

# Statistics for Modern Quantitative Biology

Monday, November 17, 2014

IBt, UNAM, Cuernavaca

## Reading

Whitlock & Schluter. The analysis of biological data.  
Roberts and Company Publishers

Siegel & Castellan. Nonparametric Statistics for the Behavioral  
Sciences. McGraw-Hill

Cobb. Introduction to design and analysis of experiments.  
Springer

Crawley. Statistics. An introduction using R.  
John Wiley & Sons

We do not “see” causes directly

(A bit of epistemology)

# Natural Sciences

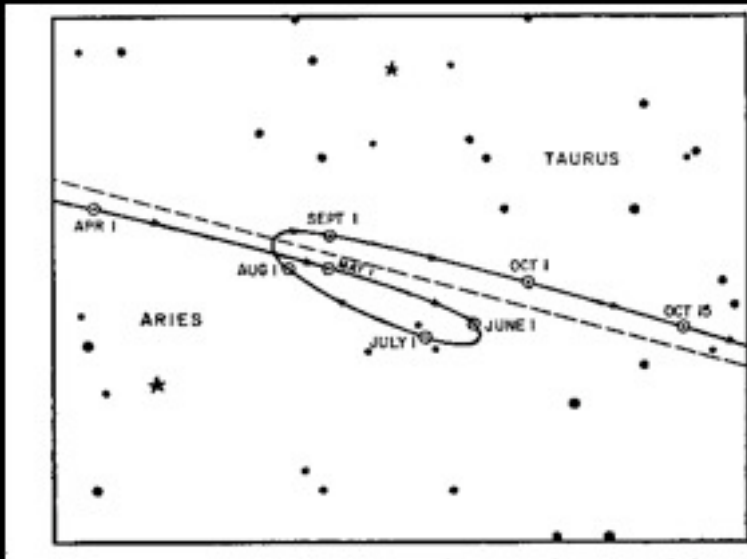
Explaining Nature

Explaining biological systems

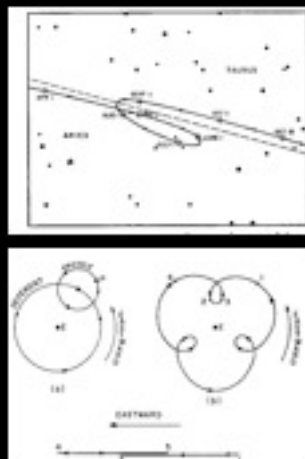
Expectation and Prediction

Statistical inference and causal inference

Ptolemaic astronomy and Newtonian astronomy



## Ptolemy deferents and epicycles



## Isaac Newton



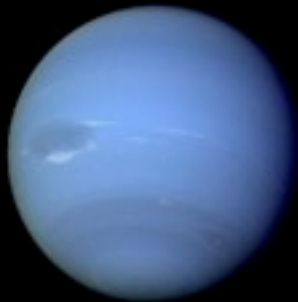
# Isaac Newton

*Philosophiæ Naturalis Principia Mathematica*

$$F = G \frac{m_1 m_2}{r^2}$$

$$F = ma$$

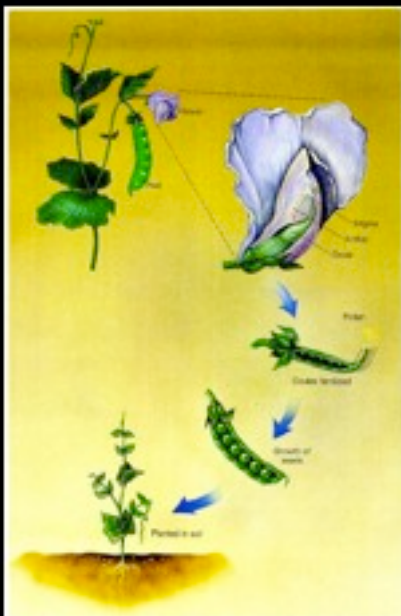
"After the discovery of Uranus, it was noticed that its orbit was not as it should be in accordance with Newton's laws. It was therefore predicted that another more distant planet must be perturbing Uranus' orbit. Neptune was first observed by Gale and d'Arrest on 1846 Sept 23 very near to the locations independently predicted by Adams and Le Verrier from calculations based on the observed positions of Jupiter, Saturn and Uranus."



Is there an essential difference between the Ptolemaic and the Newtonian explanations of the trajectory of the planets in the sky?

## Mendelian genetics and Modern synthesis

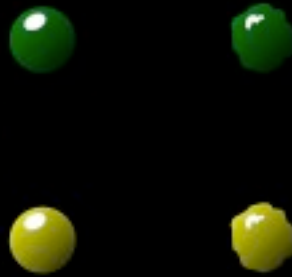
**Gregor Mendel**  
(1822-1884)



## Traits or phenotypes

Shape

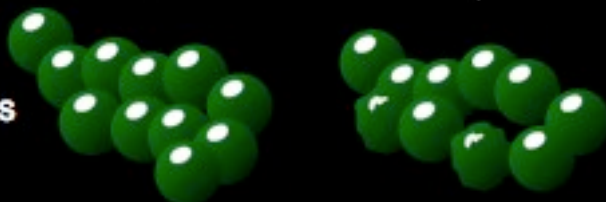
Colour



One seeds...



and collects



Parental generation

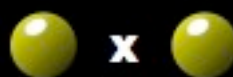


x

First generation



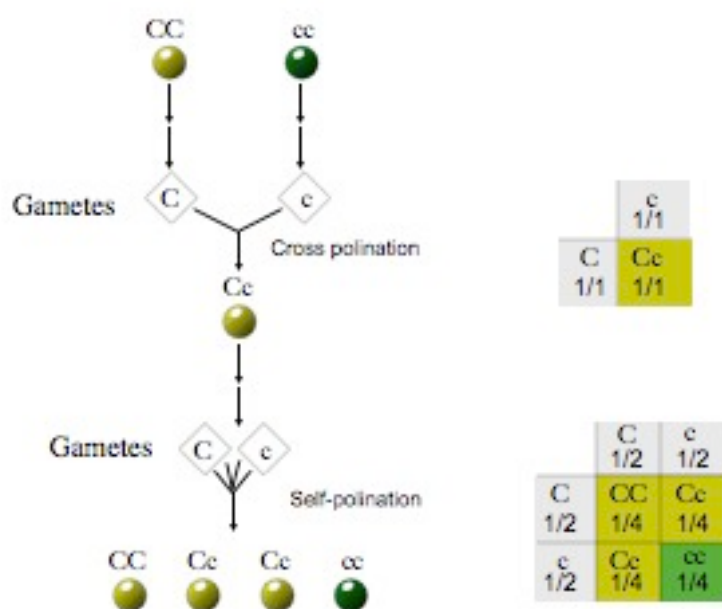
crossing



x

Second generation





## Reification of the gene

Morgan chromosomes

Muller X-ray gene mutation

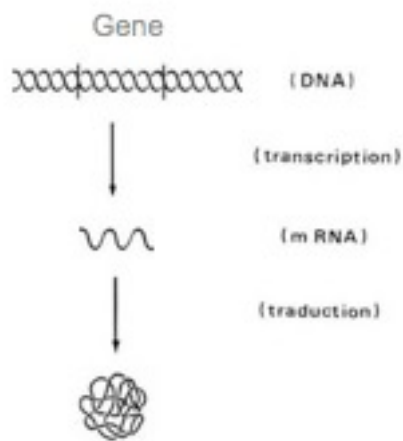
Delbrück variants precede selection

James Watson & Francis Crick DNA structure

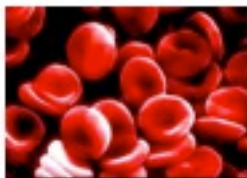
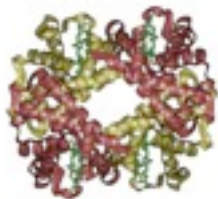
## Reification of the gene



