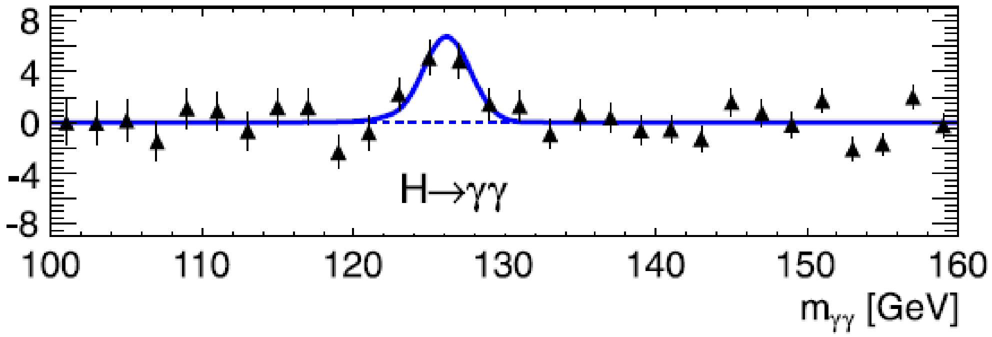
Introduction to
Statistical Data Analysis

Chapter 3

Special Distribution Functions



Nov 2023

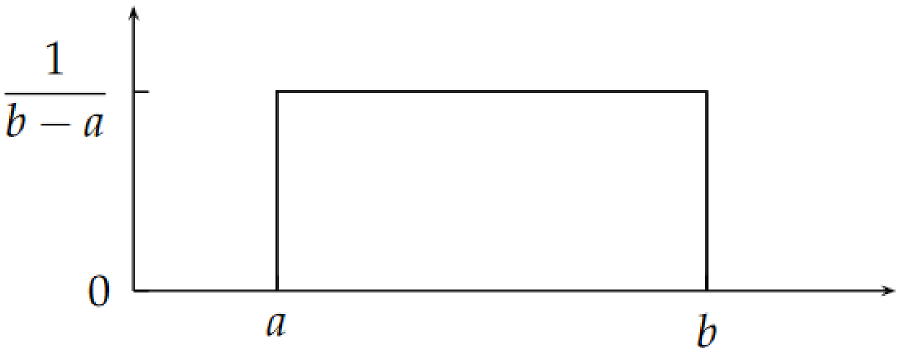
Overview

In this chapter we'll consider the following probability distribution functions used in Nuclear and Particle Physics:

Distibution Function	In Root **TRandom** Class
▪ Uniform Distribution	`Double_t Uniform(Double_t x1 = 1)`
▪ Exponential Distribution	`Double_t Exp(Double_t tau)`
▪ Binomial Distribution	`Int_t Binomial(Int_t ntot, Double_t prob)`
▪ Poisson Distribution	`Int_t Poisson(Double_t mean)`
▪ Gaussian Distribution	`Double_t Gaus(Double_t mean=0, Double_t sigma=1)`
▪ Landau Distribution	`Double_t Landau(Double_t mean=0, Double_t sigma=1)`
▪ Breit-Wigner Distribution	`Double_t BreitWigner(Double_t mean=0, Double_t gamma=1)`
▪ Student's t Distribution	
▪ X^2 Distribution	

Uniform Distribution

$$f(x;a,b)=\begin{cases}\frac{1}{b-a}, & a\le x\le b\\ 0, & \text{otherwise}\end{cases}$$



$$\langle x\rangle=\mu_X=E[X]=\frac{a+b}{2}$$

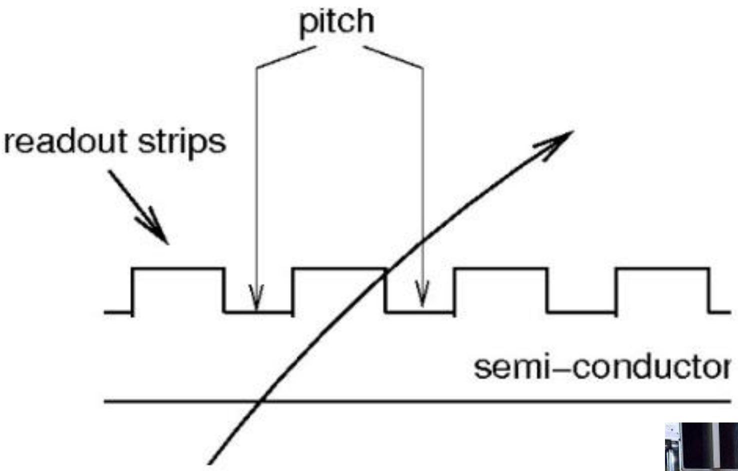
$$\sigma^2=V[X]=\frac{(b-a)^2}{12}$$

Example:

