DATA ANALYSIS

Toni Šćulac
Faculty of Science, University of Split, Croatia
visiting professor at University of Latvia, Latvia

CERN School of Computing 2023, Tartu, Estonia

LECTURES OUTLINE

1) Introduction to Data Analysis
2) Probability density functions and Monte Carlo methods
3) Parameter estimation and Confidence intervals
4) Hypothesis testing and p-value

PROBABILITY DENSITY FUNCTIONS AND MOTE CARLO METHODS

PROBABILITY DENSITY FUNCTION

- Let x be a possible outcome of an observation and can take any value from a continuous range
- We write f(x;θ)dx as the probability that the measurement's outcome lies betwen x and x + dx
- The function **f(x;θ)dx** is called the **probability density function** (**PDF**)
  - And may depend on one or more parameters θ
- If f(x;θ) can take only discrete values then f(x;θ) is itself a probability
- The p.d.f. is always normalised to a unit area (unit sum, if discrete)
- Both **x** and **θ** may have multiple components and are then written as vectors

$$P(x \in [x, x + dx] \,|\, \theta) = f(x;\theta)dx$$

$$\int_{-\infty}^{\infty} f(x;\theta)dx = 1$$

CUMULATIVE AND MARGINAL DISTRIBUTIONS

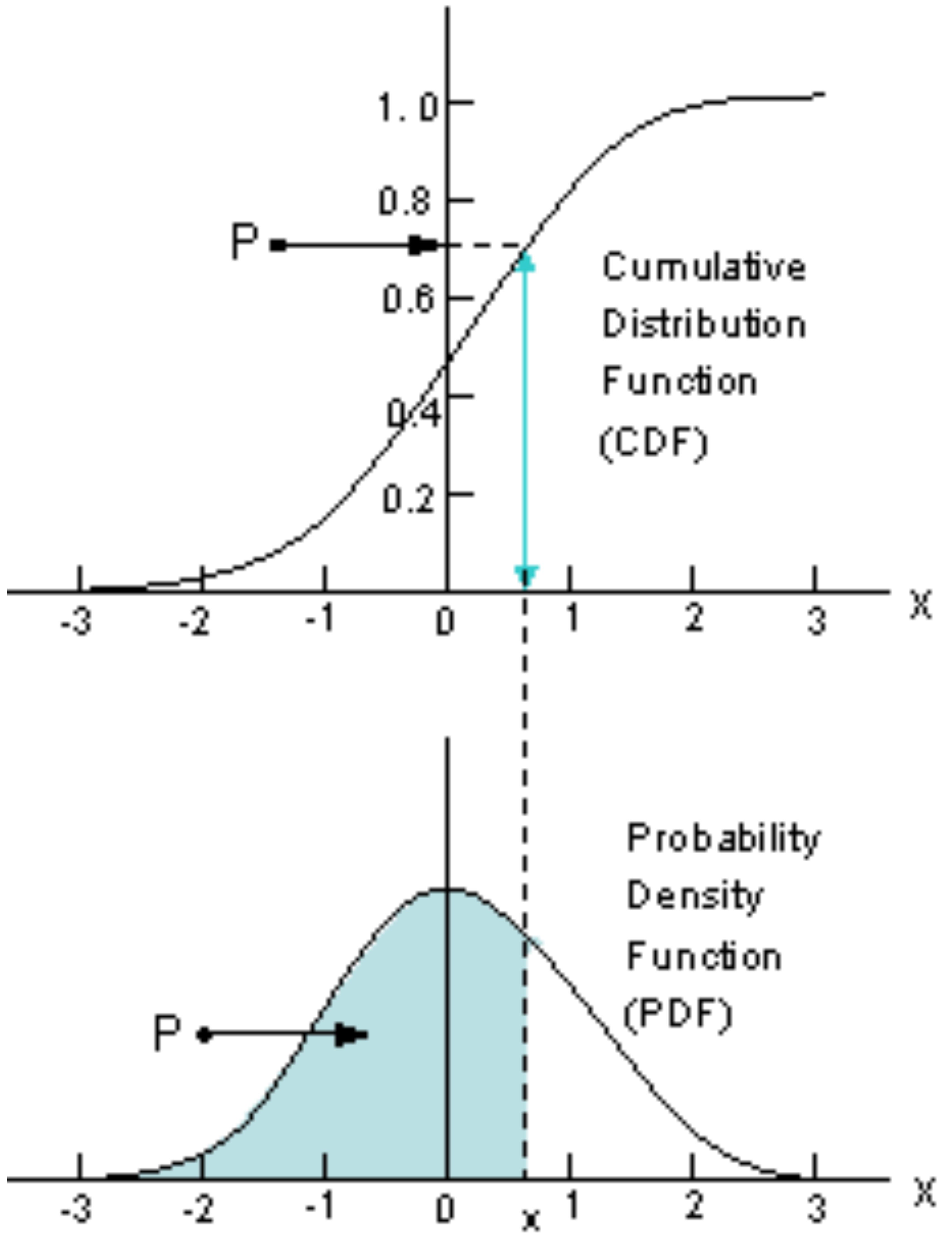
- Cumulative distribution function, CDF
  - for every real number Y, the CDF of Y is equal to the probability that the random variable x takes a value less or equal to Y

$$F(Y) = P(x \leq Y) = \int_{x_{min}}^{Y} f(x)dx$$

  - If x restricted to $x_{min} < x < x_{max}$ then $F(x_{min}) = 0$, $F(x_{max}) = 1$
  - F(x) is a monotonic function of x

- Marginal density function
  - is the projection of multidimensional density
  - Example: if f(x,y) is two-dimensional PDF the marginal density g(x) is

$$g(x) = \int_{y_{min}}^{y_{max}} f(x, y)dy$$



PROPERTIES OF THE PDF

- Probability density function (PDF) = f(x)dx
- Expectation:
  - Expectation of any random function g(x): $E(g) = \int g(x)f(x)dx$
  - Expectation of x is the **mean**: $\mu = E(x) = \int xf(x)dx$
- **Variance**: $V(x) = \sigma^2 = E[(x-\mu)^2] = \int (x-\mu)^2 f(x)dx$

- E(x) is usually a measure of the **location** of the distribution
- V(x) is usually a measure of the **spread** of the distribution

BINOMIAL DISTRIBUTION

- Probability for r successes is given by the Binomial distribution:

$$P(r;p,N) = \binom{N}{r} p^r (1-p)^{N-r}$$

- P(r;N,p) is a probability of finding exactly r successes in N trials, when probability of success in each single trial is a constant, p

- Properties of the Binomial distribution:

  - Mean: $< r > = E(r) = Np$

  - Variance: $V(r) = Np(1-p)$

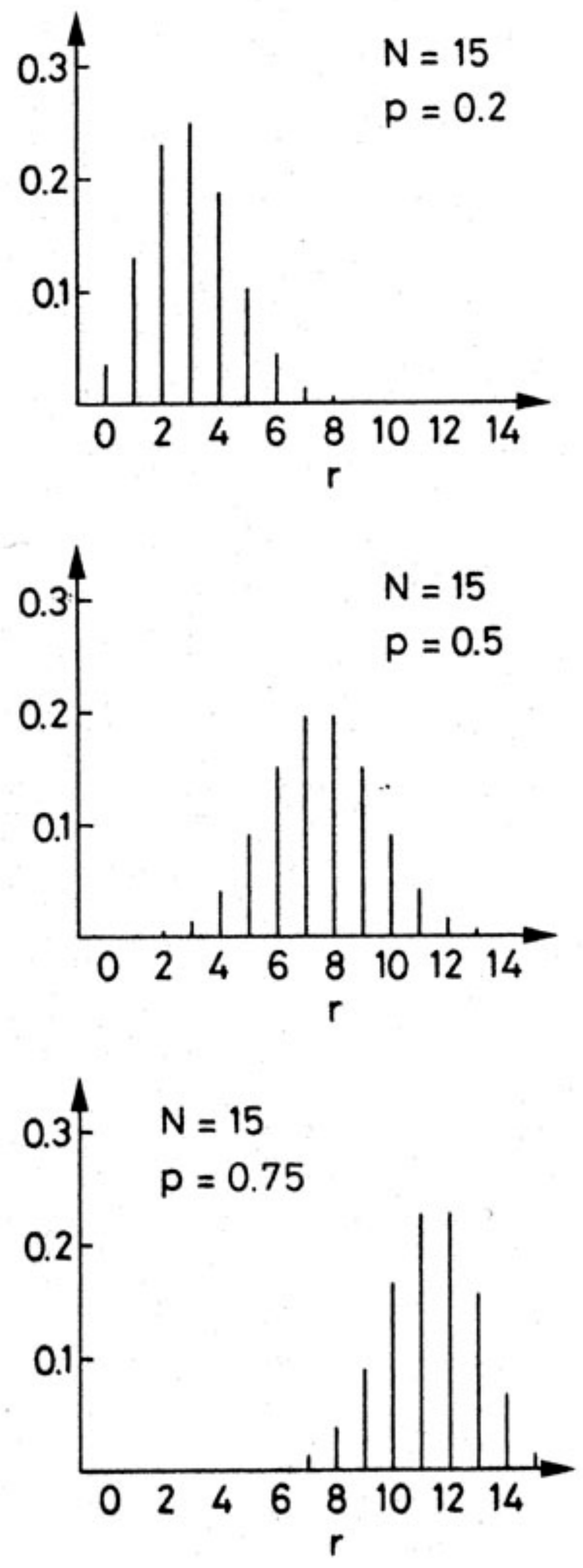
BINOMIAL DISTRIBUTION: EXAMPLE

- Usage example 1:
  - Probability for a Z boson to decay to two electrons is 3%. What is the probability to find exactly 5 Z→ee events out of 80 Z decays?

$$P(5; 0.03, 80) = \binom{80}{5} 0.03^5 (1-0.03)^{80-5} = 6\,\%$$

- Usage example 2:
  - If you flip a biased coin that has a 99% probability of landing on heads, what is the probability to get heads all 6 times from 6 throws?

$$P(6; 0.99, 6) = \binom{6}{6} 0.99^6 (1-0.99)^{6-6} = 94.15\,\%$$



FROM BINOMIAL TO POISSON

- λ events expected to occur in average during some time interval
- split interval in n very small divisions: chance of getting two events in one section can be discounted
- probability that a given section contains an event: λ/n
- use Binomial formula to calculate the probability to see r events:

$$P(r;\lambda/n,n) = \binom{n}{r}\frac{\lambda^r}{n^r}(1-\frac{\lambda}{n})^{n-r} = \frac{n!}{r!(n-r)!}\frac{\lambda^r}{n^r}(1-\frac{\lambda}{n})^{n-r}$$

- For n→∞: $(1-\frac{\lambda}{n})^{n-r} \to (1-\frac{\lambda}{n})^{n} \to e^{-\lambda}$

$$\frac{n!}{r!(n-r)!} = \frac{n(n-1)\cdots(n-r+1)\cdot\cancel{(n-r)!}}{r!\cancel{(n-r)!}} \to \frac{n^r}{r!}$$

POISSON DISTRIBUTION

- Probability of a number of events occurring in a fixed period of time if these events occur with a known average rate λ and independently of the time since the last event:

$$P(r, \lambda) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

- Compared to Binomial distribution still has particular (discrete) outcomes, but number of trials is unknown

- Properties of the Poisson distribution:
  - Mean: $< r > = E(r) = \lambda$
  - Variance: $V(r) = \lambda$

BONUS PROBLEM - 2

Some rules to follow:

1. In every lecture there will be one bonus problem presented
2. If you have good knowledge in stats and everything I am presenting is known to you feel free to start working on the problem now!
3. Otherwise, work on the problem after the lectures.
4. Solutions won't be provided, you have to come and talk to me to check if your answer is correct or if you need hints!
5. Google/AI assistance is not allowed. These are problems that I want you to think about on your own

Prove that the mean of the Binomial distribution is Np and for Poissonian is λ.

Hint! Try to find the relationship between $\binom{N}{r}$ and $\binom{N-1}{r-1}$.