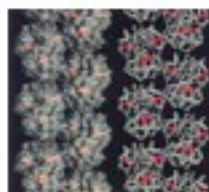
<http://abs.bio.unc.edu/Salmon/mitosis/phmit1.mov>






MVNHLTPDKSAVTALWGKVNVDVCGHIALGRLLV  
 VYPWTQRFESFGDLSTPDAMGNPKVKAHGKKV  
 LGAPSDGLAHLNLNKGTTATLSIELHCDKLIVDPEN  
 FRLLGNVLVCLAHHPGKEFTPPVQAAYQKVVAG  
 VANALAHKKYH



MVHETPVEKSAVTALWGKYNVDIEVGGEALGRLLVY  
 YPWTRQFFESFGDLSTPDNVMMGNPKVKAHGKKVVG  
 APSDGLAHLINLKGTPATLSIEICDKLIVDPENIRLL  
 GNVLNVLAIHHFGKEFTPPVQAAYQKVVAGVANAL  
 AHKYIH





Genotype	Cellular phenotype	Individual phenotype
$Hb^aHb^a$		Normal
$Hb^aHb^s$		Normally no or mild anemia
$Hb^sHb^s$		Severe anemia

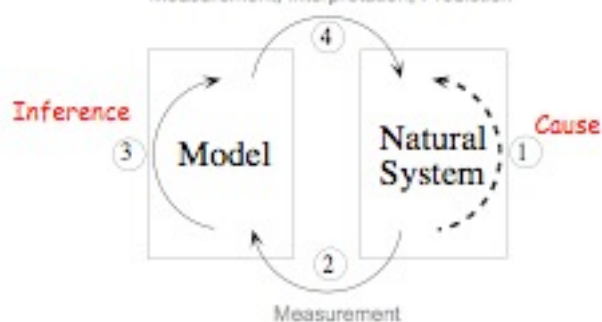
## Modelling Nature, Natural Law

At least two distinct levels of observation

"Structure and function"

### The Modelling Relation

Measurement, Interpretation, Prediction



[After: Rosen (1991) *Life Itself*. Columbia University Press. New York]

Inferential structure is mapped to causality structure

# Models in Biology

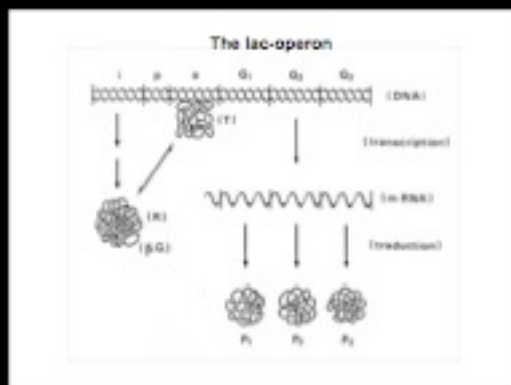
Conceptual, Propositional Models in Natural Language

(Cartoons)

Experimental models

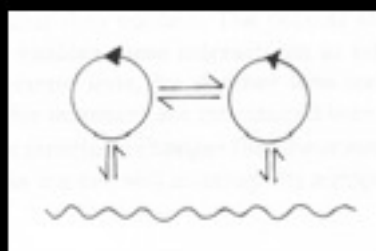
Mathematical Models

## Cartoons



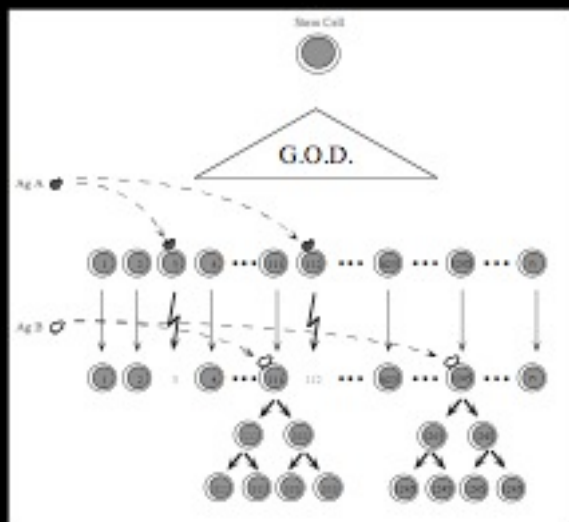
Monod's "model"

## Cartoons



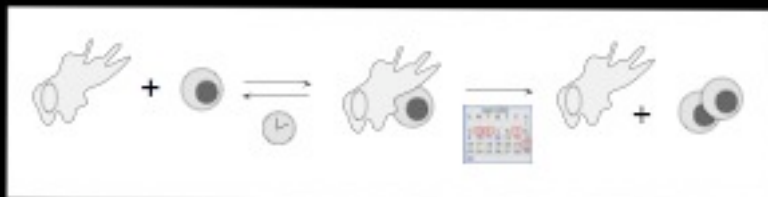
Maturana & Varela

# Cartoons



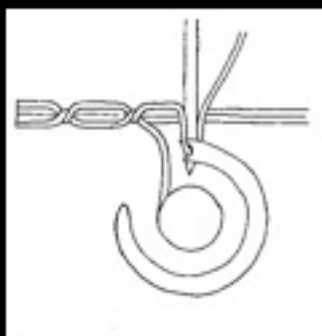
# Cartoons

What about time ?



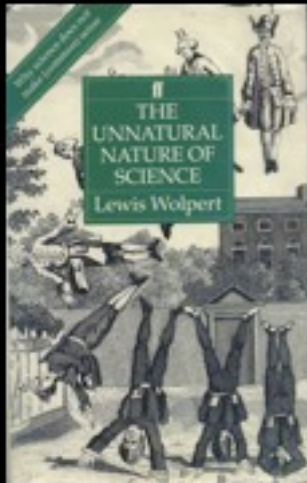
# Cartoons

What about dynamics ?



Why are mathematical descriptions and models better ?