Silicon strip detector:

resolution for one-strip clusters:

$$\sigma=\frac{pitch}{\sqrt{12}}$$

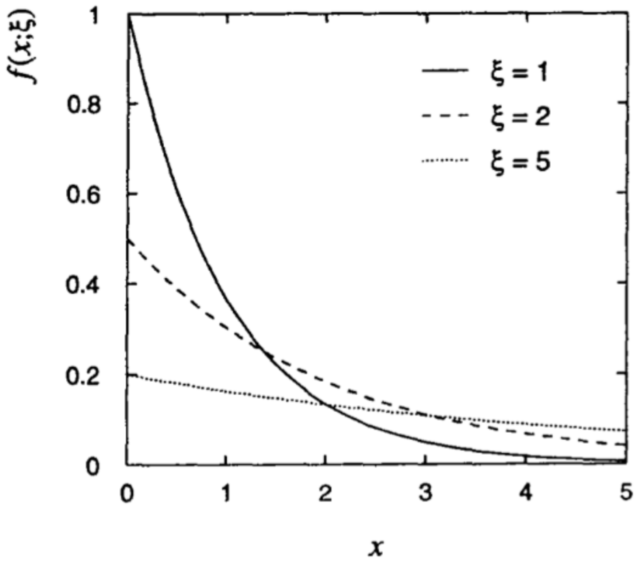


Exponential Distribution

$$f(x;\xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\langle x \rangle = \mu_X = E[X] = \xi$$

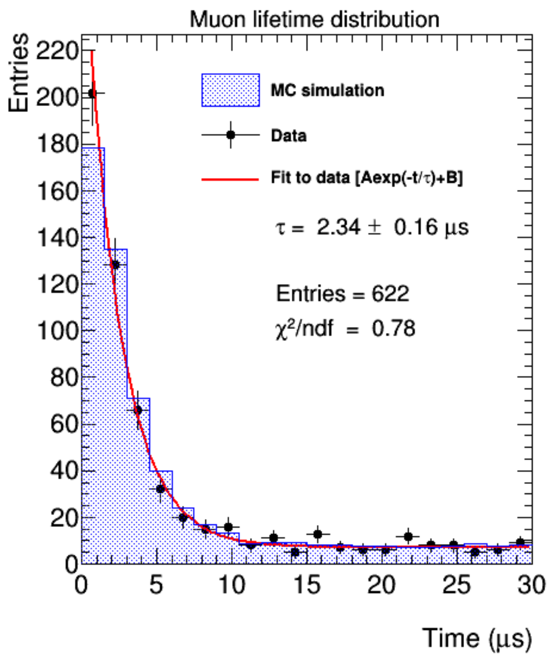
$$\sigma^2 = V[X] = \xi^2$$



Example:

Decay time of an unstable particle at rest:

$$f(t,\tau) = \frac{1}{\tau} e^{-t/\tau}$$



http://www1.gantep.edu.tr/~bingul/muon

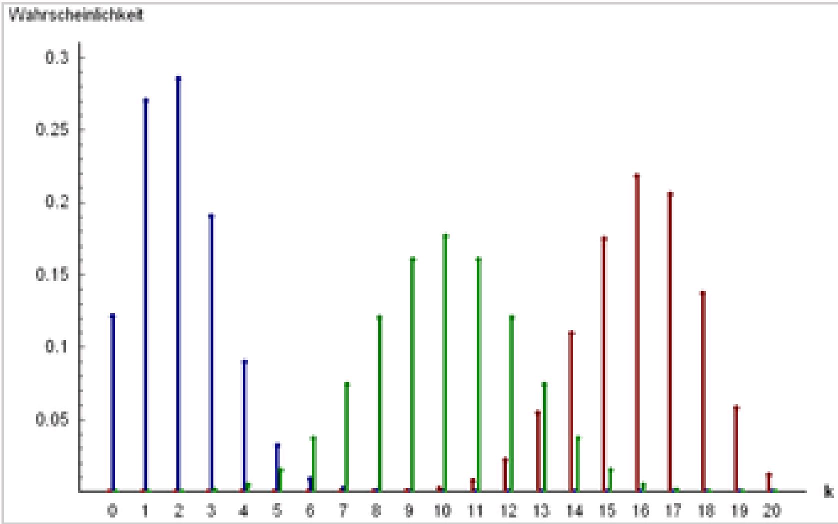
Binomial Distribution

The binomial distribution function specifies the number of times (k) that an event occurs in n independent trials. If p is the probability of the event occurring in a single trial, then:

$$f(k,n,p) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}$$

$$\langle x \rangle = \mu_X = E[X] = np$$

$$\sigma^2 = V[X] = np(1-p)$$



Binomial distribution for $n = 20$
$p = 0.1$ (blue), $p = 0.5$ (green) and $p = 0.8$ (red)

* Use binomial distribution to model processes with two outcomes: success or failure
* Trials are independent
* p is constant from one trial to another.

Example:

Detection efficiency (either we detect particle or not).

Example 1

A coin is thrown 30 times.

