the sheet of paper contained between the horizontal axis and the function f(x), we define this as **Wf**




- ρ : paper weight density
- **Ws** = (b-a) x $F_{max}$ x ρ
- **Wf** = integral x ρ
- Integral = **Wf/Ws** x (b-a) x $F_{max}$

# Integration of functions

**Method 3: estimation from the average of the function between $[a, b]$**

$$\int_a^b f(x)dx = (b-a)\langle f \rangle$$

**Assume that we choose $N$ points $\{x_i\}$ randomly distributed between $[a, b]$, then**

$$\langle f \rangle = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

**The larger the number of points, the better the estimation.**
**The typical error $\Delta$ in $\langle f \rangle$, in the sense that 63% of the estimations of $\langle f \rangle$ will be between $\langle f \rangle + \Delta$ and $\langle f \rangle - \Delta$**

$$\Delta = \sqrt{\frac{1}{N}\left(\langle f^2 \rangle - \langle f \rangle^2\right)}$$

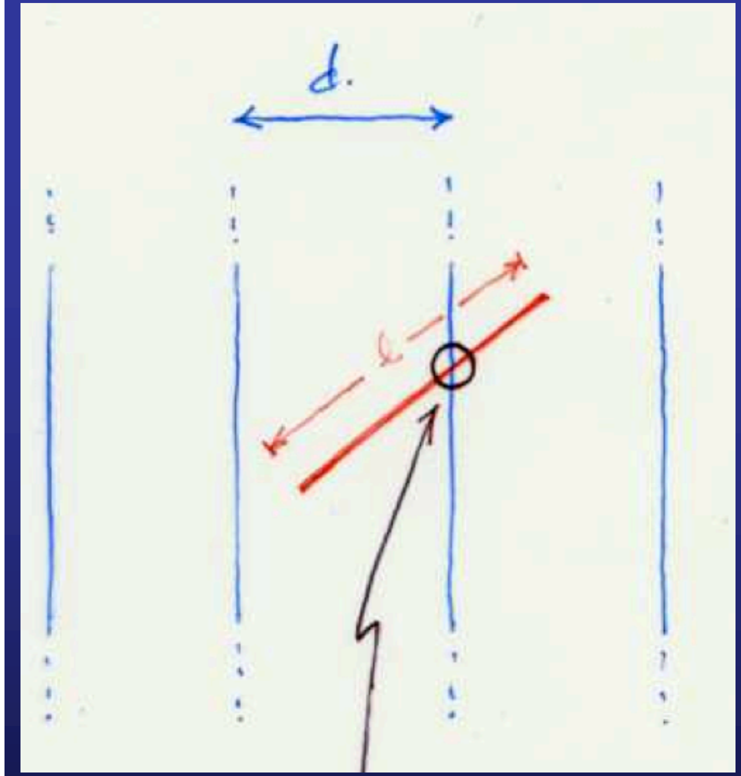
**where** $\langle f^2 \rangle = \frac{1}{N}\sum_{i=1}^{N} \left[f\left(x_i\right)\right]^2$

# Example of integration:
# How to estimate $\pi$ with a needle

**Buffon's experiment**

**If a needle of length $l$ is thrown at random onto a set of equally spaced parallel lines, $d$ apart (where $d > l$ ), the probability of the needle crossing a line is**