Through mathematics one can be cold-bloodedly objective



"It is often held that science and common sense are closely linked. Thomas Henry Huxley, Darwin's brilliant colleague, spoke of science as being nothing more than trained common sense. (...) However reasonable they may sound, such views are, alas, quite misleading. In fact, both the ideas that science generates and the way in which science is carried out are entirely counter intuitive and against common sense. (...) Science does not fit with our natural expectations."

— Lewis Wolpert

Earth moves around the Sun

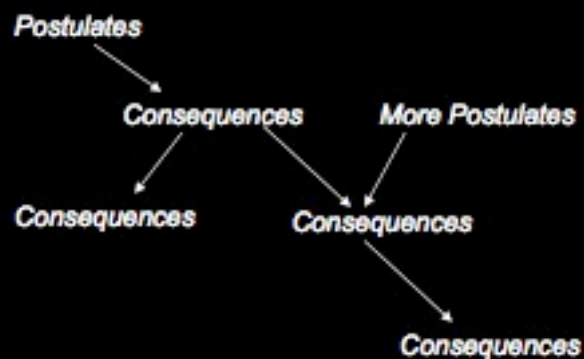
Earth is round

*"If one bullet is dropped from your hand and another is fired horizontally from a gun at exactly the same time, which will hit the ground first?"*



Uniform motion without force (Newton's first law)

Mathematics is generative



... consequences of consequences of consequences...

## Mathematics in communication and intersubjectivity

"Then, 11 months later (ample time for chimerism to develop), we measured chimerism in the blood (...). Unexpectedly, there were no donor-derived B cells and seldom any APCs; the chimeric cells were almost entirely T-cell receptor (TCR)ab<sup>+</sup>, CD4 or CD8 T cells (called 'chimeric T cells' here) (...).

A catadysmic inflammatory event, for example, from a microbial incursion, might initiate an autoimmune disease process, and any readily processed determinant which gains ascendance in either the class I or class II presentation systems will stimulate ambient T cells. In this case, the first T cell to be activated will either be the most abundant or have the most avid receptors."

$$\frac{dx}{dt} = a \cdot x - b$$

$$\dot{x} = a \cdot x - b$$

## Occam's Razor

*"Plurality non est ponenda sine necessitate (Plurality should not be posited without necessity)"*

William of Occam, XIV century

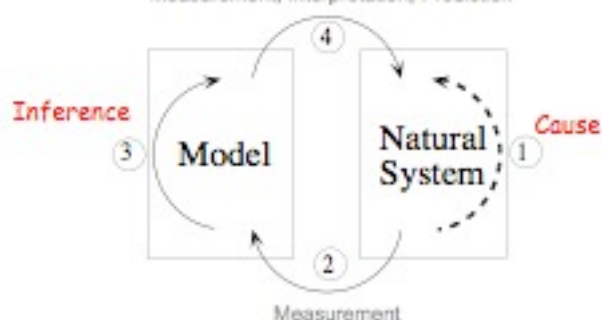
*"Everything should be made as simple as possible,  
But not simpler"*

Albert Einstein, XX century

*Complicate only when simplicity fails*  
JC

## The Modelling Relation

Measurement, Interpretation, Prediction



[After: Rosen (1991) Life Itself. Columbia University Press. New York]

Inferential structure is mapped to causality structure

## Stephen Jay Gould



see also: Lewis et al. The Mismeasure of Science: Stephen Jay Gould versus Samuel George Morison on Skulls and Race. *PLoS Biol* (2011) 9:e1001071

## Stochastic processes, probability, and the central limit theorem

Random events

Probability and probability distributions

Normal distribution and the central limit theorem

## Randomness rules the world

*Fortuna Imperatrix Mundi*

What is a random event ?

What is Probability?

Probability is a numerical measure of **uncertainty** about an event



## Laplace's Concept

$$probability = \frac{\# \text{ favorable events}}{\# \text{ all possible events}}$$

Problems:

Every event is equally likely to occur

Finite number of events



Pierre Simon Laplace  
(1749-1827)

## Frequency-based Concept

It is the relative frequency of occurrence of an event if an experiment is repeated *ad infinitum*.



J. Bernoulli (1654-1705)

## Classical Statistics



R. A. Fisher (1860-1962)



K. Pearson (1857-1936)

## Subjective Concept

It is a measure of the degree of belief of an individual on the occurrence of a certain event.



Rev. Thomas Bayes (1701-1761)

## Bayesian Statistics



J. Haldane (1852-1934)



H. Jeffreys (1881-1949)

## Random Variable

### Random experiment:

Every experiment in which the outcome is not deterministic.  
e.g. tossing a dice, inheritance of a paternal allele

### Event space :

The set of all possible outcomes of the random experiment.  
tossing a dice:  $\{1, 2, 3, 4, 5, 6\}$   
paternal allele  $\{ \text{grandfather allele}, \text{grandmother allele} \}$

### Random variable X:

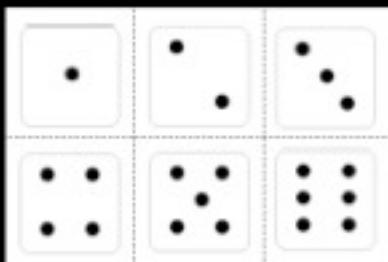
A mapping that assigns a probability to every possible outcome.

tossing a dice :  $P[X = x] = 1/6$ , where  $x = 1, 2, 3, 4, 5, 6$

A few properties of probability  
and  
it's computation

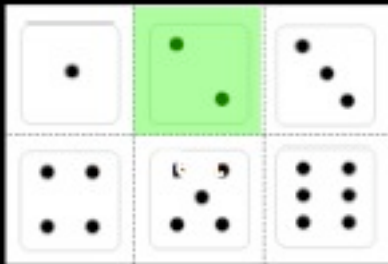


Probability of an event



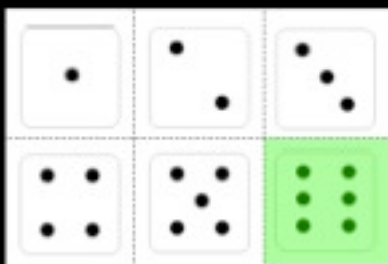
(Venn diagram)

Probability that the "outcome is 2"



$$P(X=2) = 1/6$$

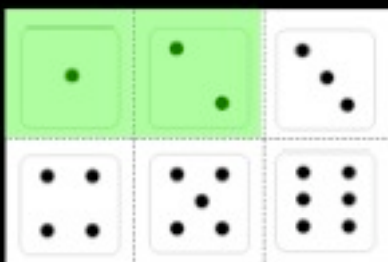
Probability that the "outcome is 6"



$$P(X=6) = 1/6$$

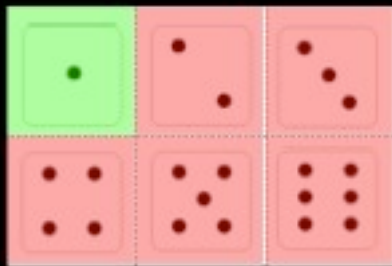
Addition of probability of mutually exclusive events:  
either this or that

$$P[X==1 \text{ AND } X==2] = 0$$



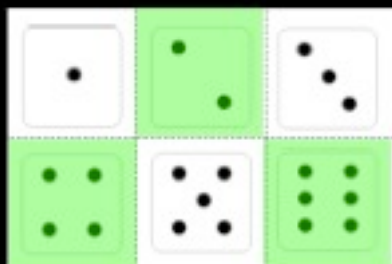
$$P(X=1 \text{ OR } X=2) = P(X=1) + P(X=2) = 2/6$$