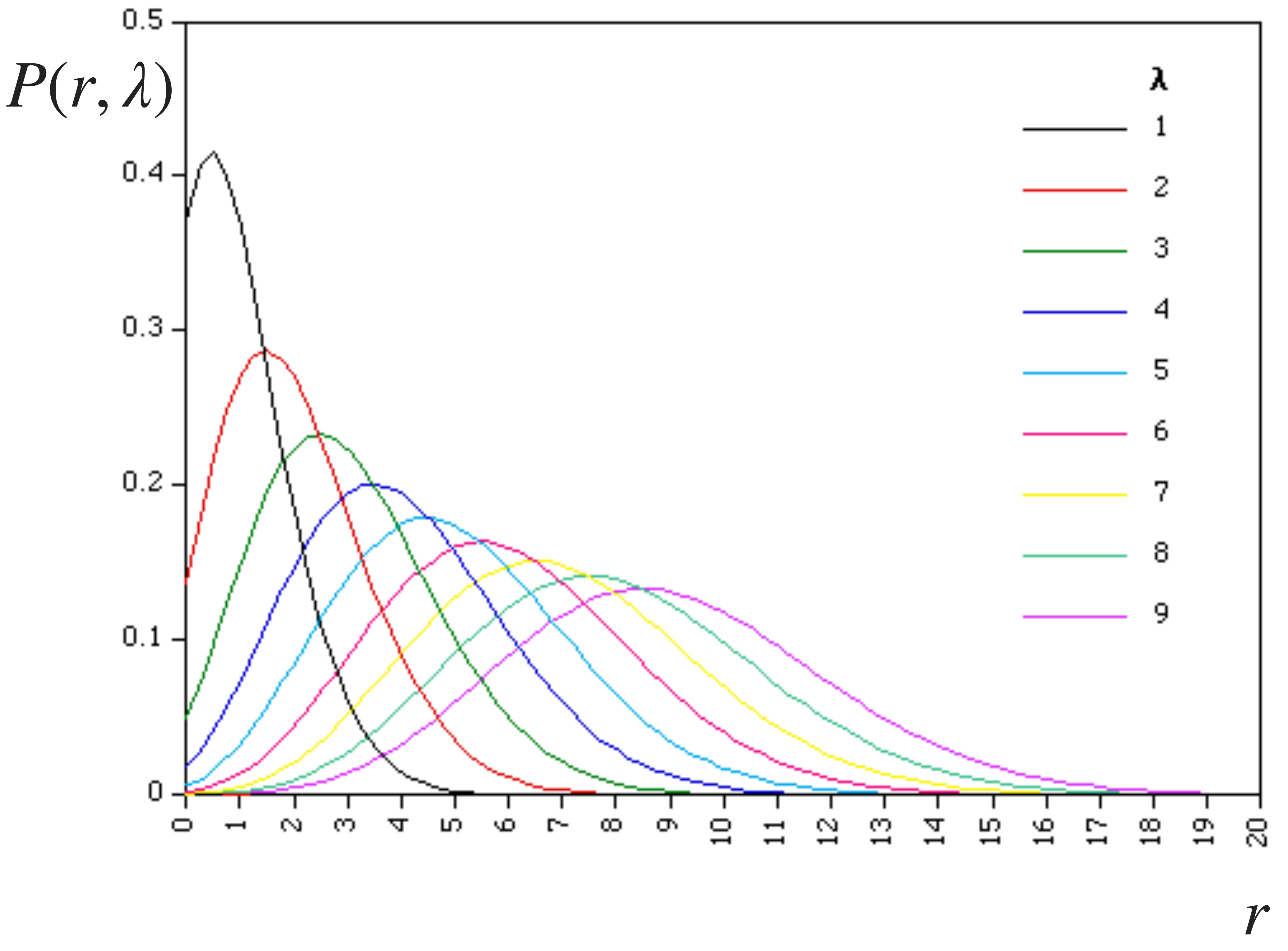
POISSON DISTRIBUTION: EXAMPLE

- Usage example 1:
  - FIFA reports that the average number of goals in a World Cup soccer match is approximately 2.5. What is the probability to have 5 goals in a match?

$$P(5,2.5) = \frac{e^{-2.5}2.5^5}{5!} = 6.7\,\%$$

- Usage example 2:
  - If expect to detect one extremely high energy gamma ray every year, what is the probability not to detect any in a year?

$$P(0,1) = \frac{e^{-1}1^0}{0!} = 36.8\,\%$$



POISSON OR BINOMIAL?

- A student is trying to hitch a lift. Cars pass at random intervals at an average rate of 1 per minute. The probability of a car giving a lift is 1%.What is the probability that the student will still be waiting:
  (1) after 60 cars have passed?
  (2) after 1 hour?

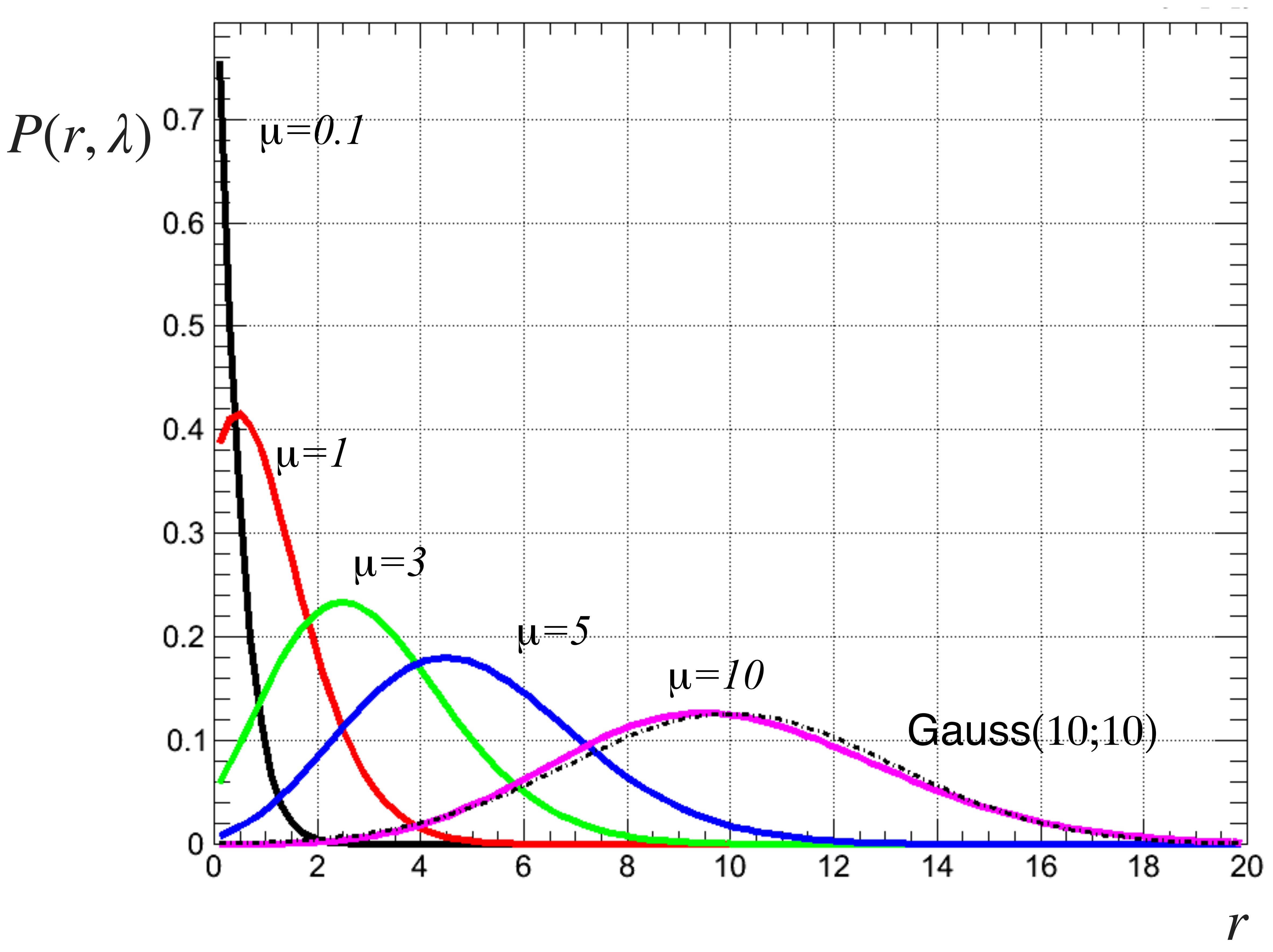
(1) $$P(60;0.99,60) = \binom{60}{60} 0.99^{60}(1-0.99)^{60-60} = 54.71\,\%$$

(2) $$P(0,0.6) = \frac{e^{-0.6}0.6^{0}}{0!} = 54.88\,\%$$

FROM POISSON TO NORMAL DISTRIBUTION

- For a large λ Poisson distribution converges towards a Gaussian distribution

$$P(r,\lambda) = \frac{e^{-\lambda}\lambda^r}{r!} \xrightarrow{\text{large } \lambda} Gauss(r;\mu \equiv \lambda)$$



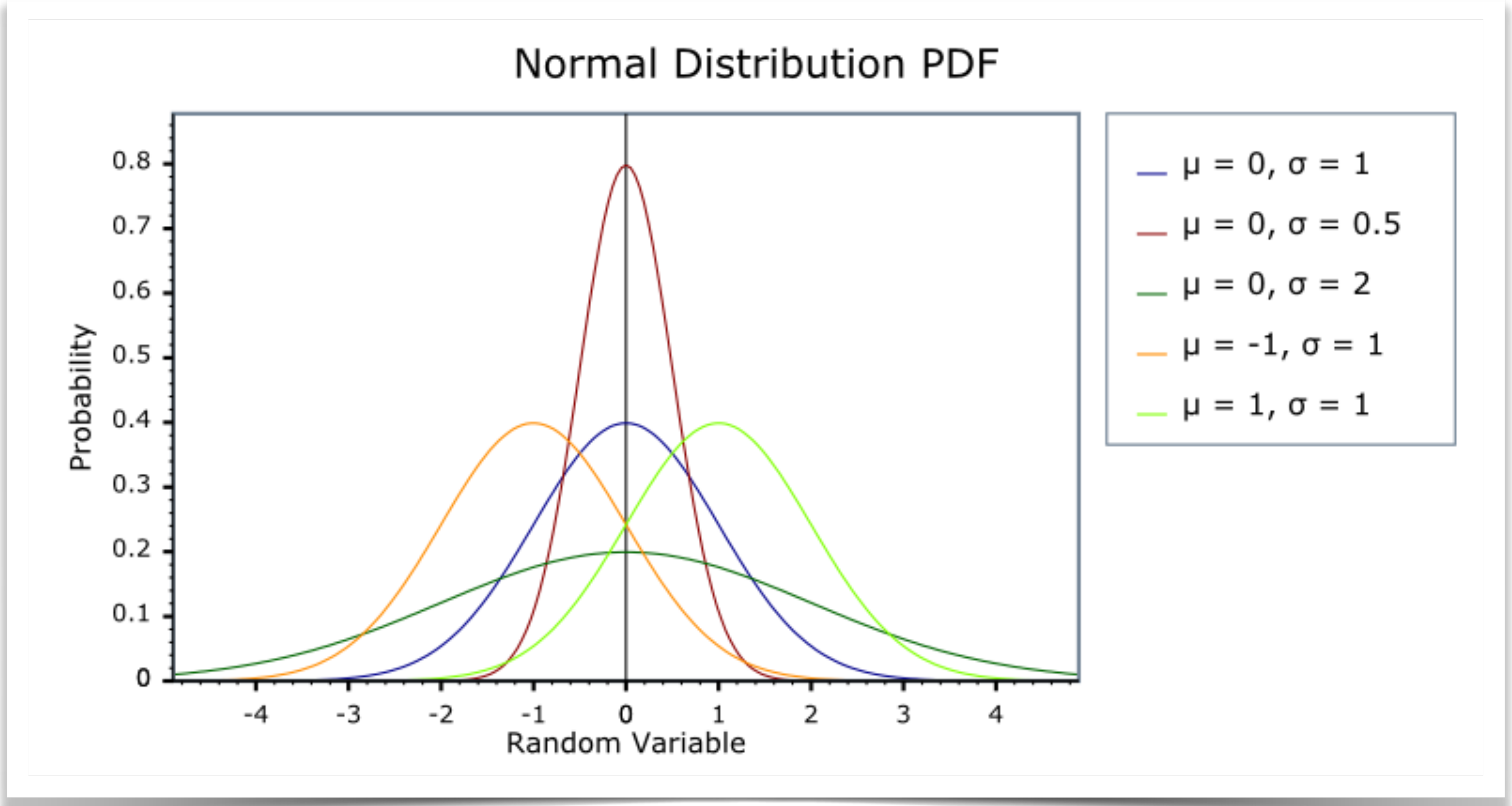
NORMAL OR GAUSSIAN DISTRIBUTION

- The most important distribution in statistics because of the Central Limit Theorem:

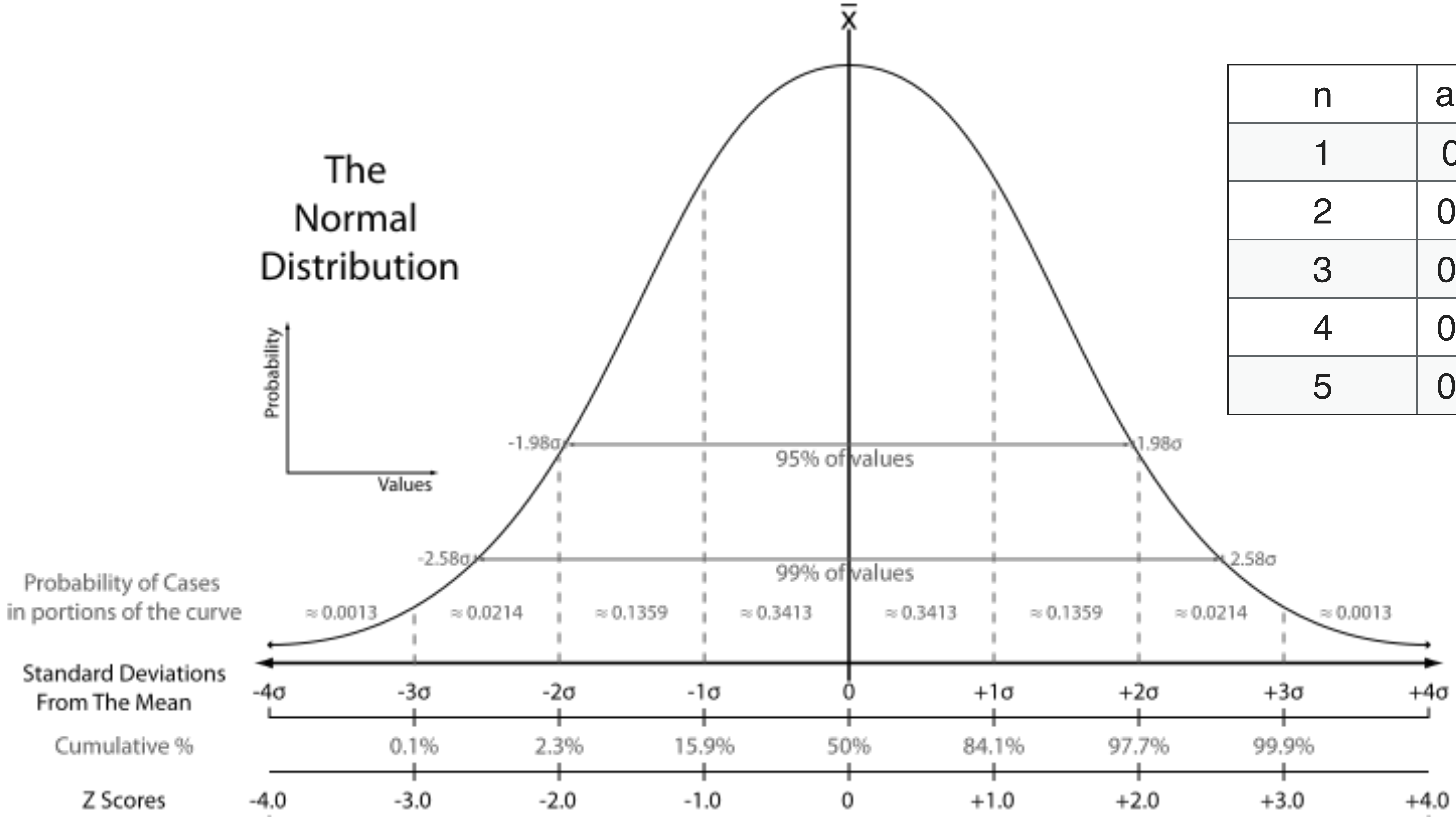
$$N(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- N(0,1) is called standard Normal density

- Properties of the Gaussian distribution:
  - Mean: $< r > = E(r) = \mu$
  - Variance: $V(r) = \sigma^2$

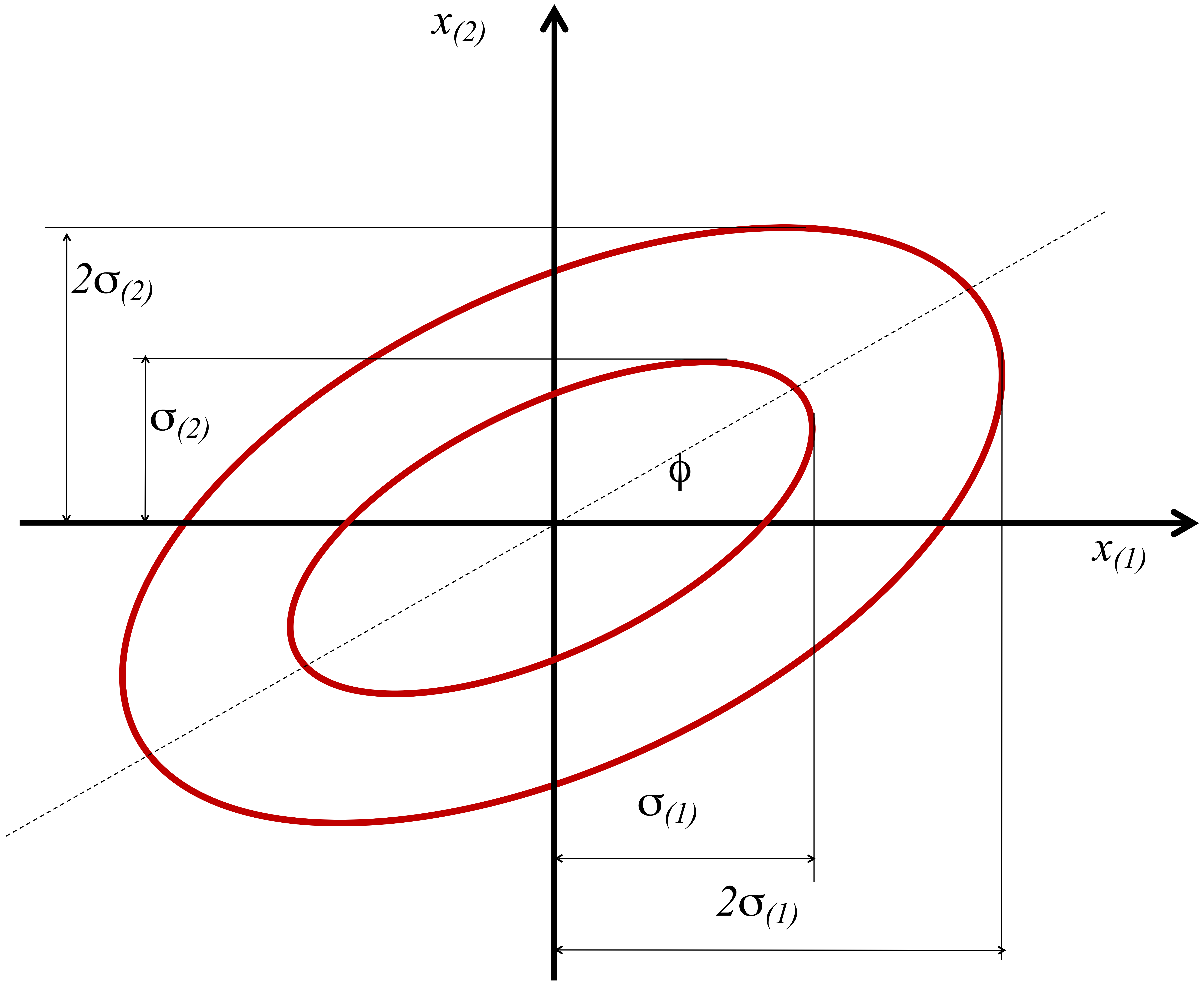


NORMAL DISTRIBUTION PROPERTIES



n	area ± nσ
1	0.682689
2	0.954499
3	0.997300
4	0.999936
5	0.999999

2D GAUSSIAN



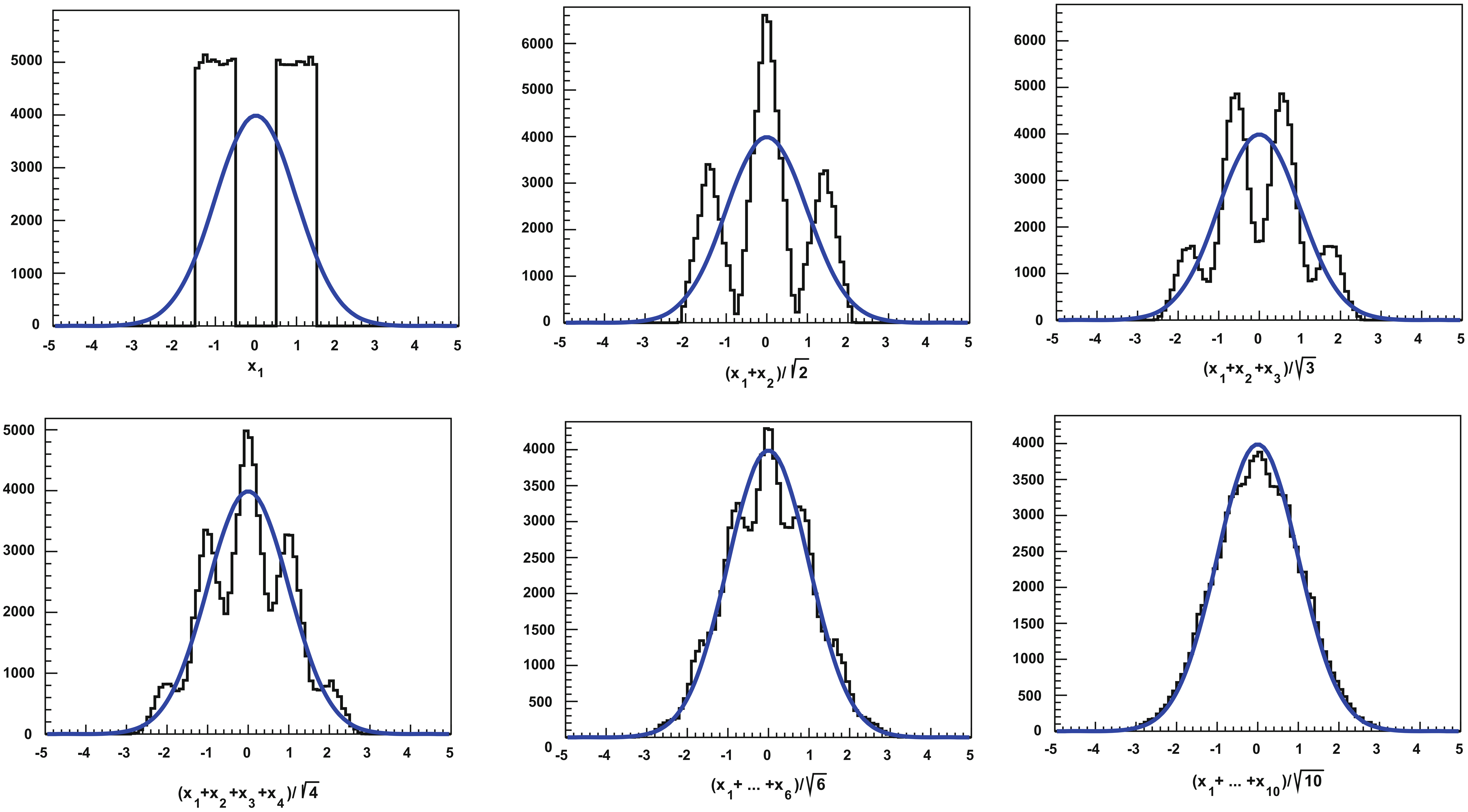
n	P_{1D}	P_{2D}
1σ	**0.6827**	0.3934
2σ	**0.9545**	0.8647
3σ	**0.9973**	0.9889
1.515σ		**0.6827**
2.486σ		**0.9545**
3.439σ		**0.9973**

CENTRAL LIMIT THEOREM

- **Central limit theorem**:
  - If we have a set of N independent variables x_i, each from a distribution with mean μ_i and variance σ_i^2, then the distribution of the sum $X = \Sigma\, x_i$
    - has a mean $\langle X \rangle = \Sigma\, \mu_i$,
    - has a variance $V(X) = \Sigma\, \sigma_i^2$,
    - becomes Gaussian as $N \rightarrow \infty$.

- Therefore, no matter what the distributions of original variables may have been, their sum will be Gaussian in a large N limit

- Example:
  - measurements errors
  - human heights are well described by a Gaussian distribution, as many other anatomical measurements, as these are due to the combined effects of many genetic and environmental factors
  - student test scores



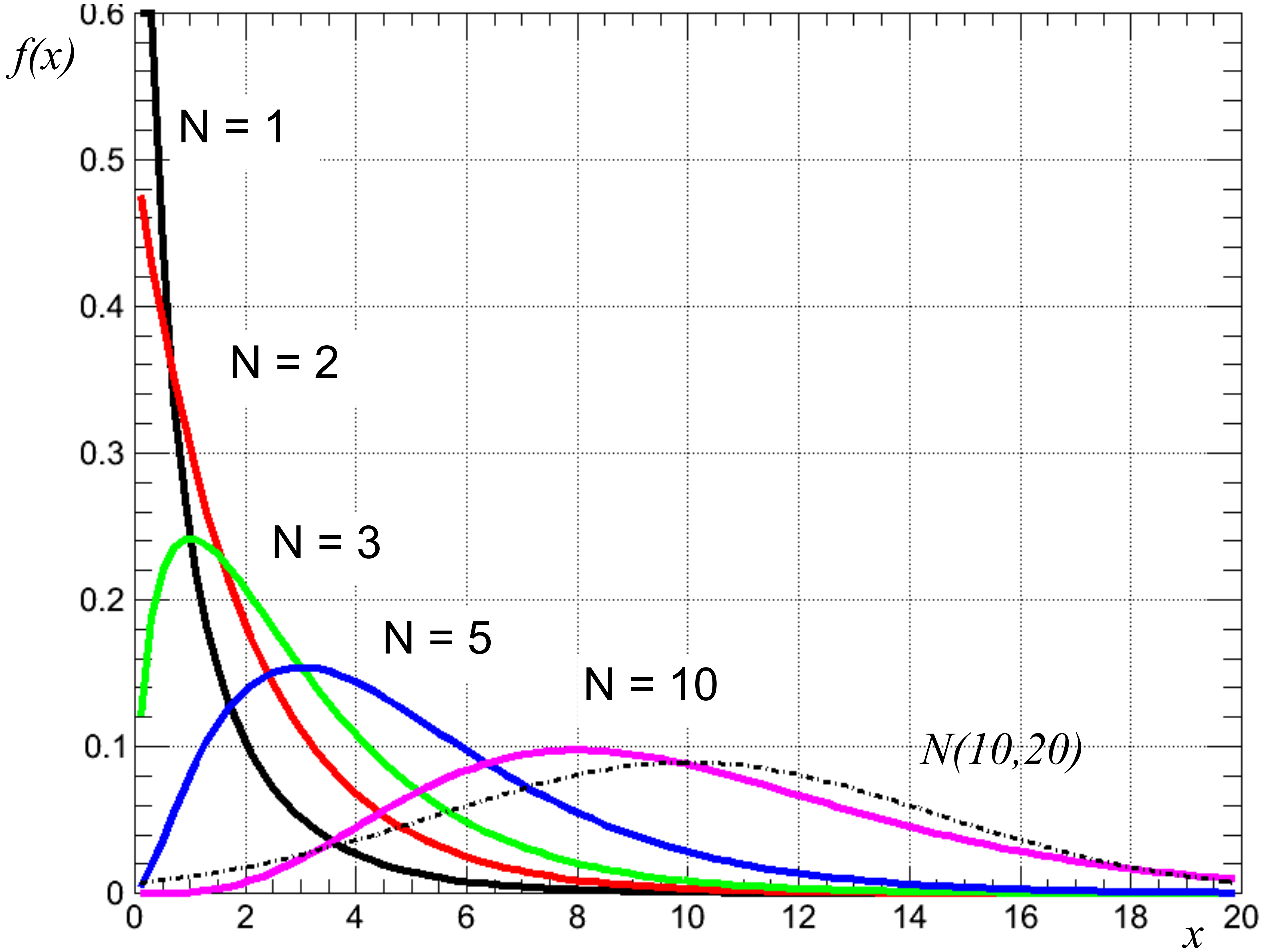
5000
4000
3000
2000
1000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
x_1
6000
5000
4000
3000
2000
1000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
$(x_1+x_2)/\sqrt{2}$
6000
5000
4000
3000
2000
1000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
$(x_1+x_2+x_3)/\sqrt{3}$
5000
4000
3000
2000
1000
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
$(x_1+x_2+x_3+x_4)/\sqrt{4}$
4000
3500
3000
2500
2000
1500
1000
500
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
$(x_1+ \ldots +x_6)/\sqrt{6}$
4000
3500
3000
2500
2000
1500
1000
500
0
-5 -4 -3 -2 -1 0 1 2 3 4 5
$(x_1+ \ldots +x_{10})/\sqrt{10}$