$$\mathcal{P} = \frac{2l}{\pi d}$$



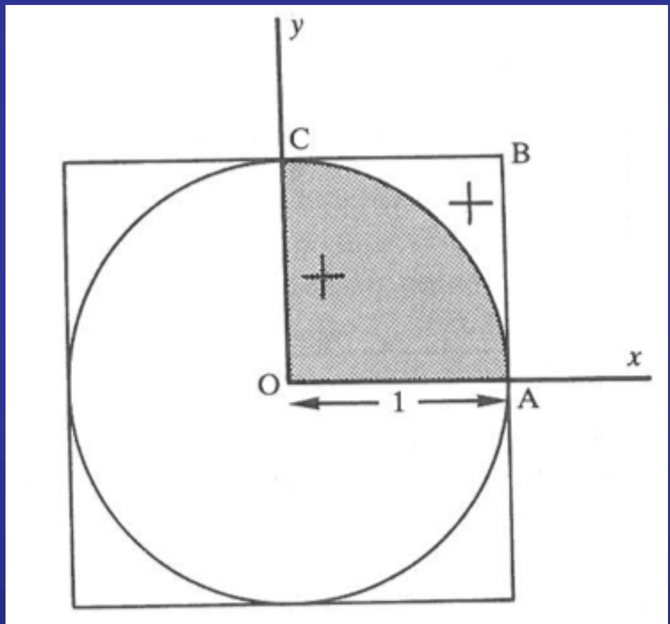

**Lazzerini (1901):**
**Spinning round and dropping a needle 3407 times:**

$$\pi \approx 3.1415929$$

# Example of integration:
# How to estimate $\pi$ in a rainy day

**Let us rephrase the problem:**
**how to estimate the area of a circle of radius 1**

**A circle centred at the origin and inscribed in a square**




**A number of trial shots are generated in the square OABC**

**At each trial two independent random numbers are chosen from a uniform distribution on $(0, 1)$**

**These numbers are taken as the coordinates of a point (marked as + in the figure)**

**The distance from the random point to the origin is calculated**

**If the distance is less or equal to one, the shot has landed in the shaded region and a hit is scored**

**If a total of $\tau_{\text{shot}}$ are fired and $\tau_{\text{hit}}$ hits scored, then**

$$\frac{\tau_{\text{hit}}}{\tau_{\text{shot}}} = \frac{\text{Area under the curve CA}}{\text{Area of the square OABC}} = \frac{\pi R^2/4}{R^2} = \frac{\pi}{4}$$

$$\pi \approx \frac{4 \times \text{Area under the curve CA}}{\text{Area of the square OABC}} = \frac{4\tau_{\text{hit}}}{\tau_{\text{shot}}}$$

# Example of integration: How to estimate $\pi$ in a rainy day

**In mathematical words, we have estimated the integral**

$$\int\int f(x,y)dxdy$$

**where the function** $f(x,y)$ **is**

$$f(x,y) = \begin{cases} 1 \text{ if } x^2 + y^2 \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

**In the two-dimensional interval**

$$[0,1] \times [0,1]$$

# Example of integration:
# How to estimate $\pi$ from the hit and miss method

```
program pi
!
! Compute pi by the hit and miss Monte Carlo method
!
real :: x, y
integer :: n, i
integer :: sum
print *, 'Number of points to use?'
read *, n
sum = 0.0
do i = 1, n
   call random_number(x)
   call random_number(y)
   if (x*x+y*y <= 1.0) sum = sum + 1
enddo
print *, 'pi = ', 4.0 * sum / real(n)
end program pi
```