Probability of all possible mutually exclusive events adds to 1



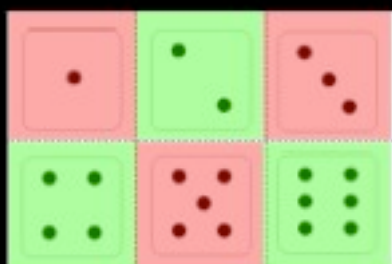
$$P(X \neq 1) = 1 - P(X = 1)$$

Probability that the "outcome is an even number"



$$P(X \text{ is even}) = 3/6$$

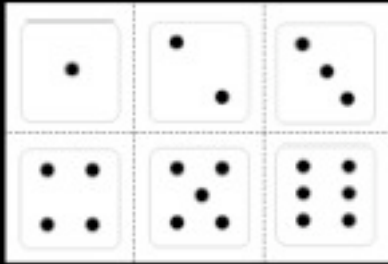
Complementary events



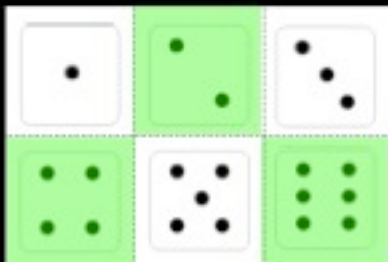
$$P(X \text{ is even}) + P(X \text{ is odd}) = 1$$

$$P(X \text{ is even AND odd}) = 0$$

## Adding probabilities of non-mutually exclusive events

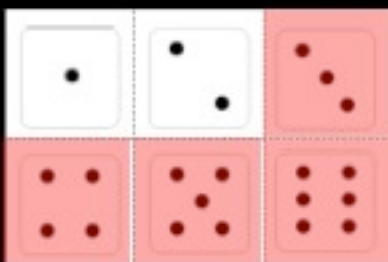


Probability that the outcome "is an even number"



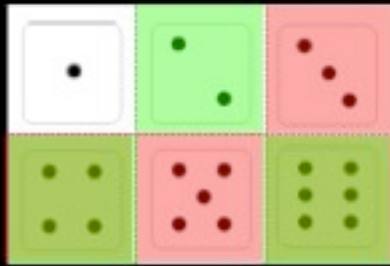
$$P(X \text{ is even}) = 3/6$$

Probability that the outcome "is equal to or greater than 3"



$$P(X \geq 3) = 4/6$$

Probability that the outcome "is equal to or greater than 3"  
OR "is an even number"



$$P(X \geq 3 \text{ OR } X \text{ is even}) = P(X \geq 3) + P(X \text{ is even}) \\ - P(X \geq 3 \text{ AND } X \text{ is even})$$

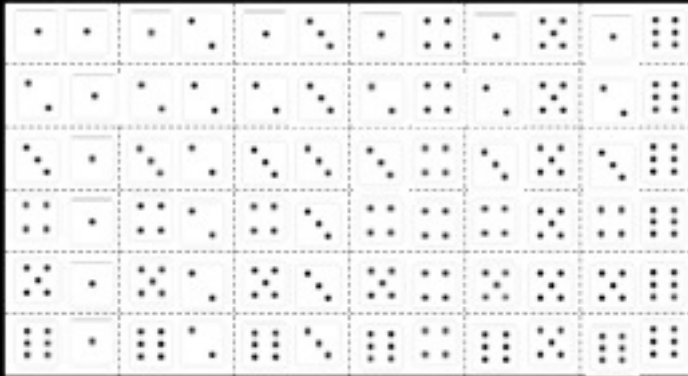
### Probability Addition Rule

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

### Independence and the multiplication rule



## The event space

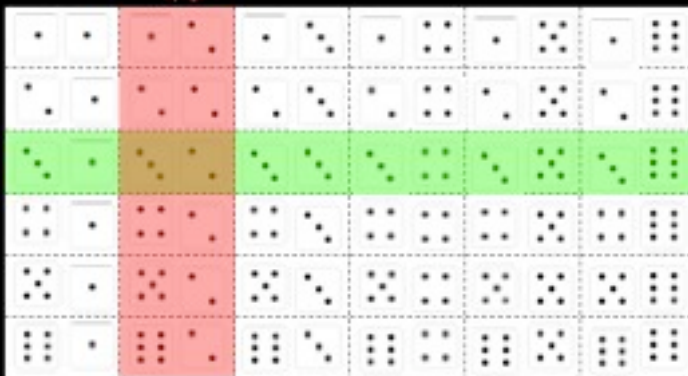


1-1	1-2	1-3	1-4	1-5	1-6
2-1	2-2	2-3	2-4	2-5	2-6
3-1	3-2	3-3	3-4	3-5	3-6
4-1	4-2	4-3	4-4	4-5	4-6
5-1	5-2	5-3	5-4	5-5	5-6
6-1	6-2	6-3	6-4	6-5	6-6

Probability of drawing "three in the first dice"  
AND "two in the second dice"

1/6

1/6



1-1	1-2	1-3	1-4	1-5	1-6
2-1	2-2	2-3	2-4	2-5	2-6
3-1	3-2	3-3	3-4	3-5	3-6
4-1	4-2	4-3	4-4	4-5	4-6
5-1	5-2	5-3	5-4	5-5	5-6
6-1	6-2	6-3	6-4	6-5	6-6

$$P(X_1=3 \text{ AND } X_2=2) = P(X_1=3) \times P(X_2=2) = (1/6) \times (1/6) = 1/36$$

## Probability Multiplication Rule

If two events A and B are independent

$$P(A \text{ AND } B) = P(A) \times P(B)$$

What events are interdependent ?

### Conditional probability

Is the probability of an event given that another event occurs,  
i.e. the probability of an event given that a condition is met

$$P(A|B)$$

### Law of total probability

The probability of an event X is :

$$P(X) = \sum_Y P(Y) P(X|Y)$$

where Y represents all possible mutually exclusive values of the  
conditions



## Probability Multiplication Rule

$$\begin{aligned}P(A \text{ AND } B) &= P(A|B) \times P(B) \\ &= P(B|A) \times P(A)\end{aligned}$$

## Bayes' Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

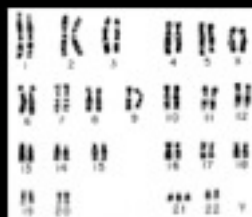
### Example

#### Bayes' Theorem and the diagnosis of Down syndrome



Down syndrome is a chromosomal condition that occurs in about 1:1000 pregnancies. The most accurate test for DS requires amniocentesis, which is not devoid of risk of miscarriage (1:200). It would be good to have an accurate non invasive test without risks. Recently a method has been proposed involving the quantification of three enzymes in the blood.