CHI-SQUARE DISTRIBUTION

- If x_i are k independent, normally distributed random variables with mean 0 and variance, then the random variable $Q = \Sigma x_i^2$ is distributed according to the chi-square distribution with k degrees of freedom

$$f(x) = \frac{\frac{1}{2}(\frac{X}{2})^{\frac{N}{2}-1}e^{-\frac{x}{2}}}{\Gamma(\frac{N}{2})}$$

- Properties of the Chi-Square distribution:
  - Mean: $< x > = E(x) = N$
  - Variance: $V(x) = 2N$



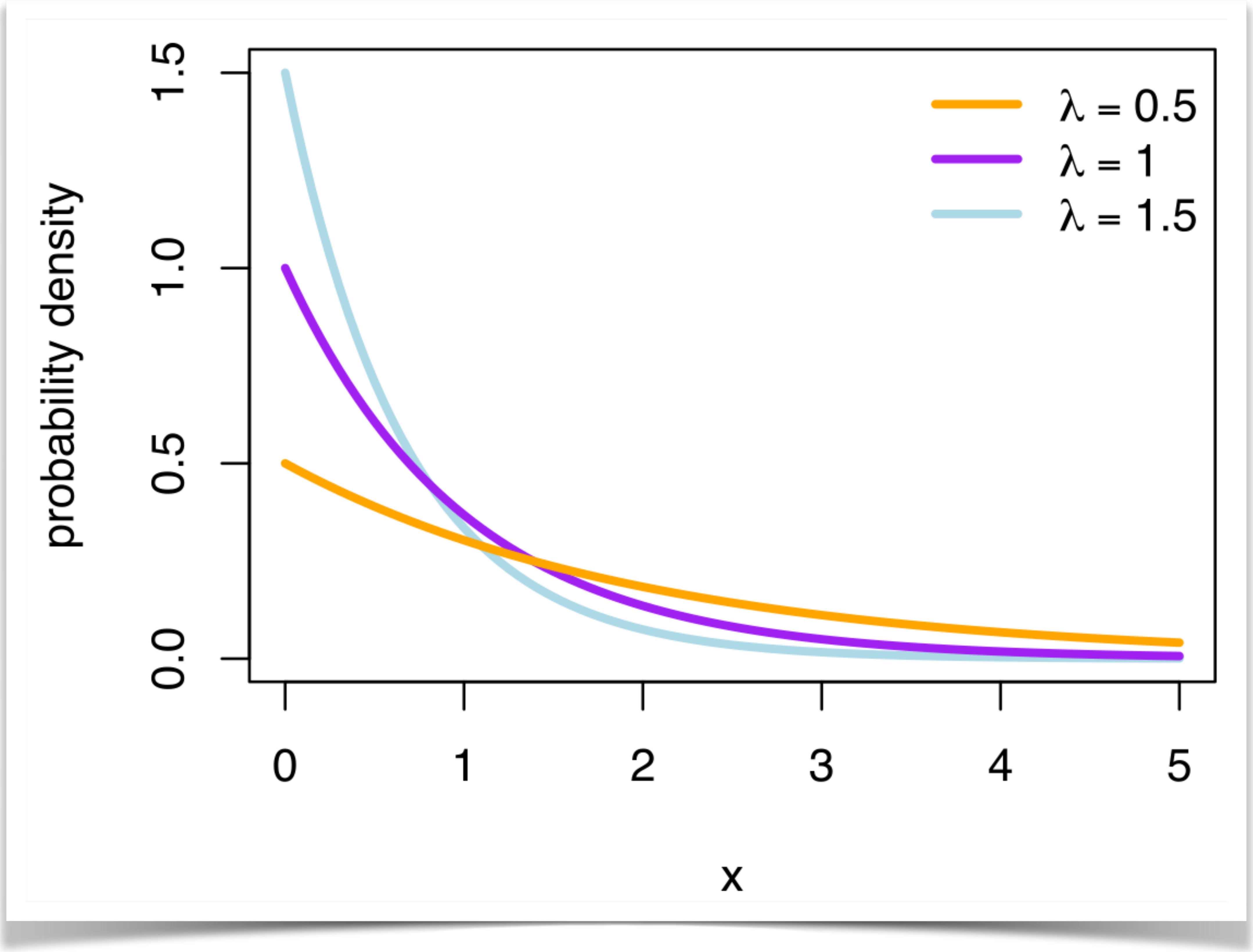
EXPONENTIAL DISTRIBUTION

- Exponential probability density of the continuous variable x > 0:

$$N(x;\lambda) = \lambda e^{-\lambda x}$$

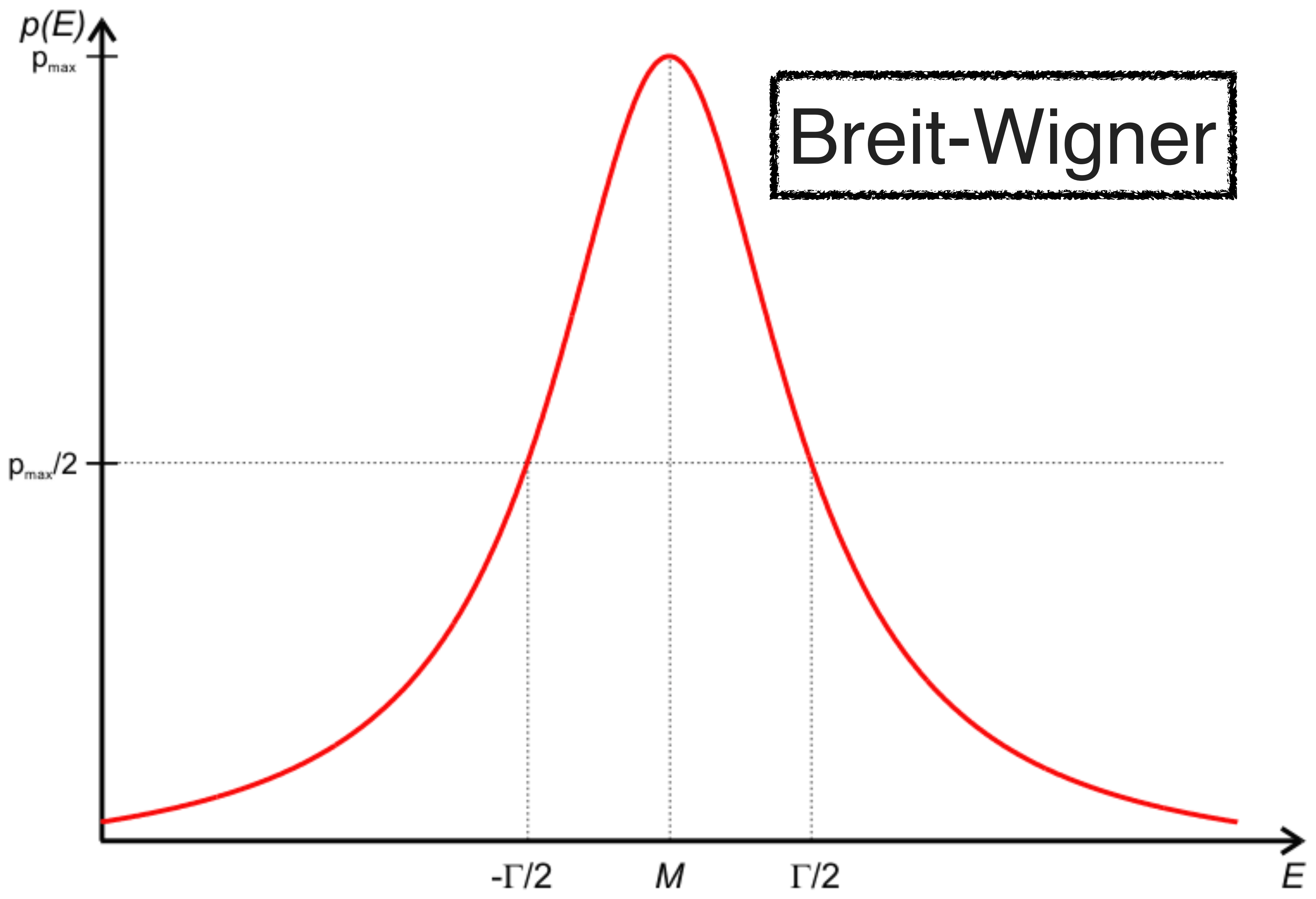
- Example: decay time of an unstable particle measured in its rest frame
  - $\lambda = 1/\tau$(particle)

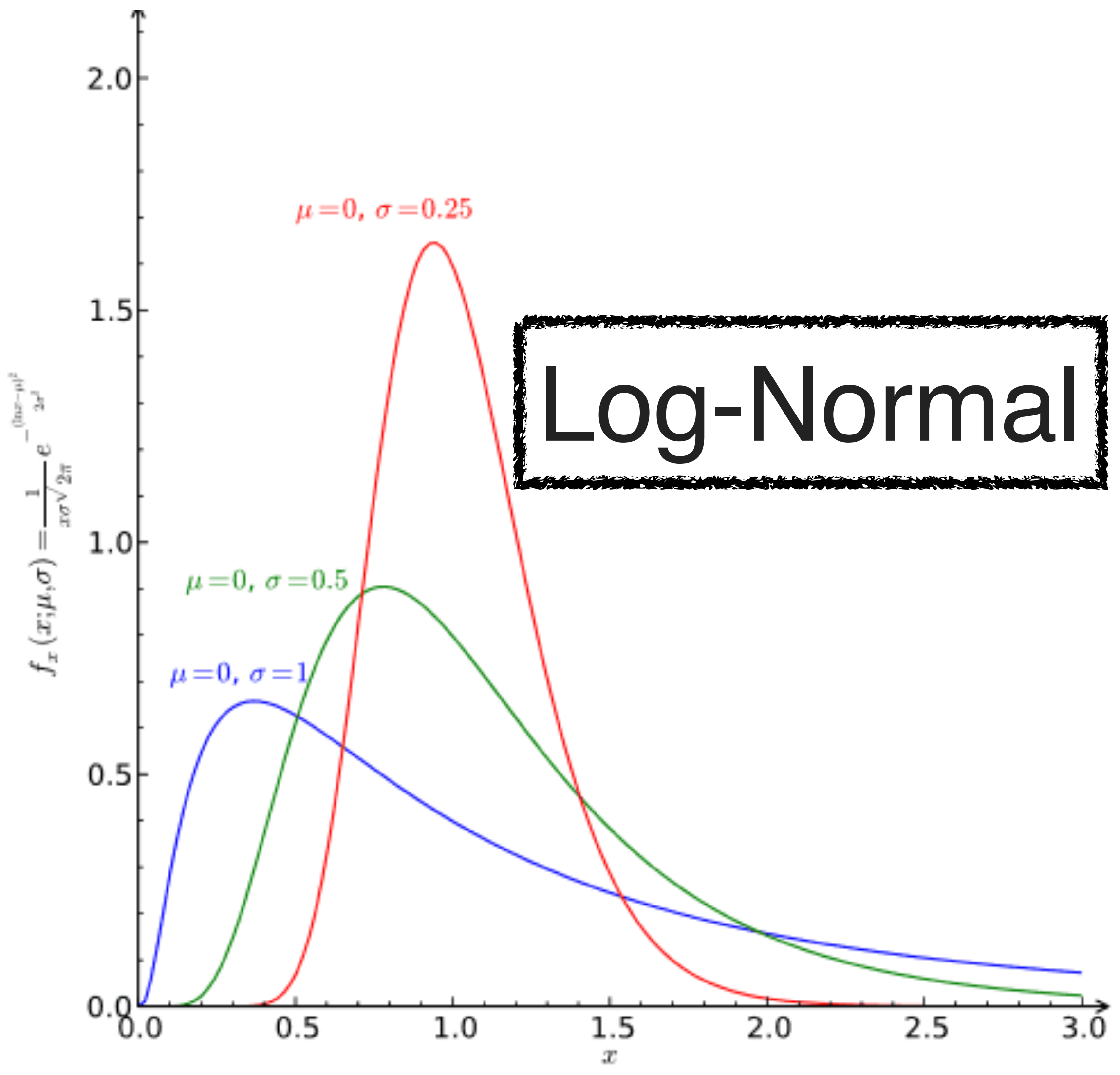
- Properties of the Exponential distribution:
  - Mean: $$< x > = E(x) = \frac{1}{\lambda}$$
  - Variance: $$V(r) = \frac{1}{\lambda^2}$$