(a) Calculate the mean (expected) number heads and standard deviation

$$\langle x \rangle = np = (30)(0.5) = 15$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{(30)(0.5)(0.5)} = 2.74$$

(b) Imagine you observed 20 heads. Compute how many standard deviations your observation differ from the mean value. Is the coin fair?

$$N = \frac{20-15}{2.74} = 1.83$$

N < 3 sigma

20 heads is consistent with 15 => the coin is fair

(c) Imagine you observed 30 heads. Compute how many standard deviations your observation differ from the mean value. Is the coin fair?

$$N = \frac{30-15}{2.74} = 5.47$$

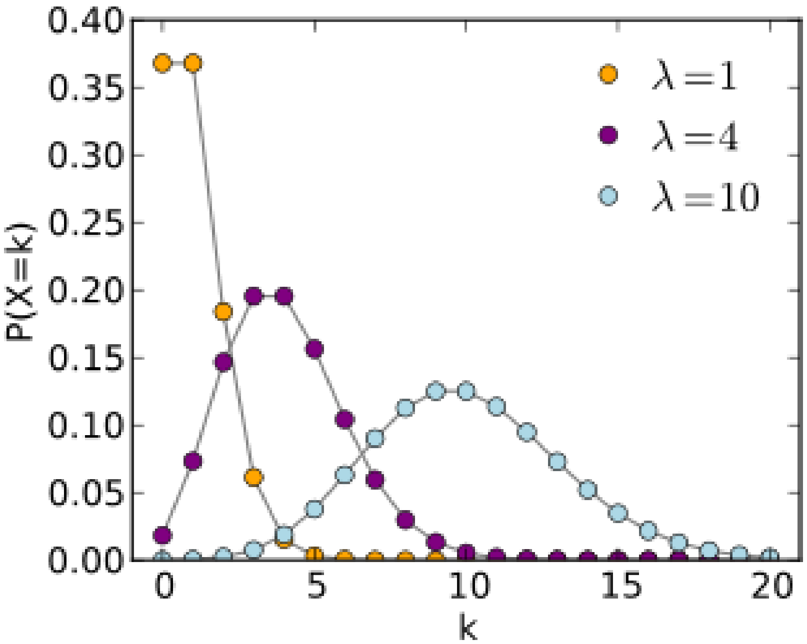
N > 5 sigma

20 heads is not consistent with 15.

Discovery => the coin is not fair

Poisson Distribution

In the binomial equation, if the probability p is so small then the distribution of events can be approximated by the Poisson distribution.



$$f(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

$$\langle x \rangle = \mu_X = E[X] = \lambda$$

$$\sigma^2 = V[X] = \lambda$$

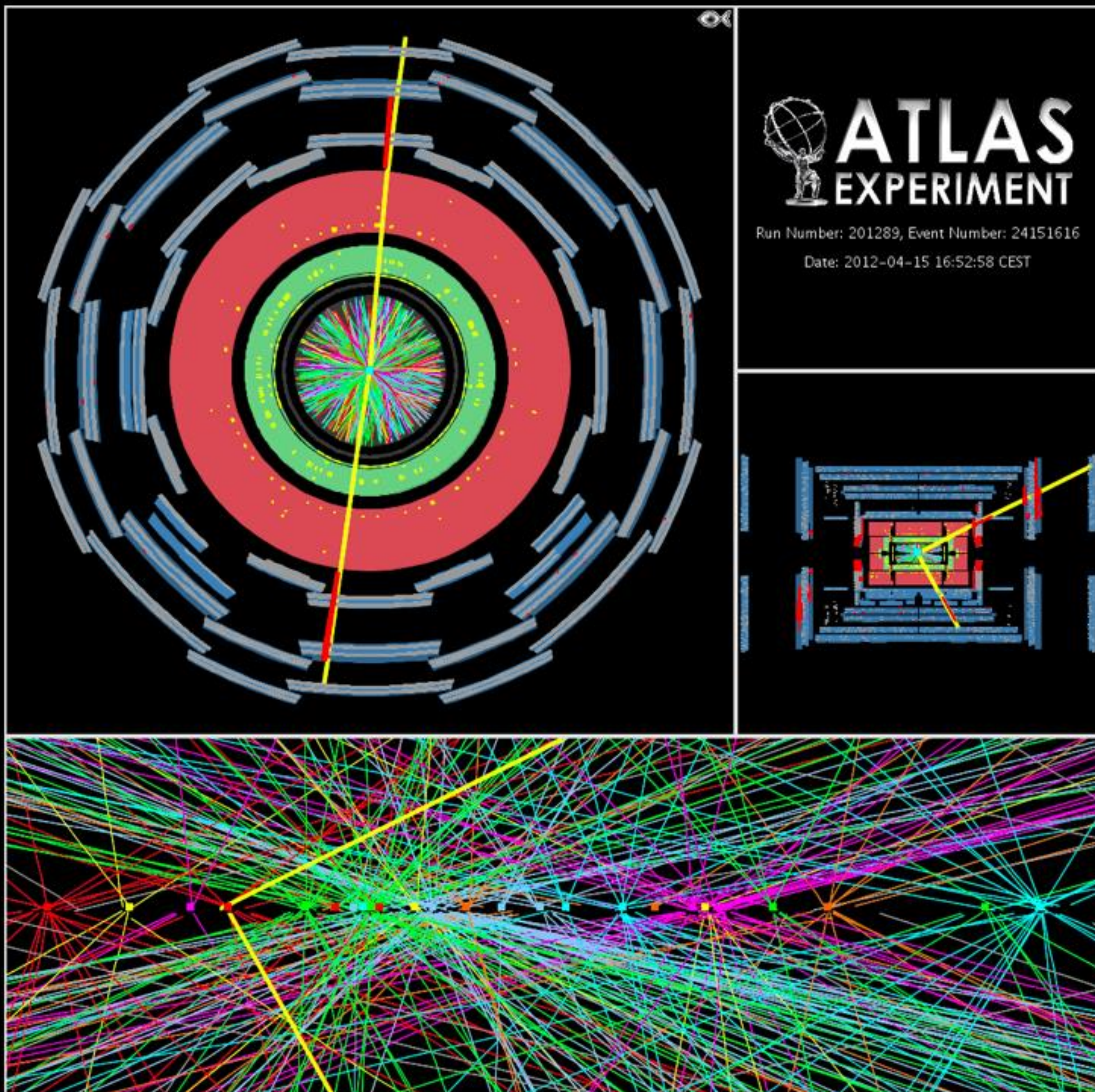
Examples:

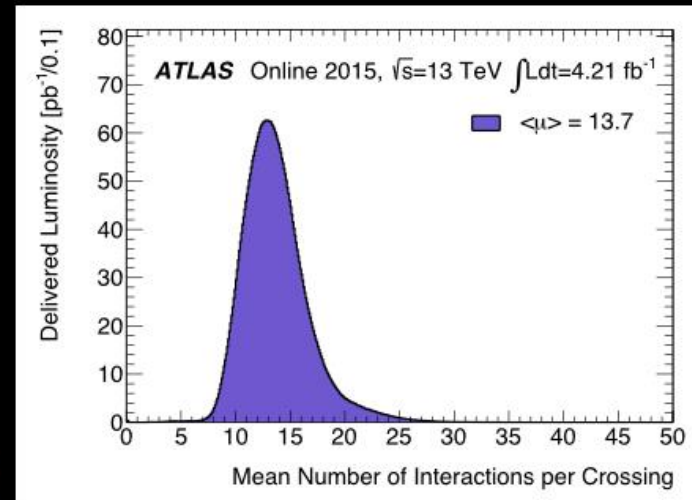
* Clicks of a Geiger counter in a given time interval

* Mean number of p-p interactions per bunch crossing at LHC (pile-up events)

* Number of atmoshperic muons passing through unit area per unit time

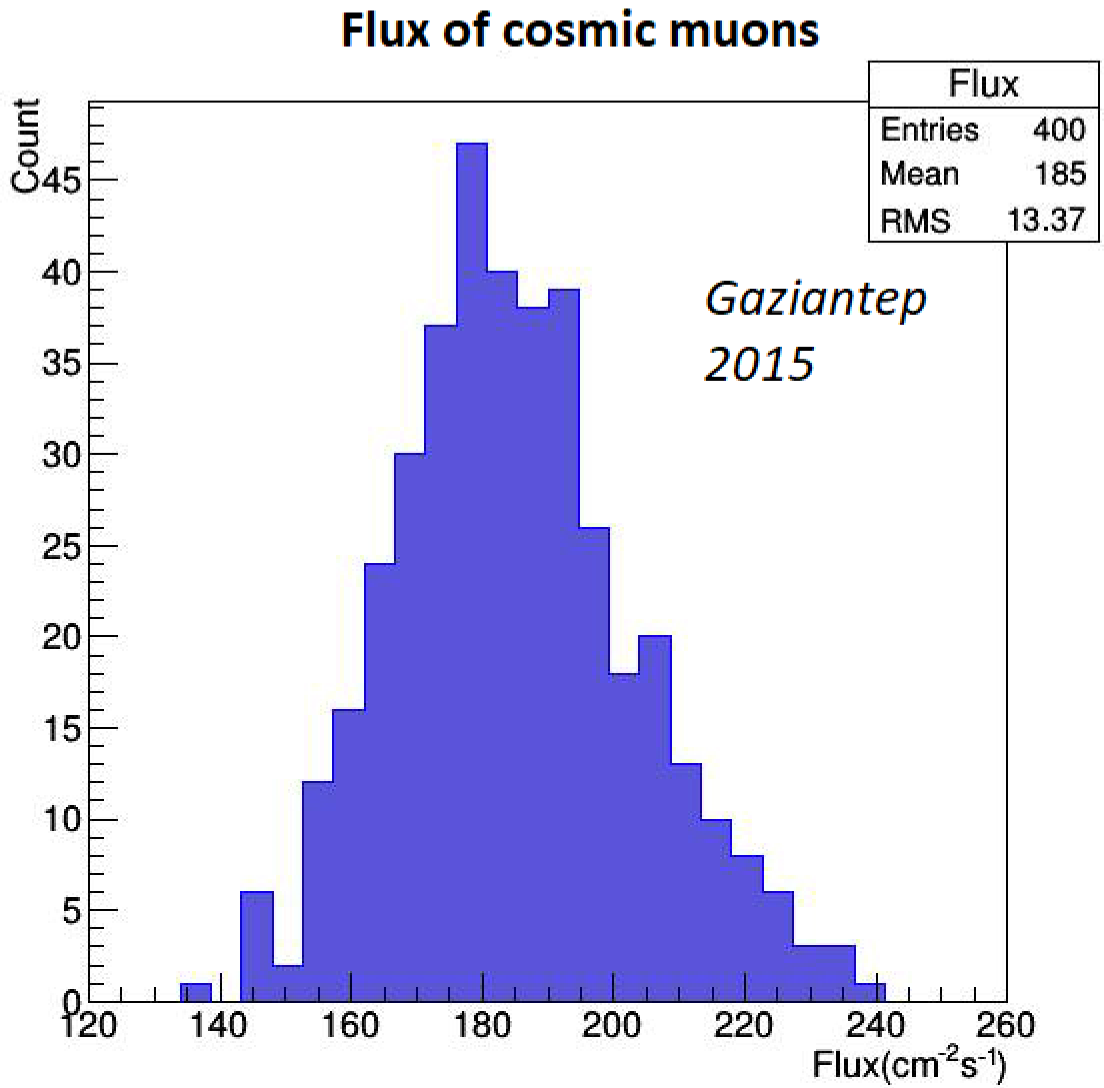
* Number of photons generated in Cherenkov Radiation process