This so-called triple test does not always identify a fetus with DS (false negative), and some times it incorrectly identifies a fetus with a normal set of chromosomes (false positive). Under normal conditions, the detection rate of the triple test (the probability that a fetus with DS is identified correctly) is 0.60. The false positive rate is 0.05 [Newberger et al. 2000 American Family Physician]



For most people's intuition these numbers are acceptable.  
Make a guess of the rate at which a fetus positively identified has DS.

$$P(DS) = 0.001$$

$$P(TT+ | DS) = 0.60$$

$$P(TT+) = 0.001 \times 0.60 + (1 - 0.001) \times 0.05 = 0.05055$$

$$\begin{aligned} P(DS | TT+) &= P(TT+ | DS) \times P(DS) / P(TT+) \\ &= 0.60 \times 0.001 / 0.05055 \\ &= 0.012 \end{aligned}$$

### Recapitulating

#### Addition

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

#### Multiplication

$$\begin{aligned} P(A \text{ AND } B) &= P(A|B) \times P(B) \\ &= P(B|A) \times P(A) \end{aligned}$$

#### Total probability

$$P(X) = \sum_i P(Y_i) P(X|Y_i)$$

#### Bayes' Theorem

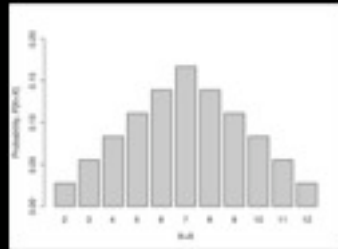
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

ATCGG CCATT CCG

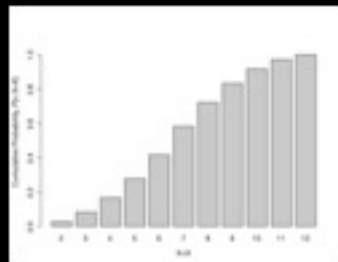
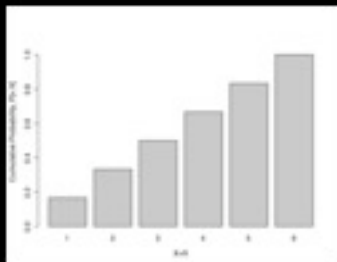
What is the probability of finding the following sequence in a genome?

And in a *Plasmodium falciparum* with a GC content of 20%?

## Probability Distributions



## Probability Distributions

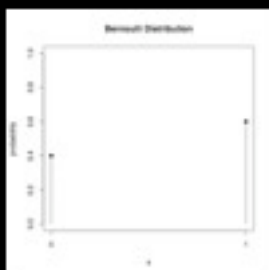


## Discrete Probability Distributions

Event space has a countable number of possible outcomes

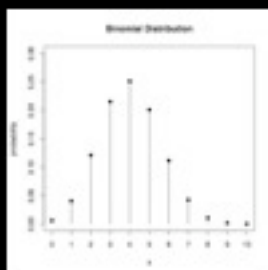
### Bernoulli

Tossing a coin  
Inheritance of a paternal allele



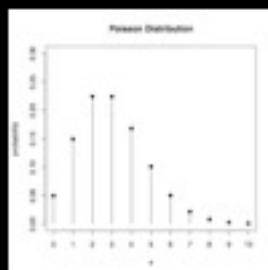
### Binomial

Tossing a coin n times  
Number of cured animals after treatment



### Poisson

Number of births in a certain period  
Number of mutations in a generation



# Continuous Probability Distributions

Event space has a non-countable number of possible events

## Normal

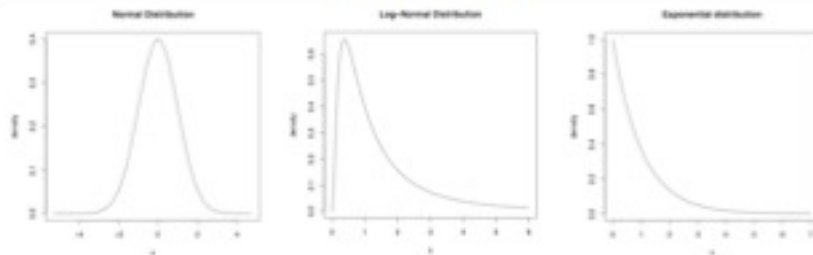
Height of human population

## Log-normal

Age at disease onset  
Survival after cancer diagnosis  
Specific protein content per cell

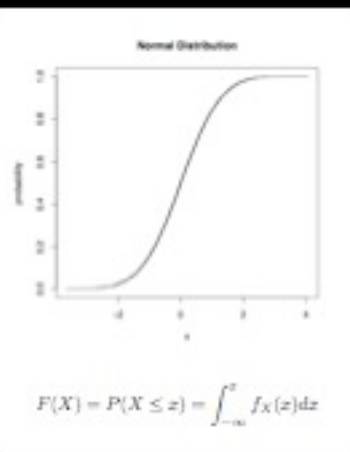
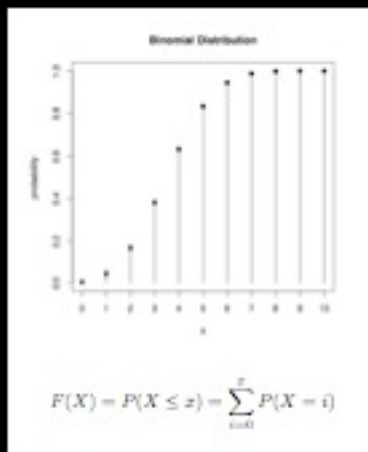
## Exponential

Time between births  
Lymphocyte life span



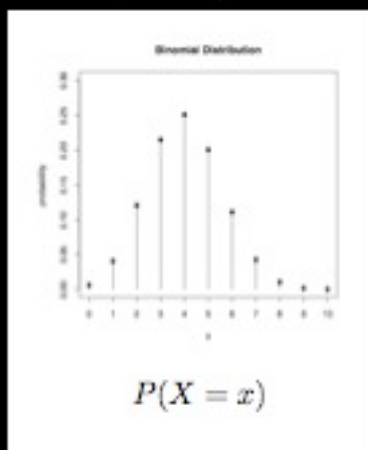
**Probability is an area under the curve**