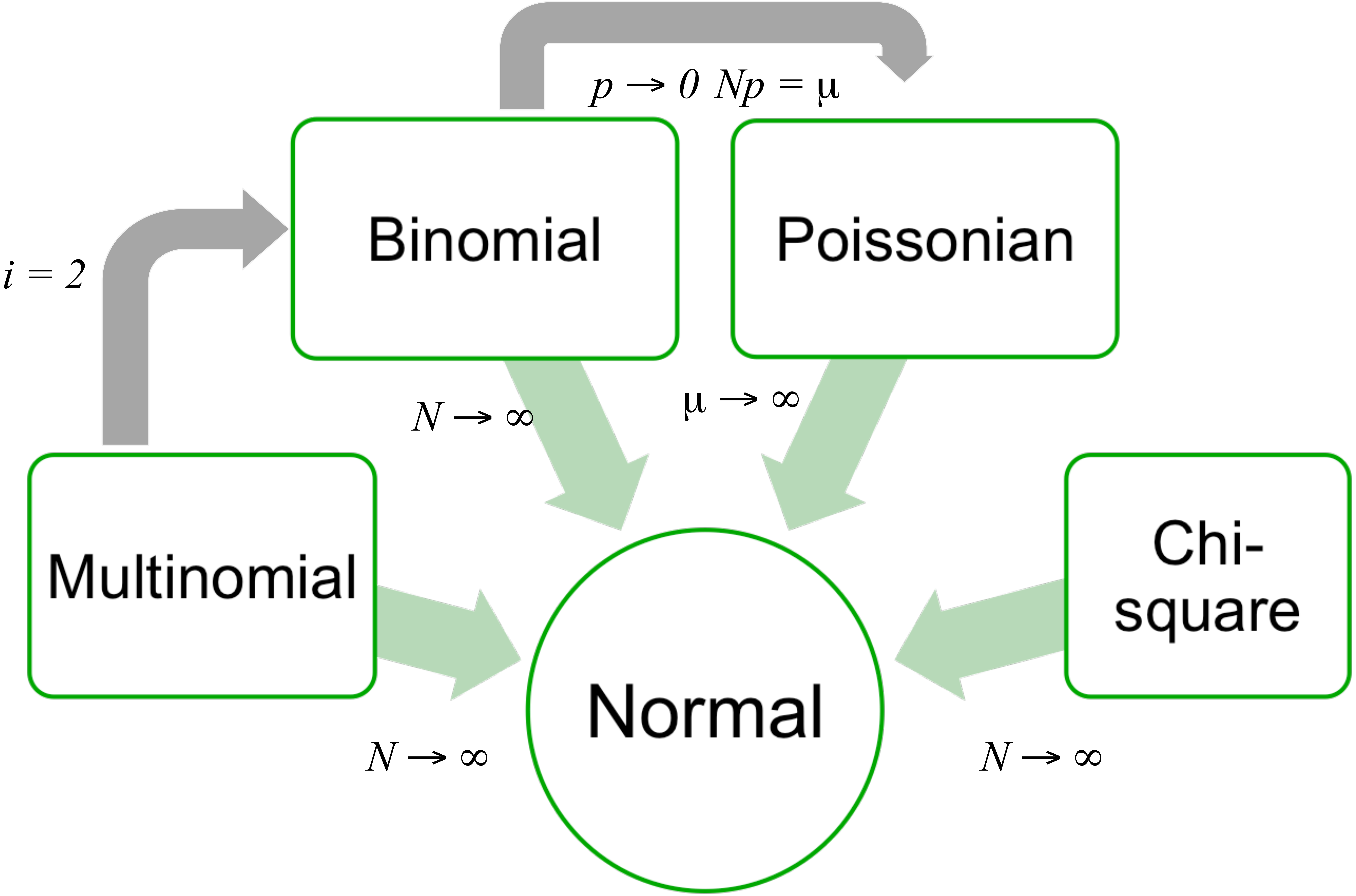
SOME OTHER DISTRIBUTION

- **Uniform** distribution
  - Basic distribution for pseudo-random number generators
- **Gamma** distribution
  - Probability model for waiting time
- **Cauchy** or **Lorentz** or **Breit-Wigner** distribution
  - A solution to the differential equation describing a resonance
  - Energy distribution of a resonance
- **Crystal Ball** distribution
  - Adds an asymmetric power-law tail to a Gaussian PDF
- **Landau** distribution
  - Used to model the fluctuations in the energy loss of particles in thin layers
- **Log-Normal** distribution
  - Used when including systematic errors in the analysis
  - If x is Log-Normally distributed, than log(x) is Normally distributed





ALL ROADS LEAD TO ROME



GENERAL PICTURE REMINDER

2 Sampling reality

EXPERIMENT

1 **Physical phenomena**

Described by a theory

Described by PDFs, depending on unknown parameters with true values

$\theta^{true}=(m_H^{true},\Gamma_H^{true},\ldots,\sigma^{true})$

4 DATA

ANALYSIS

3 **Data sample**

$\mathbf{x} = (x_1,x_2,\ldots,x_N)$

x is a multivariate random variable

5 **Results**

- parameter estimates
- confidence limits
- hypothesis tests

WHAT ARE MONTE CARLO METHODS?

- a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results
- the underlying concept is to use randomness to solve problems that might be deterministic in principle
- mainly used in three problem classes:
  - optimisation
  - numerical integration
  - generating draws from a PDF
- invented in the late 1940s by physicists while he was working on nuclear weapons projects Los Alamos National Laboratory
- the name Monte Carlo, which refers to the Monte Carlo Casino in Monaco

IMPORTANCE OF MC IN SCIENCE

EXPERIMENT

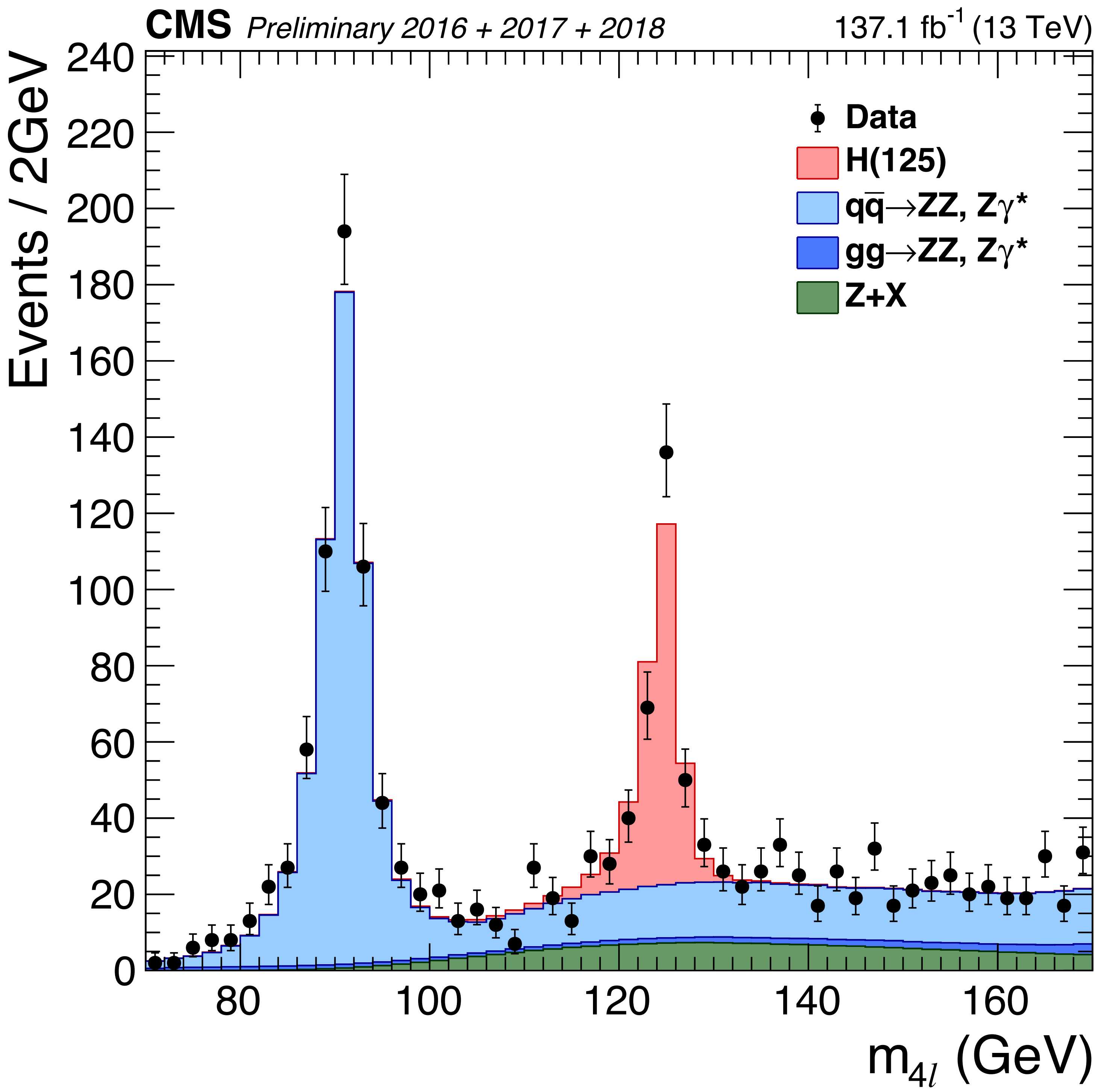
AGREEMENT?