# Example of integration:
# How to estimate $\pi$ in a rainy day or with a needle

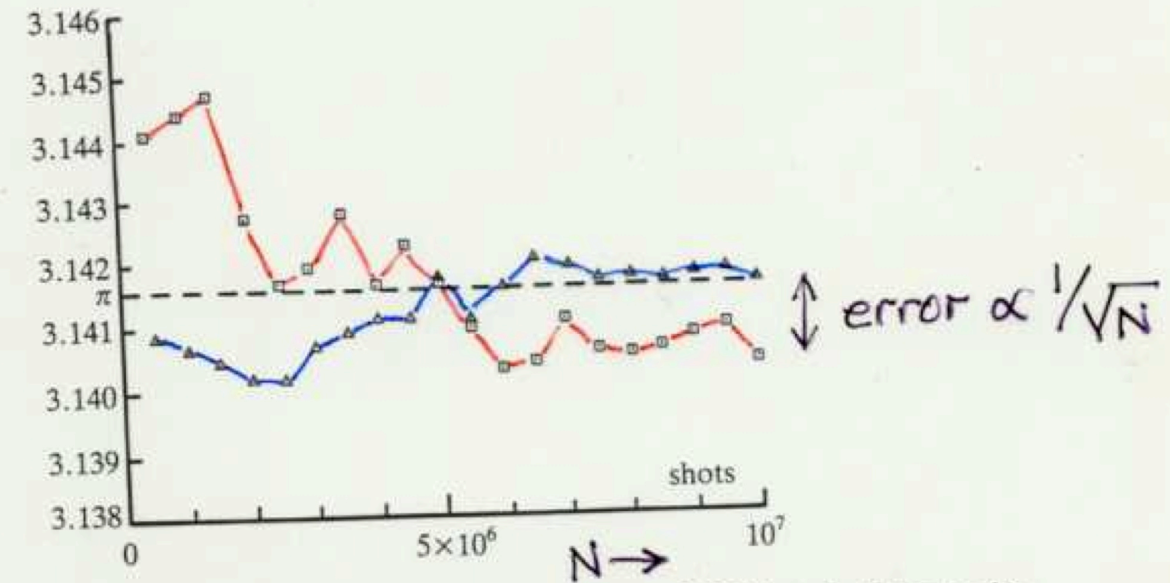



**Fig. 4.2** The cumulative estimate of $\pi$ as a function of the number of MC shots by hit-and-miss area of a circle (triangles) and the Buffon needle experiment (squares).

**To gain one significative figure (i.e. to reduce the typical error by one order of magnitude) we should to increase the number of integration points by 100**

# Why Monte Carlo?

***"Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse."*** Alan Sokal

*Monte Carlo methods in statistical mechanics*, 1996

The **<u>error</u>** is only shrinking as $1/\sqrt{N}$

Other simple methods for integrations in two dimensions: $1/N^{3/2}$

In more than two dimensions (2D), e.g. in three dimensions (3D) (volume of a sphere,… ):

$$1/N^{3/D}$$

<u>Which method is better for **D=10** ?</u> Answer = … [ FILL IN] ….

# Applications to Statistical Mechanics

**Fundamental postulate of Statistical Physics:**
**All the microstates of a closed system in equilibrium are equally probable**

**Closed means that the total energy $U$, the number of particles $N$, and the volume $V$ are constant**

**In thermodynamics, these are the natural variables in the entropic representation**

$$S = S\left(U, V, N\right)$$

**The fundamental connection between Statistical Mechanics and Thermodynamics**
**The entropy can be computed from the total number of microstates of a system**

$$\Omega = \Omega\left(U, V, N\right)$$

$$S\left(U, V, N\right) = k_{\mathrm{B}} \ln \Omega\left(U, V, N\right)$$

# Applications to Statistical Mechanics


**Boltzmann's grave in the Zentralfriedhof, Vienna, with bust and entropy formula**

## Any macroscopic physical quantity can be computed as an statistical average over accesible microstates

$$M = \frac{1}{\Omega}\sum_{j=1}^{\Omega} M_j$$

**Therefore, to study the properties of any closed macroscopic system in equilibrium, it should be enough (in principle) to determine all the microscopic states and evaluate the corresponding averages taking only those microstates compatible with the thermodynamic variables $U$, $N$, and $V$**

**In practice, it is impossible to find all the microscopic states available**

## Any macroscopic physical quantity can be computed as an statistical average over accesible microstates

$$M = \frac{1}{\Omega}\sum_{j=1}^{\Omega} M_j$$

**Therefore, to study the properties of any closed macroscopic system in equilibrium, it should be enough (in principle) to determine all the microscopic states and evaluate the corresponding averages taking only those microstates compatible with the thermodynamic variables $U$ , $N$ , and $V$**

**In practice, it is impossible to find all the microscopic states available**

# Number of possible microstates might be inconceivable large

Let us imagine a system of $N$ spins with two possible states (up or down).
How many microstates are possible?