## Cumulative probability functions

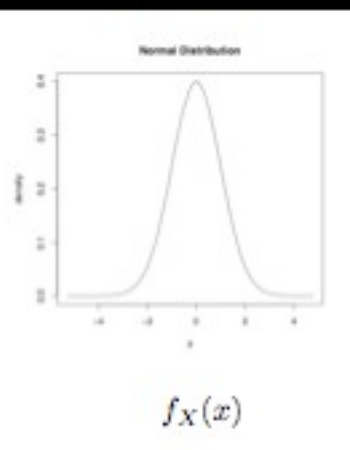


## Probability functions

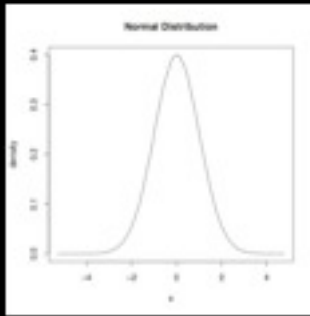
### Mass probability function



### Probability Density Function



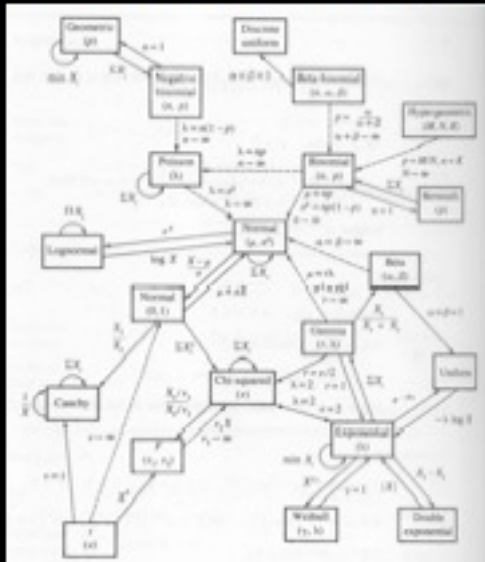
## Normal distribution



Gauss (1777-1855)

Symmetric around the mean  
Mean=Median=Mode

## Why is Normal Distribution so important?



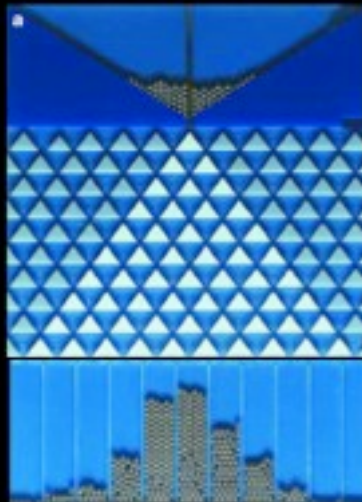
Can you single out a major

concept, experiment, or result in science

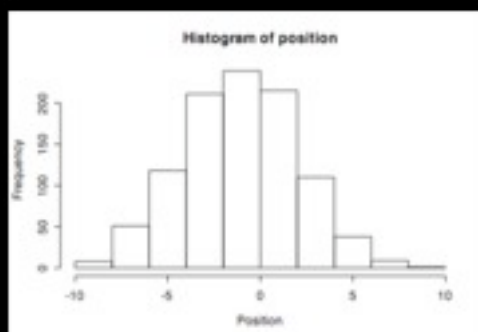
as THE one ?

# The Central Limit Theorem

## The Central Limit Theorem



BioScience (2001) 51,341

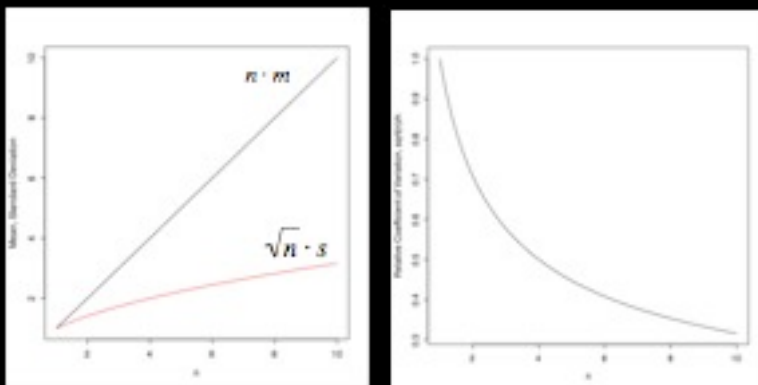


## The Central Limit Theorem

Let  $X_1, X_2, X_3, \dots, X_n$  be identical and independently distributed random variables with mean  $m$  and variance  $s^2$ .

Let  $\sum X$  be the sum of the values of the  $n$  variables.

When  $n$  is very large  $\sum X$  follows approximately a Normal distribution with mean  $n \cdot m$  and variance  $n \cdot s^2$

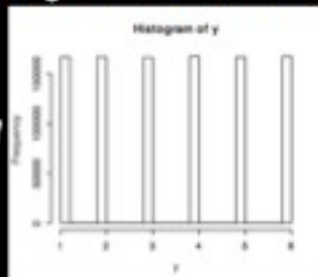


## As a corollary

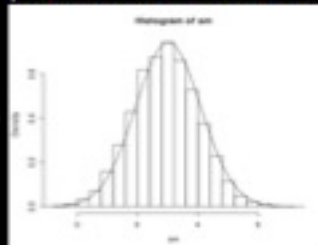
A sample mean tends to be normally distributed irrespective of the distribution of the population one is sampling from



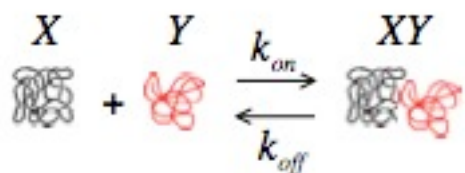
Histogram of 100000 dice draws



Histogram of 1000 means of 10 draws



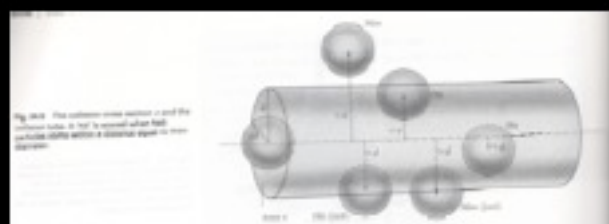
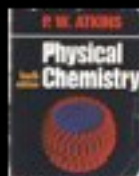
What about some additional implications  
of the Central Limit Theorem ?