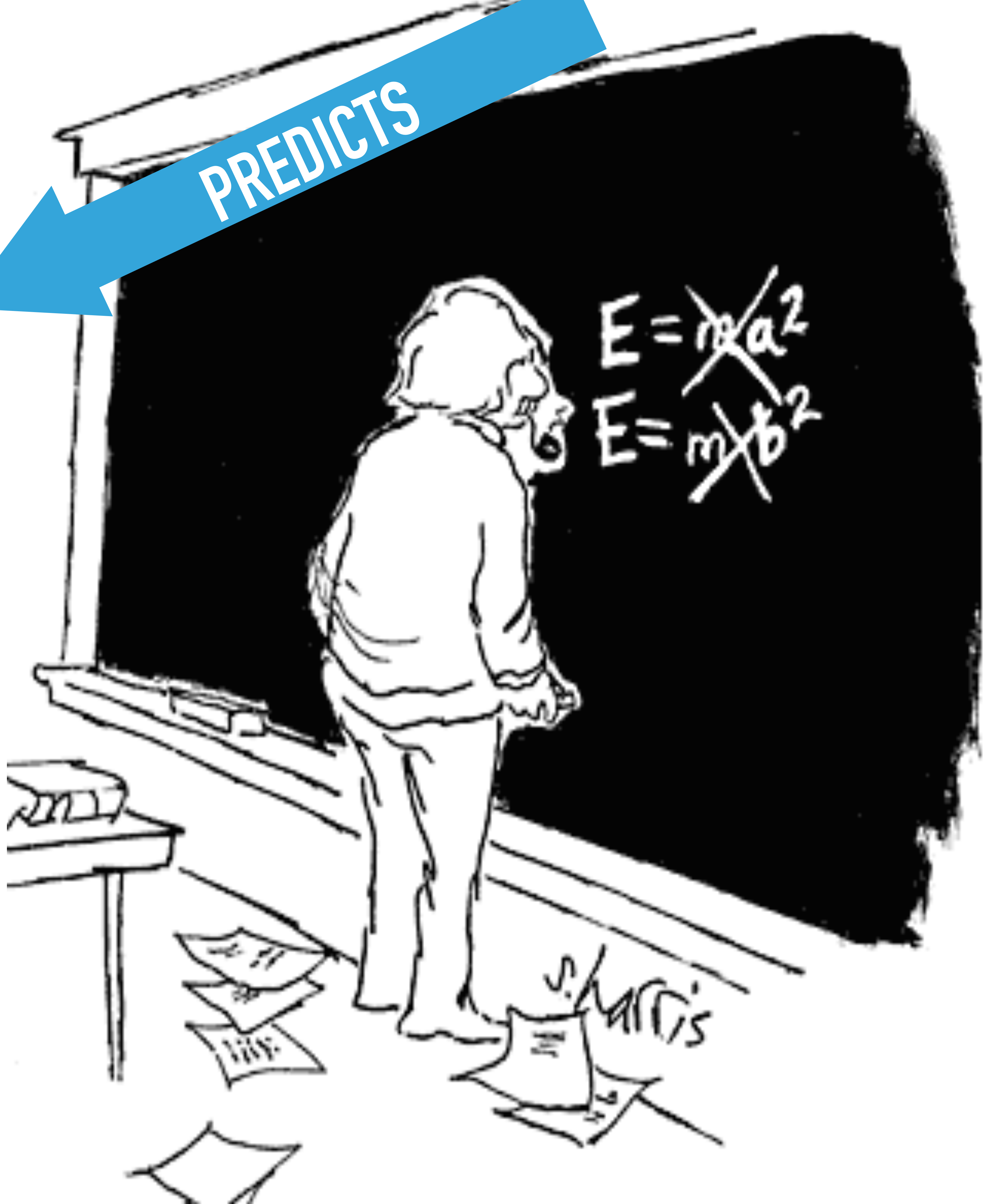
THEORY

MEASURED BY

OBSERVABLES

PREDICTS



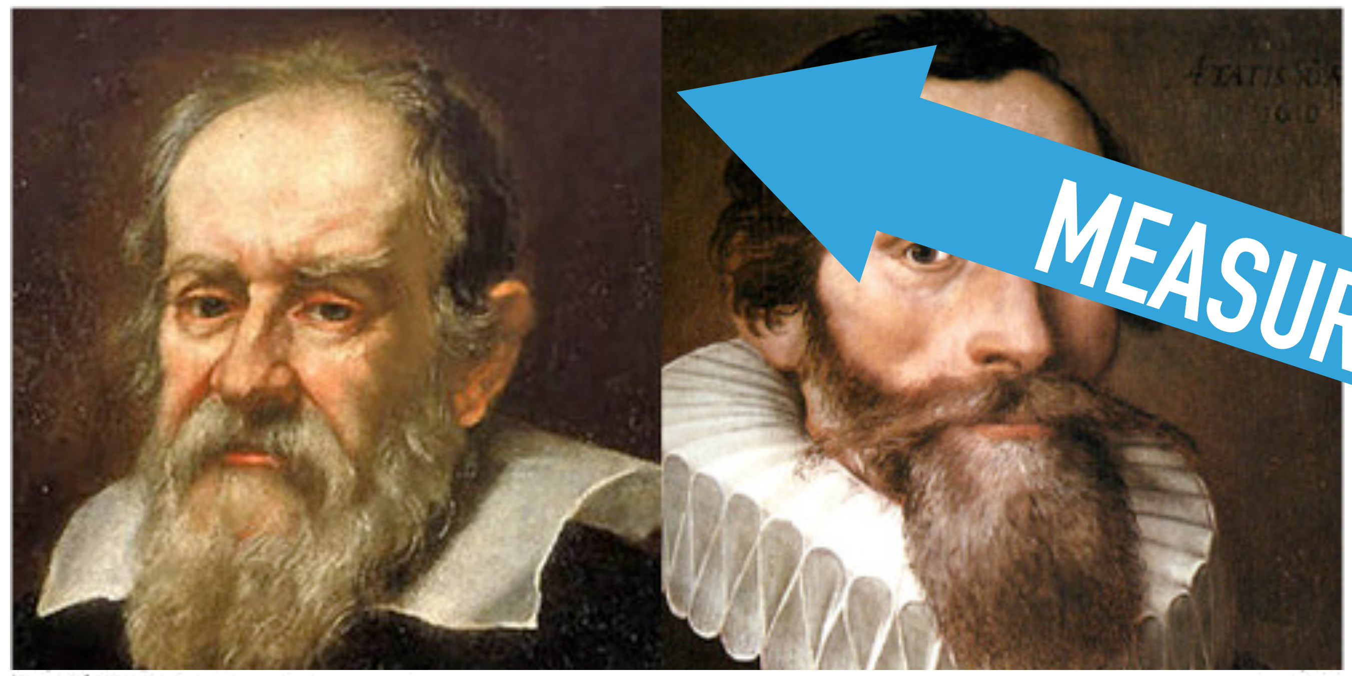


EXAMPLE

EXPERIMENT

AGREEMENT?