[4] The beat.

Something deep and subtle, with the number of taps per second based on "pileup", the average number of collisions per proton-bunch crossing in 2015. Here it's 16.


Flux of cosmic muons
Flux
Entries 400
Mean 185
RMS 13.37
Gaziantep
2015
Count
0
5
10
15
20
25
30
35
40
45
120
140
160
180
200
220
240
260
Flux(cm-2s-1)

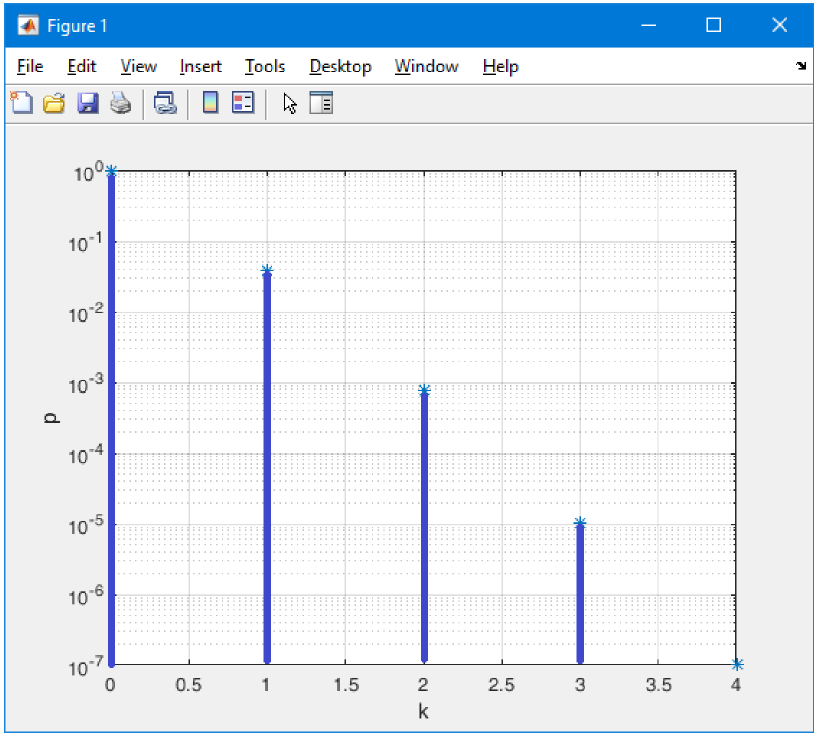
Example 2

Inefficency of a muon detector is 1%. Determine the probability of detecting single Higgs Boson from the decay channel: *H* → *ZZ** → $\mu^+\mu^-\ \mu^+\mu^-$

$$\lambda = np = (4)(0.01) = 0.04$$

$$p(k,\lambda) = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-0.04}0.04^4}{4!} = 1\times10^{-7}$$

k	Probability
0	0.96078943915232
1	0.03843157756609
2	0.00076863155132
3	0.00001024842068
4	0.00000010248420



Hence, we can detect *H* with %96 probability.

Example 3

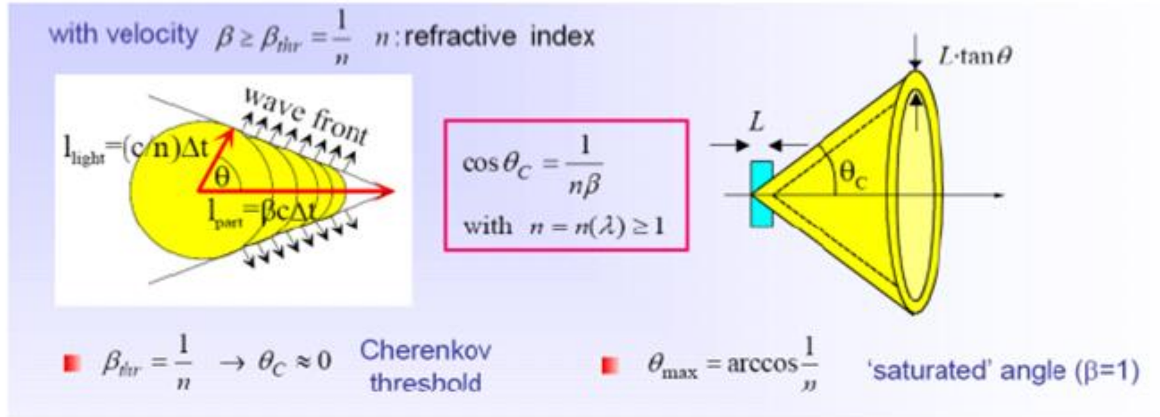
If the velocity of a charged particle is larger than the velocity of light in the medium $v > c / n$ (n: refractive index), it emits 'Cherenkov Radiation' with cone angle:

$$\cos\theta_c = \frac{1}{n\beta} \qquad (\beta = v/c)$$

Number of photons generated (N) per unit length (dx) for the wavelength λ can be found from:

$$\frac{d^2N}{dx\,d\lambda} = \frac{2\pi\alpha}{\hbar c\lambda^2}\left(1 - \frac{1}{\beta^2 n^2}\right)$$

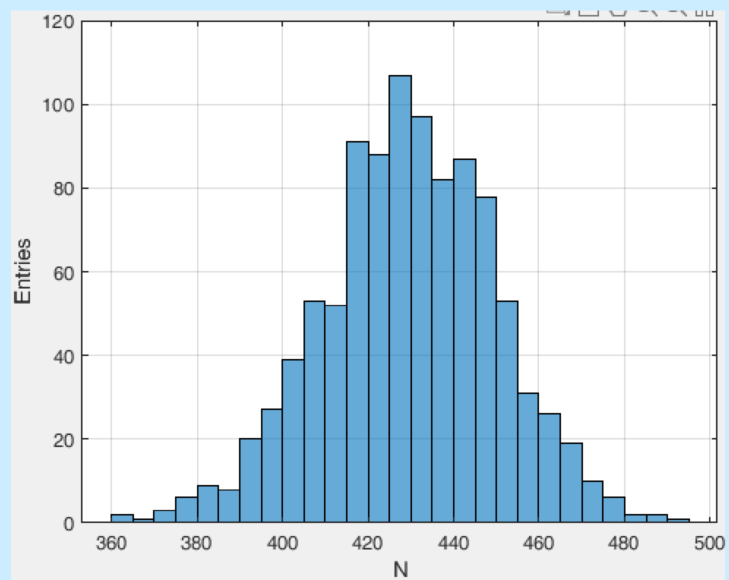
Do not forget dispersion,
i.e. $n = n(\lambda)$



Problem: Calculate number of generated photons/cm for the visible light (400-700 nm) in water (n=1.33) for charged particle of veloctiy beta ~ 1.

$$\frac{dN}{dx} = \int_{400nm}^{700nm} \frac{2\pi}{137}\left(1 - \frac{1}{1.33^2}\right)\frac{d\lambda}{\lambda^2} = 215 / cm$$

If dx = 2 cm => N = 430. Dist. of N for 1000 particles:



Gaussian (Normal) Distribution

In Statistics, if the number of events is very large (*n*>30), then the Gaussian (normal) distribution function may be used to describe nearly all events.

The Gaussian distribution is a continuous Random Variable of the form:

$$p_{gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\text{mean}: \quad \mu$$

$$\text{std.dev}: \quad \sigma$$

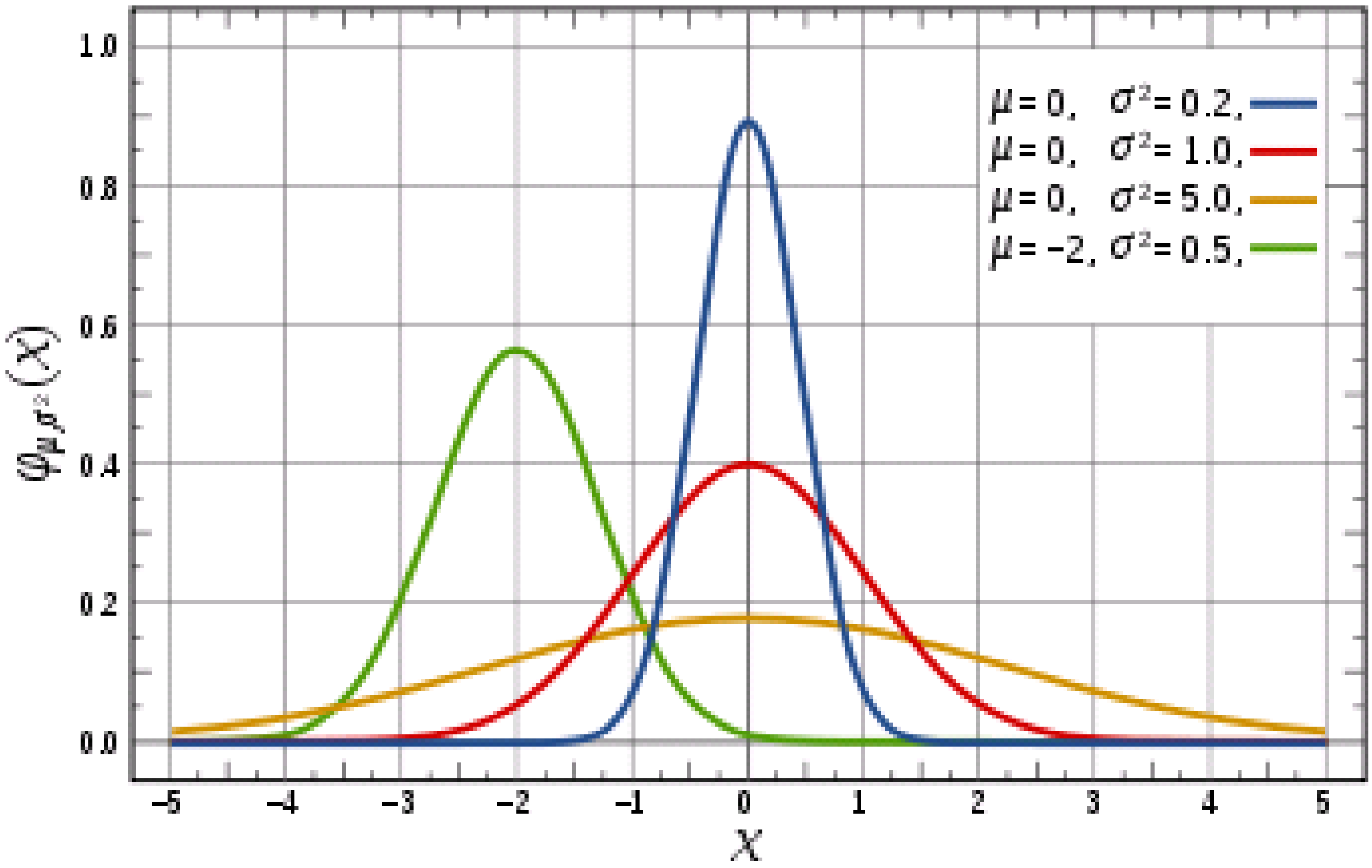
$$p_{gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ = mean