$$\frac{d[X]}{dt} = -k_{on}[X][Y] + k_{off}[XY]$$



# Collision Theory of Reaction



**Collision frequency**  
(collisions per molecule per unit of time)

$$z = \frac{2^{1/2} \sigma \bar{c} N}{V} = \frac{2^{1/2} \sigma \bar{c} p}{kT}$$

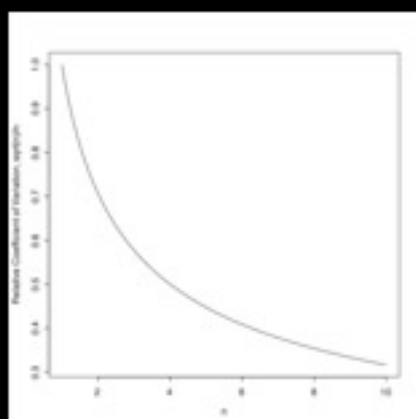
**Collision density**  
(total collisions per unit of volume)

$$Z_{AA} = \sigma \left( \frac{4kT}{\pi m} \right)^{1/2} N_A^2 [A]^2$$

## Collision Theory of Reaction

$$\frac{d[X]}{dt} = -k_{on}[X][Y] + k_{off}[XY]$$

Valid when we have many, many molecule



## WHAT IS LIFE? ERWIN SCHRODINGER

PHYSICAL LAWS REST ON ATOMIC STATISTICS AND ARE THEREFORE ONLY APPROXIMATE

And why could all this not be fulfilled in the case of an organism composed of a moderate number of atoms only and sensitive already to the impact of one or a few atoms only? Because we know all atoms to perform all the time a completely disorderly heat motion, which, so to speak, opposes itself to their orderly behaviour and does not allow the events that happen between a small number of atoms to enrol themselves according to any recognizable laws. Only in the co-operation of an enormously large number of atoms do statistical laws begin to operate and control the behaviour of these assemblies with an accuracy increasing as the number of atoms involved increases. It is in that way that the events acquire truly orderly features. All the

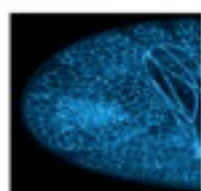
## Schrodinger on diffusion laws ...

Being based on pure chance, its validity is only approximate. If it is, as a rule, a very good approximation, that is only due to the enormous number of molecules that co-operate in the phenomenon. The smaller their number, the larger the quite haphazard deviations we must expect and they can be observed under favourable circumstances.

**How many copies of each molecule  
does a cell have ?**



How are the data produced ?



Cell Tissue

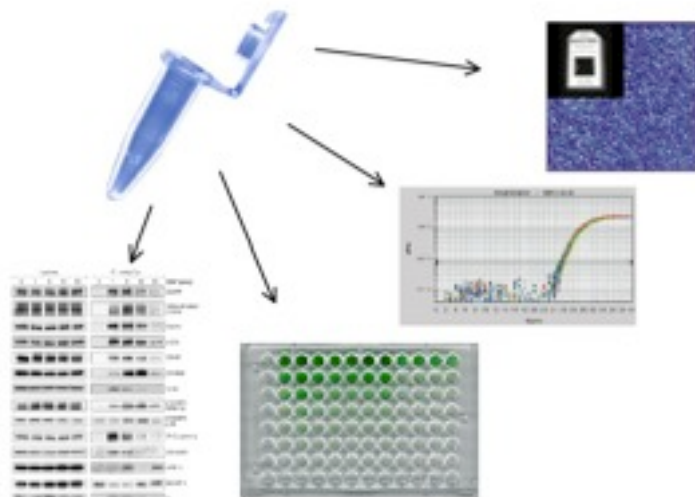


Blending

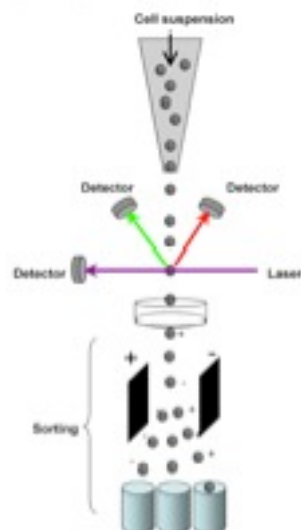


Homogenate

## How are the data produced ?

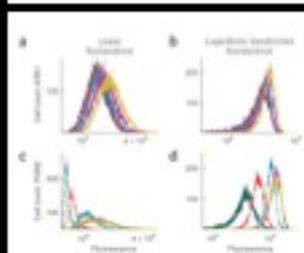


## Flow cytometer (FACS)

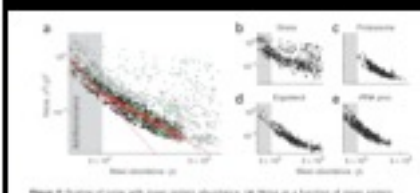


## Noise in protein expression scales with natural protein abundance

Arren Bar-Evan<sup>1</sup>, Johan Paulsson<sup>2,3</sup>, Narendra Maheshri<sup>2</sup>, Miri Carmi<sup>1</sup>, Erin O'Shea<sup>2</sup>, Yitzhak Pilpel<sup>2</sup> & Naama Barkai<sup>1,3</sup>



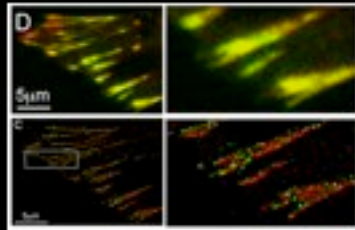
**Figure 1** Log-normal distributions of fluorescence levels. (a) Cells expressing the high-abundance protein PAB1 were sorted from synthetic computer models to produce histograms (1% ethanol). (b) Cells expressing the low-abundance protein MB1 were sorted from the same populations. The cells were subjected to five sequential sorts at different time points after the transfer. Fluorescence distributions are shown as lines (left) and as histograms (right). (c) Cells expressing the high-abundance protein PAB1 were sorted from the same populations. The cells were subjected to five sequential sorts at different time points after the transfer. Fluorescence distributions are shown as lines (left) and as histograms (right). (d) Cells expressing the low-abundance protein MB1 were sorted from the same populations. The cells were subjected to five sequential sorts at different time points after the transfer. Fluorescence distributions are shown as lines (left) and as histograms (right).



**Figure 2** Scaling of noise with mean protein abundance. (a) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Thick curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ . (b) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Thin curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ . (c) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Dotted curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ . (d) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Dashed curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ . (e) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Dotted curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ . (f) Noise as a function of mean protein abundance for genes in all conditions and time points are shown. Dashed curve corresponds to  $\log(\sigma^2) = 0.175 - \log(\mu)$ .

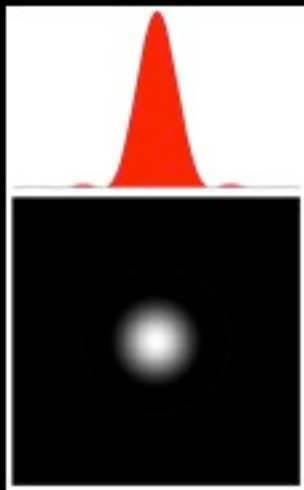
## Super resolution microscopy:

PALM, STORM beyond the Rayleigh diffraction limit



But before ...