THEORY

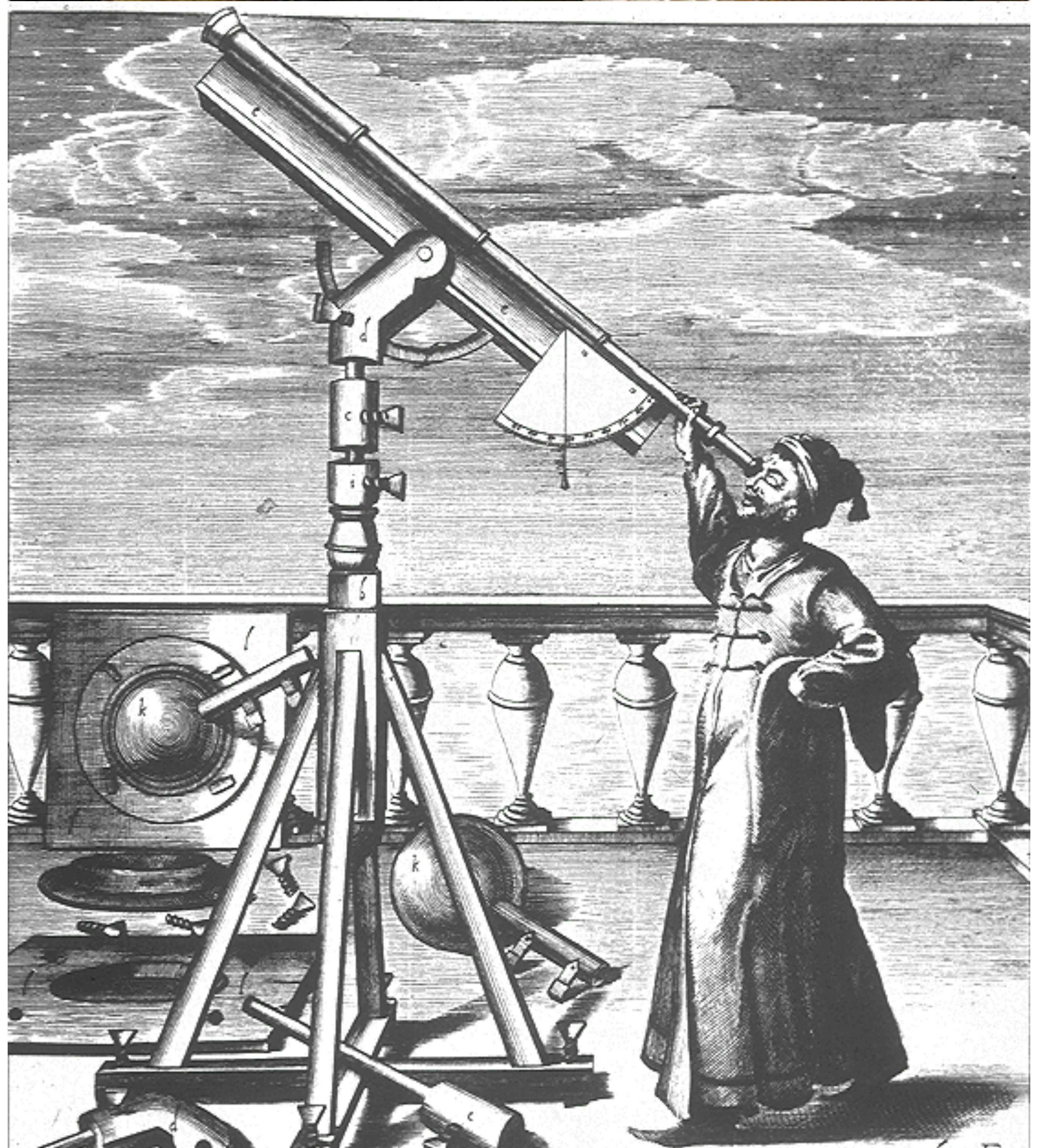


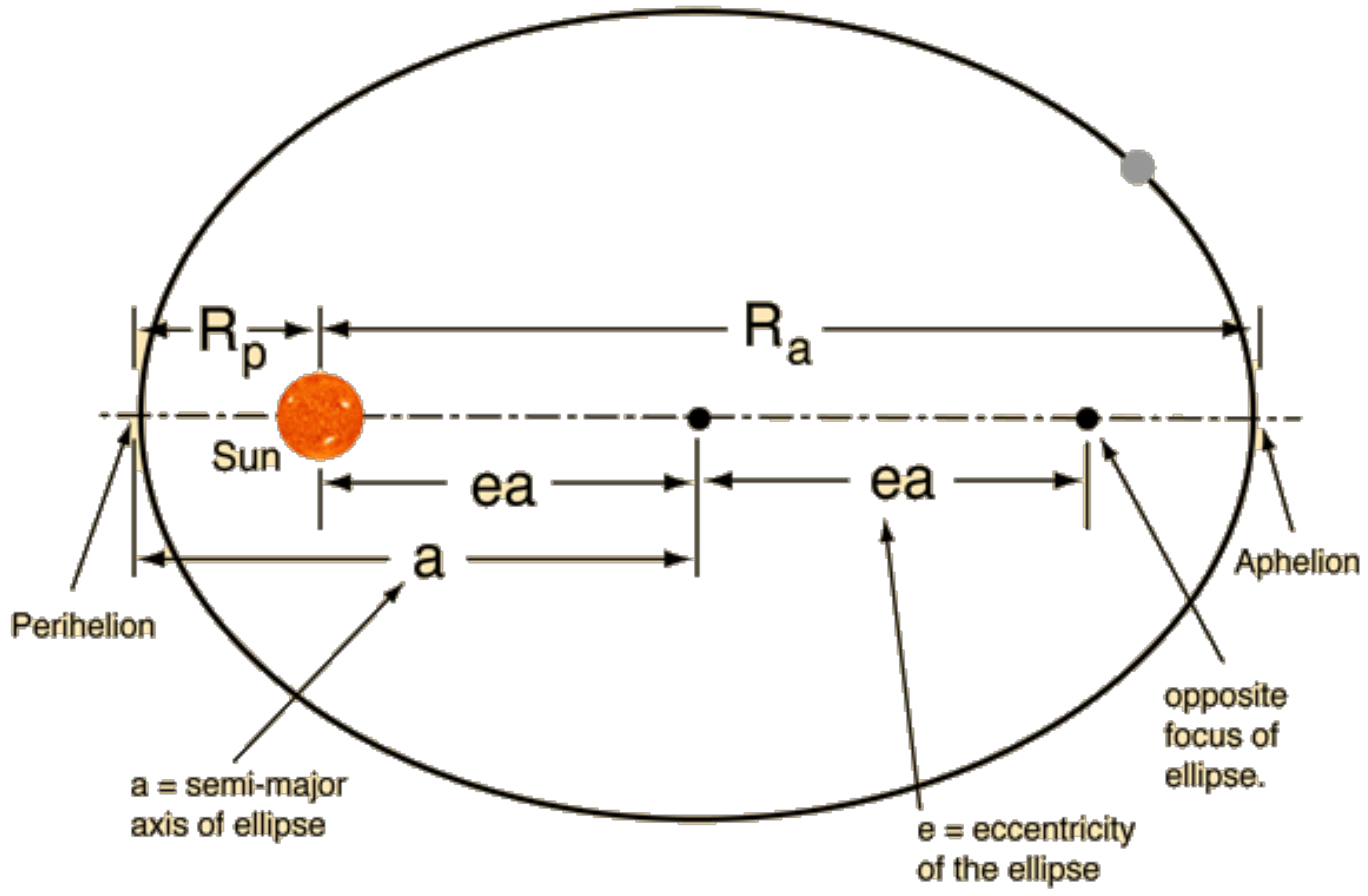
MEASURED BY

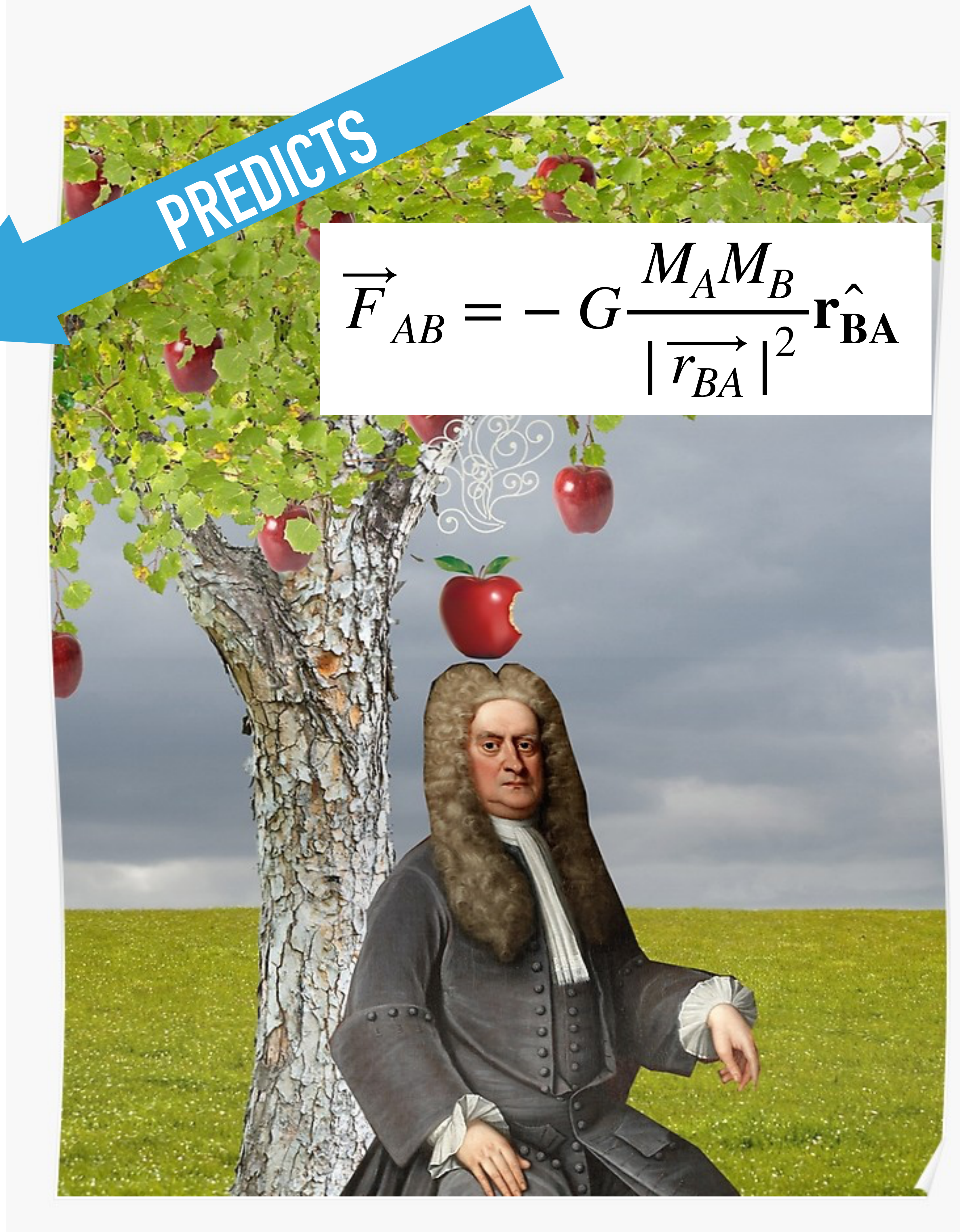
OBSERVABLES

PREDICTS

$$\overrightarrow{F}_{AB} = -G\frac{M_A M_B}{|\overrightarrow{r_{BA}}|^2}\hat{\mathbf{r}}_{\mathbf{BA}}$$







MONTE CARLO WORKFLOW

DEFINE A DOMAIN OF POSSIBLE INPUTS

↓

GENERATE INPUTS RANDOMLY FROM THE DOMAIN

↓

PERFORM A DETERMINISTIC COMPUTATION USING THE INPUTS

↓

AGGREGATE THE RESULTS OF THE INDIVIDUAL COMPUTATIONS INTO THE FINAL RESULT

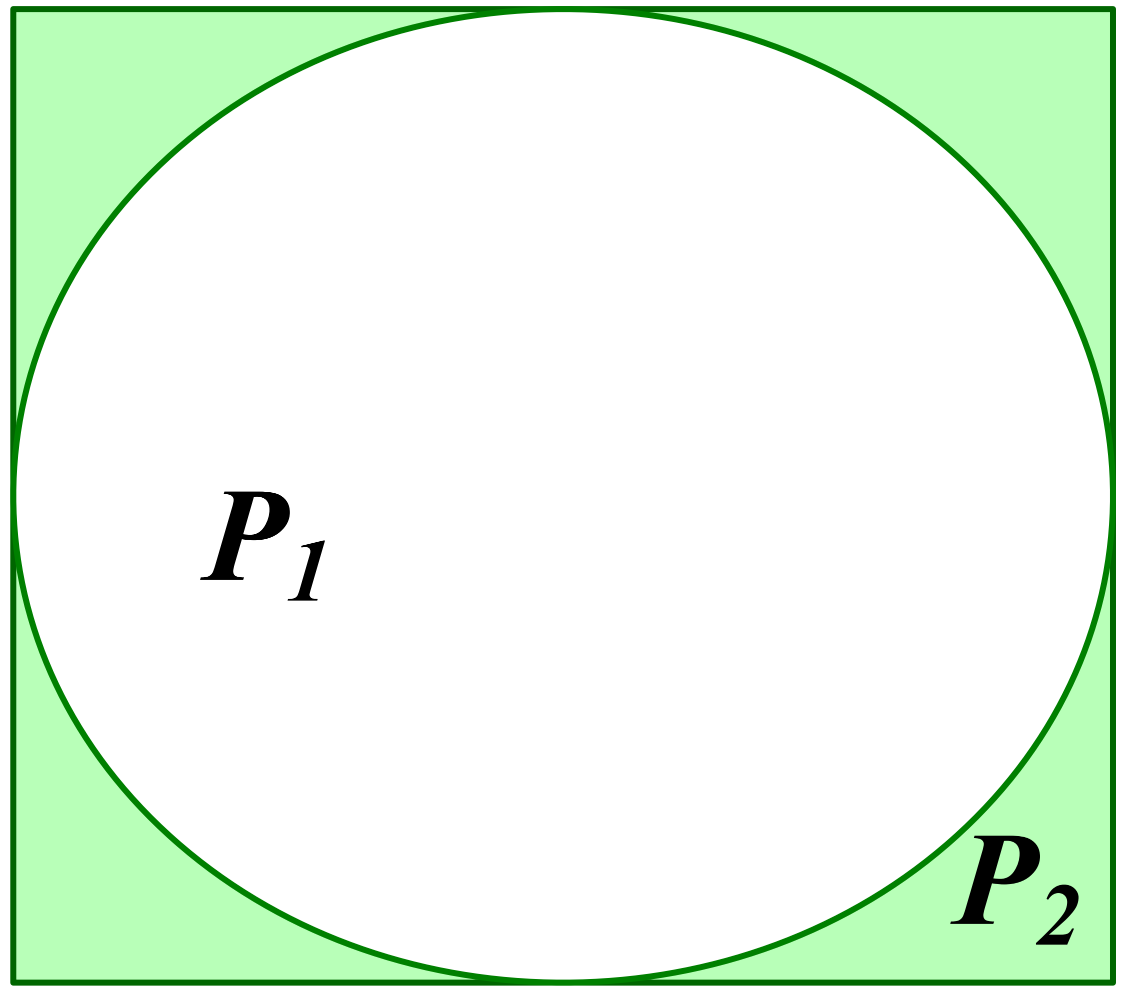
ESTIMATING π

DRAW A SQUARE ON THE GROUND, THEN INSCRIBE A CIRCLE WITHIN IT

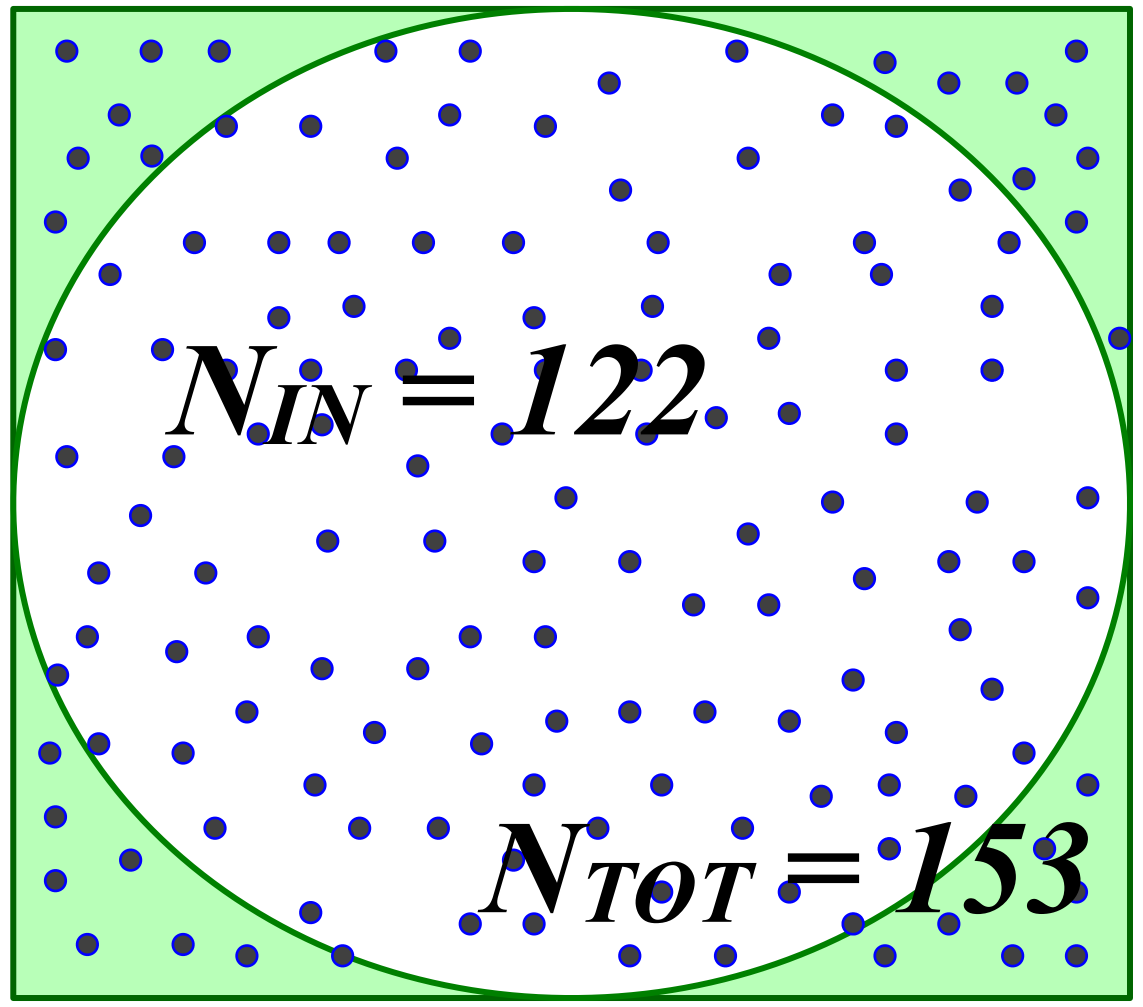
UNIFORMLY SCATTER N OBJECTS OF UNIFORM SIZE THROUGHOUT THE SQUARE.