σ = standart deviation

π = 3.141593

e = 2.718281



Properties of Gaussian Function

$$p_{gauss}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

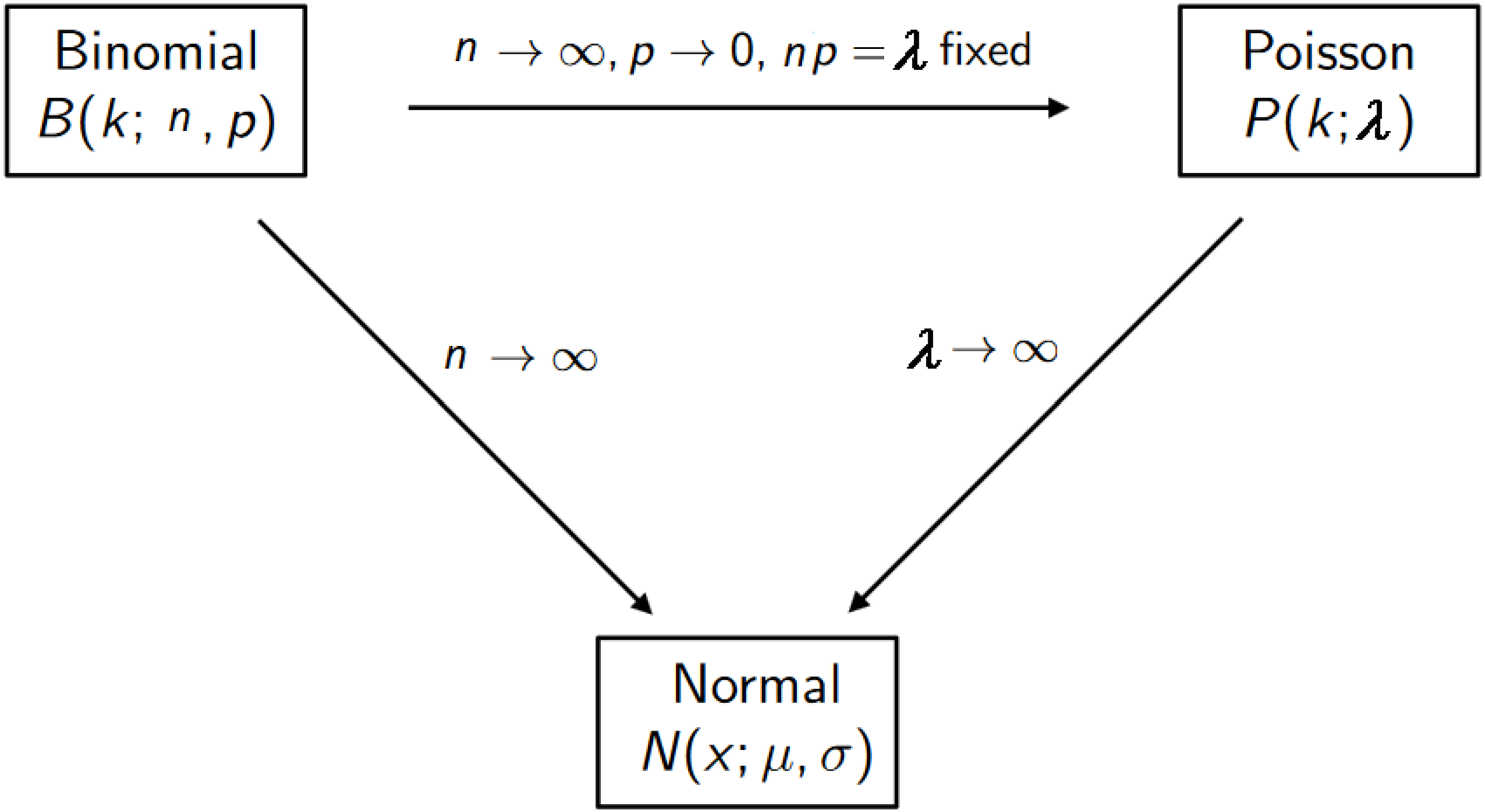
$$p(x) \geq 0$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