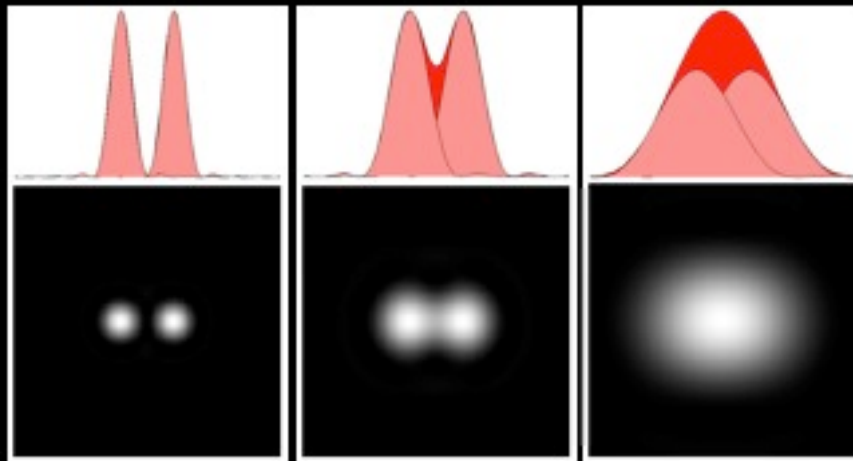
**Airy disc** is the central bright disc present in the diffraction pattern generated by a perfect, aberration-free lens.



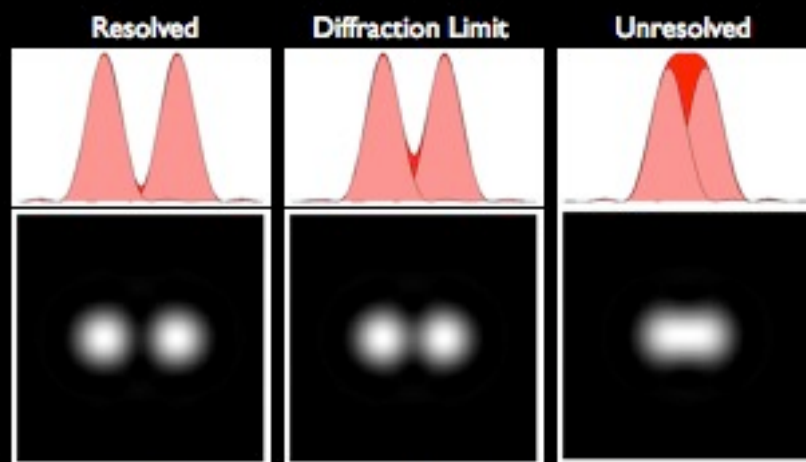
### Radius of the Airy Disc

$$radius_{Airy} = 1.22\lambda / 2NA_{obj}$$

$\lambda$  wavelength and NA is the numerical aperture of the objective



**Rayleigh criterion** is used to establish the minimum resolvable distance between two light sources (independent of the magnification)



<http://www.microscopyu.com/tutorials/java/imageformation/airyna/>

### Resolution: Rayleigh criterion

Laterally

$$\sim \frac{\lambda}{2NA} \quad (x, y)$$

Axially

$$\sim \frac{2\lambda\eta}{(2NA)^2} \quad (z)$$

Conventional fluorescence microscope

with visible light

$$450nm < \lambda < 700nm$$

and high numerical aperture objective

$$NA = 1.4$$

$$\sim 200nm \quad (x, y) \quad 500 - 800nm \quad (z)$$



## The central limit theorem

Let  $X_1, X_2, X_3, \dots, X_n$  be identical and independently distributed random variables with mean  $m$  and variance  $s^2$ .

Let  $\Sigma_n$  be the sum of values of the  $n$  variables.

When  $n$  is very large  $\Sigma_n$  follows approximately a Normal distribution with mean  $nm$  and variance  $ns^2$

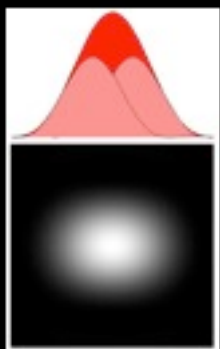
## Resolution in PALM, STORM

$$\frac{\sigma}{\sqrt{N}} \quad (x, y)$$

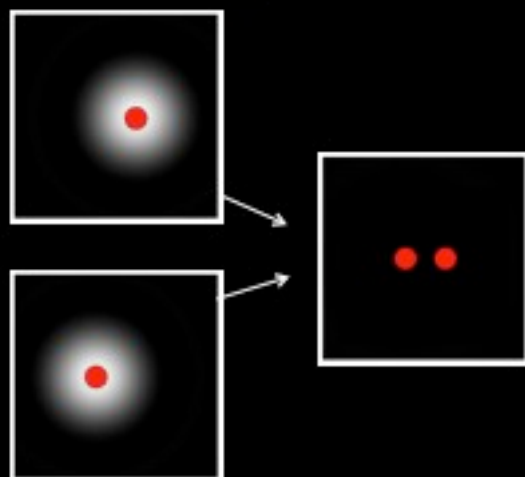
**Statistics:** Standard error of the sample mean

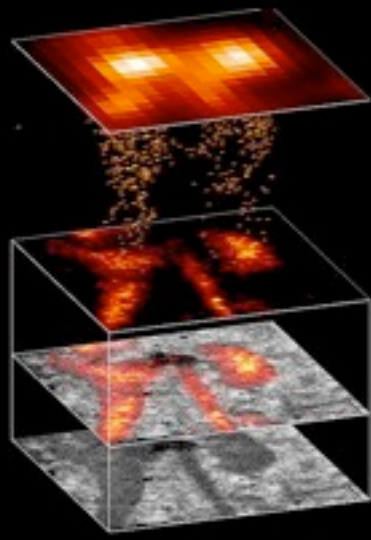
Henriques & Milanga. PALM and STORM: What hides beyond the Rayleigh limit. *Biotechnol J* (2009) 4, 846

Unresolved by  
conventional wide-field  
microscopy



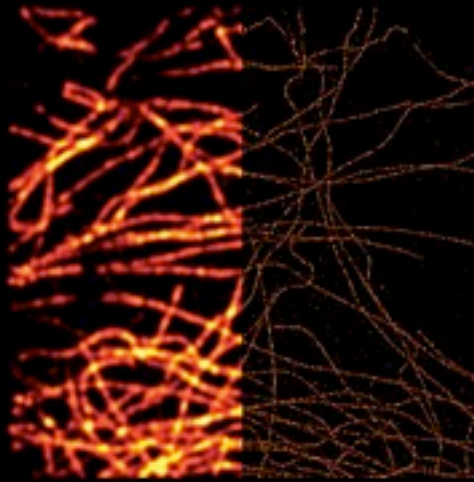
But would be resolved if one could look at  
one fluorophore at a time





[http://images.pennnet.com/articles/bow/thm/th\\_268711.jpg](http://images.pennnet.com/articles/bow/thm/th_268711.jpg)

### Photo-activated Localization Microscopy (PALM)



TIRF image (left) and PALM image (right) of antibody staining for tubulin in a cultured cell. Spadman: S. Nish, University of Tokyo, Japan.

"Mathematics is Biology's next microscope, only better.  
Biology is Mathematics'next Physics, only better."

— Joel E. Cohen PLoS Biology