$$2^N$$

With a few dozens of particles, this number might be very, very large

It is impossible to compute these averages, $M = \frac{1}{\Omega}\sum_{j=1}^{\Omega} M_j$ , exactly

But we can estimate them by a partial sampling of all the possible microstates, in the same way as sociologist prepare the poll.
We should employ a non-bias method to sample the configuration space

# The Metropolis algorithm

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

## Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
*Los Alamos Scientific Laboratory, Los Alamos, New Mexico*

AND

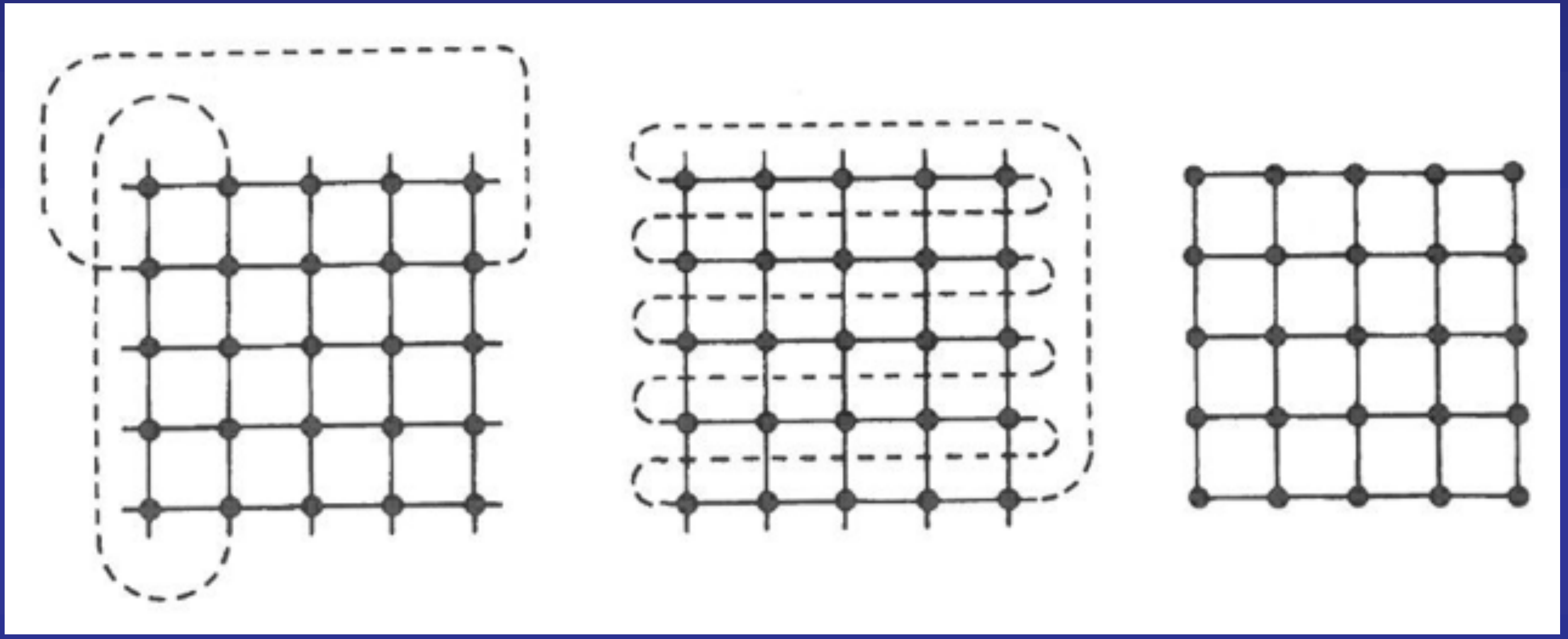
EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

# Boundary conditions

**Typical boundary conditions for the two-dimensional Ising model**



Periodic boundary contsitions

Screw periodic

Free edges