COUNT NUMBER OF OBJECTS IN THE CIRCLE = N_{IN}

ESTIMATE FINAL RESULT Π ~ $4 \times N_{IN} / N_{TOT}$



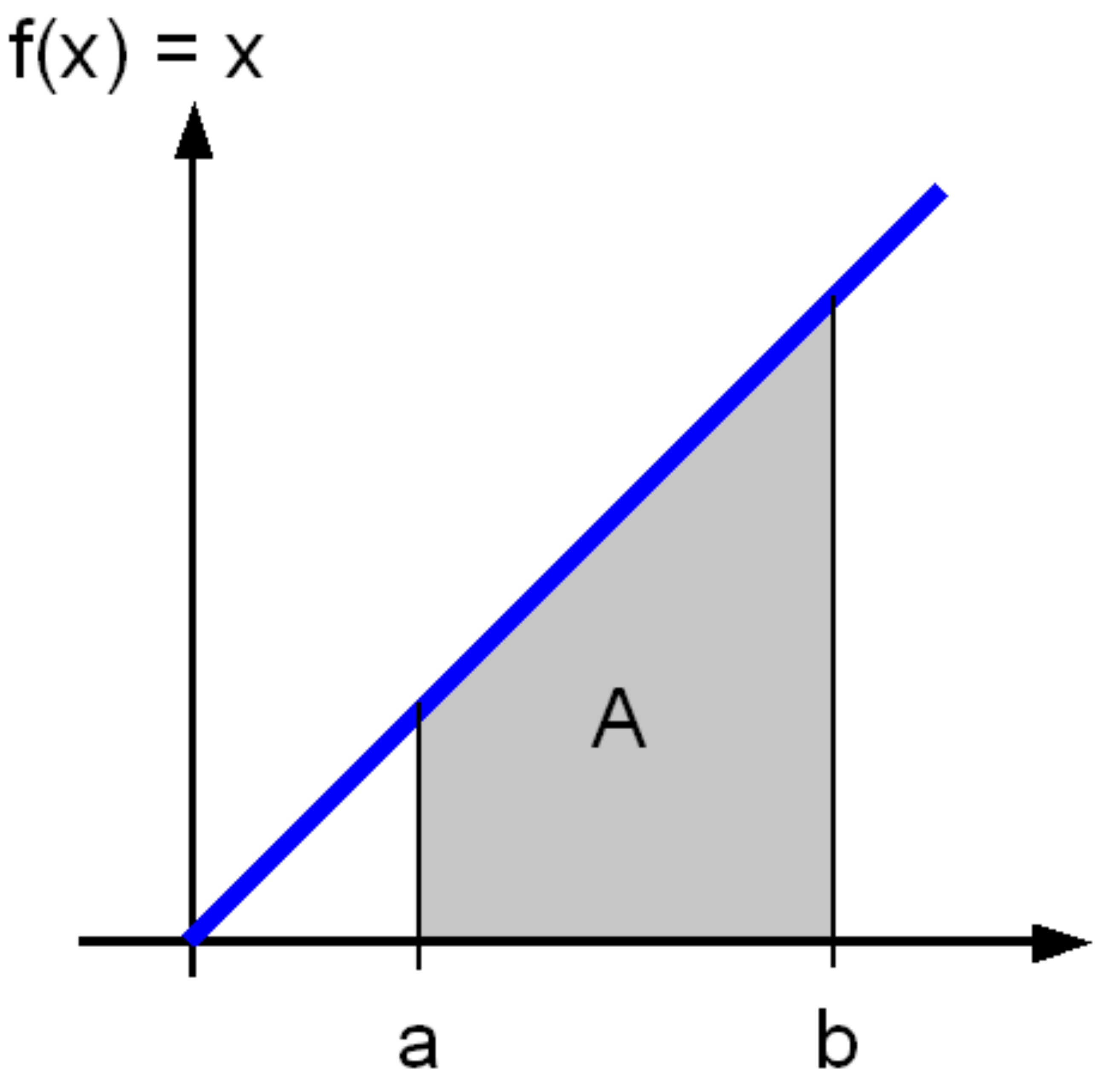
$$P_1 / P_2 = \pi / 4$$



$$\pi \approx 4 \times 122/153 = 3.19$$

SOLVING INTEGRALS

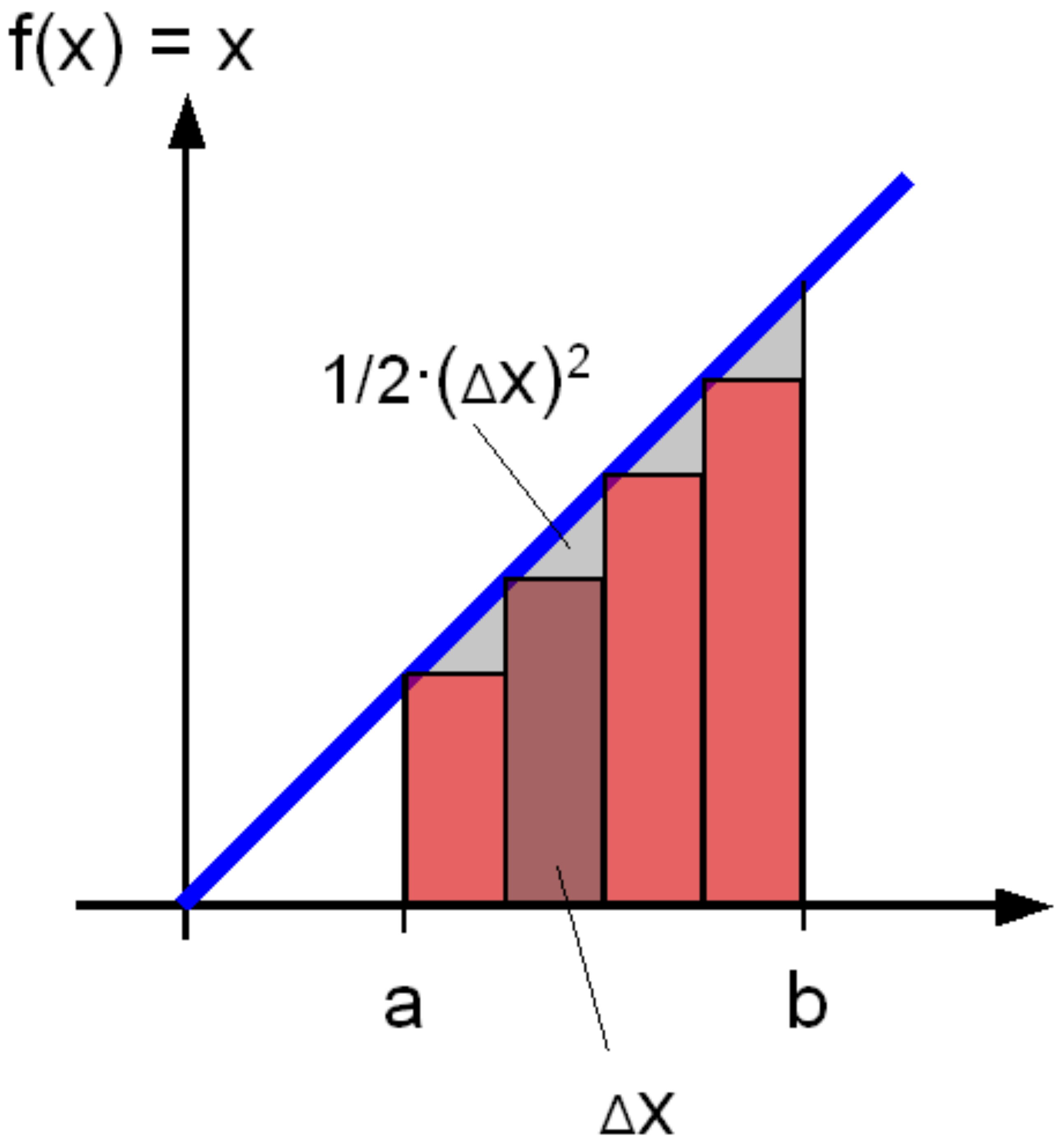
ANALYTICAL SOLUTION



$$\int x dx = \frac{1}{2}x^2$$

$$A = \frac{1}{2}(b^2 - a^2)$$

DETERMINISTIC ALGORITHM

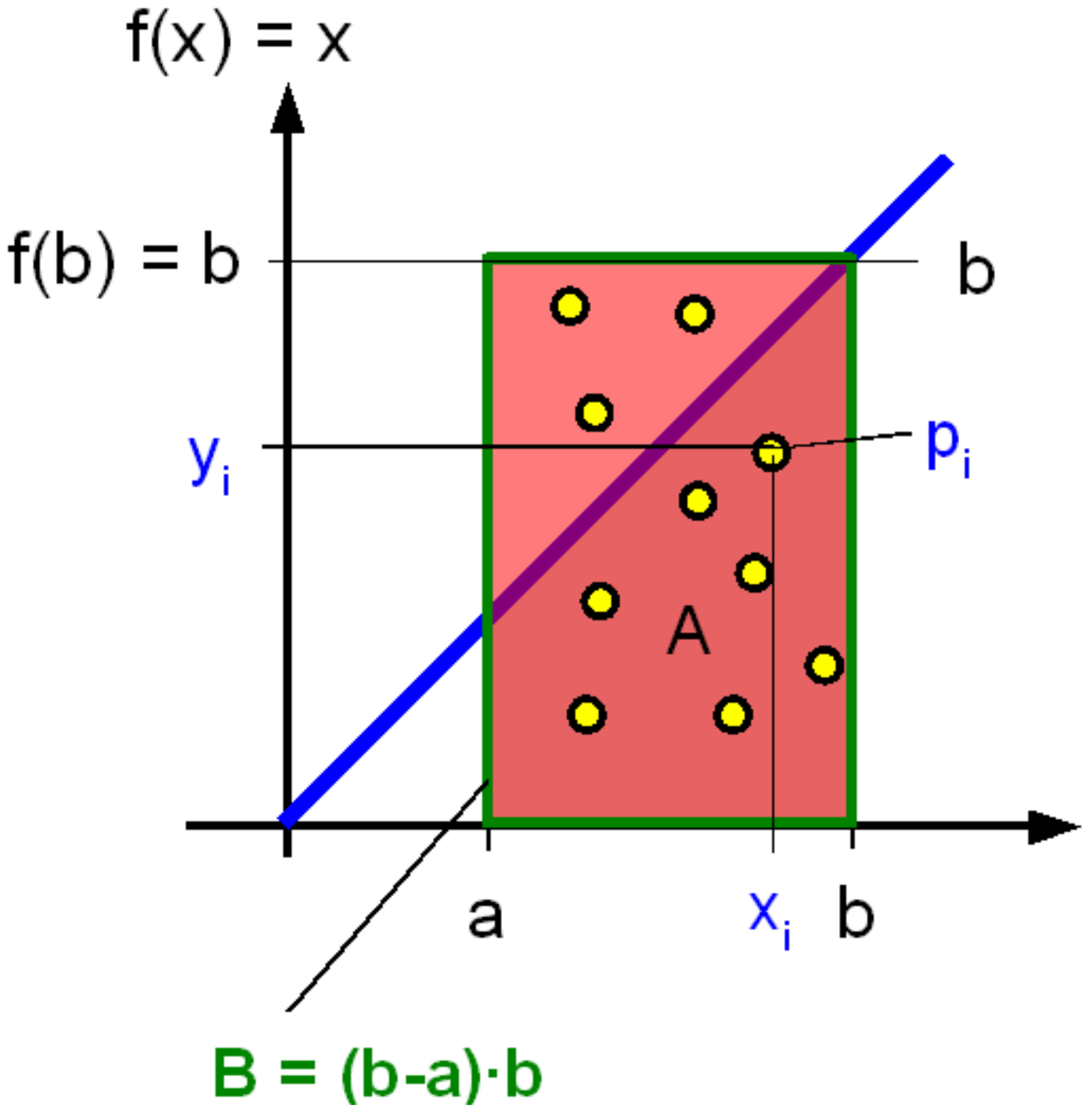


$$\Delta x = \frac{b-a}{n}$$

$$A \approx \sum f(a + i\Delta x)\Delta x$$

gets more precise with more steps (n)

MONTE CARLO

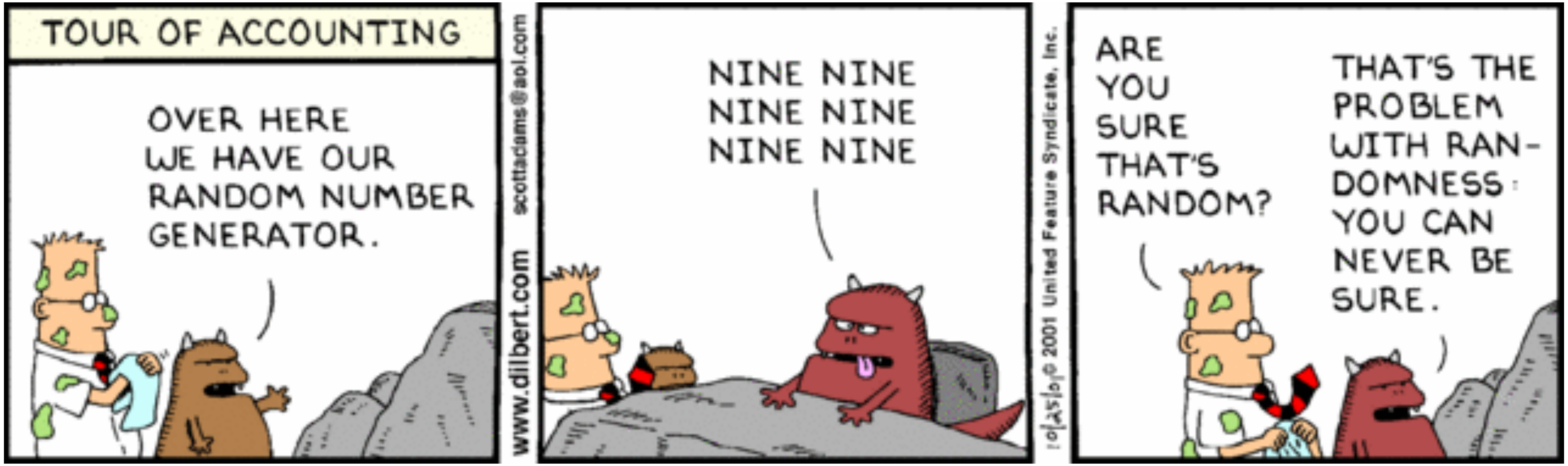


$$A \approx B\frac{N_{IN}}{N_{TOT}}$$

gets more precise with more random number pairs $p_i(x_i,y_i)$

RANDOM NUMBER GENERATION

- Physical methods:
  - “true” random numbers from “unpredictable” process
  - Example: dice, coin flipping, roulette
- True random numbers from random atomic or subatomic physical phenomena:
  - Example: radioactive decay, amplitude of noise in radio
- Computational methods:
  - Pseudo-random number generators create long runs (for example, millions of numbers long) with good random properties but eventually the sequence repeats
  - Example: Linear congruential generator



MC SIMULATION VS REAL LIFE

EVENT GENERATION
TOOLS: MC GENERATORS (PYTHIA, . . .)
OUTPUT: FINAL STATE PARTICLES

DETECTOR SIMULATION
TOOLS: MC SIMULATORS (GEANT)
OUTPUT: SIMULATED DETECTOR RESPONSE

COLLISIONS
TOOLS: ACCELERAORS
OUTPUT: FINAL STATE PARTICLES

DATA ACQUISITION
TOOLS: DETECTORS (CMS, ATLAS,...)
OUTPUT: DETECTOR RESPONSE

EVENT RECONSTRUCTION
TOOLS: DETECTOR SOFTWARE PACKAGES (CUSTOM MADE; MC USED IN ALGORITHMS)
OUTPUT: RECONSTRUCTED PHYSICAL OBJECTS (ELECTRONS, MUONS, JETS . . .)

DATA ANALYSIS
TOOLS: STATISTICS (ROOT, ...; MC USED IN ALGORITHMS; F.G. TOY MC)
OUTPUT: NEW KNOWLEDGE (PARAMETER/INTERVAL ESTIMATES, COST/PERFORMANCE ESTIMATE, HYPOTHESIS TESTS, ARTICLE, TALKS ...)