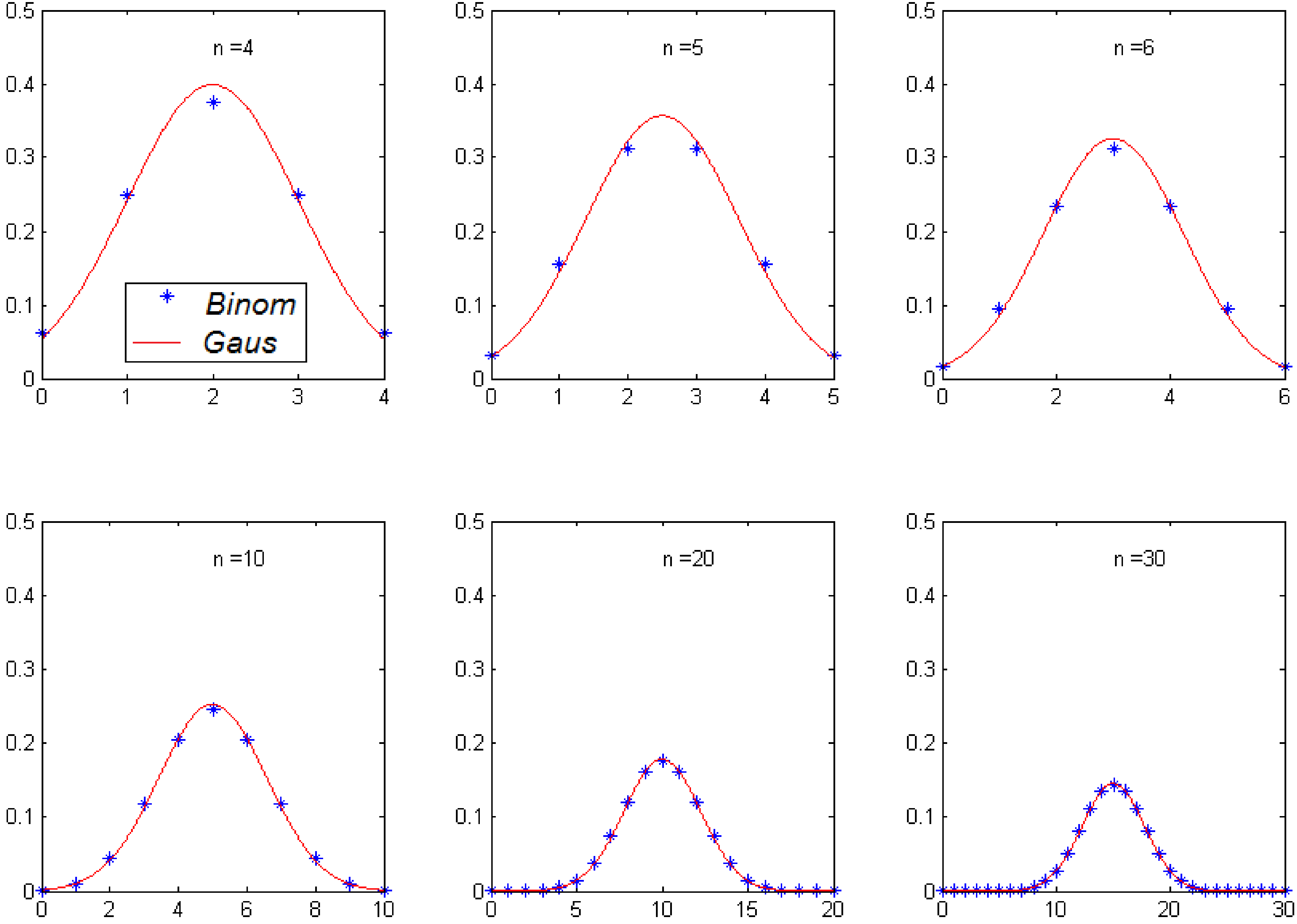
$$< x >= E[X] = \int_{-\infty}^{+\infty} xp(x)dx = \mu$$

$$\int_{a}^{b} p(x)dx = P(a \leq x \leq b)$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx$$


Binomial
$B(k; n, p)$
$n \to \infty, p \to 0, np = \lambda$ fixed
Poisson
$P(k; \lambda)$
$n \to \infty$
$\lambda \to \infty$
Normal
$N(x; \mu, \sigma)$

Binomial Approximation to Gaussian Function for p = 0.5

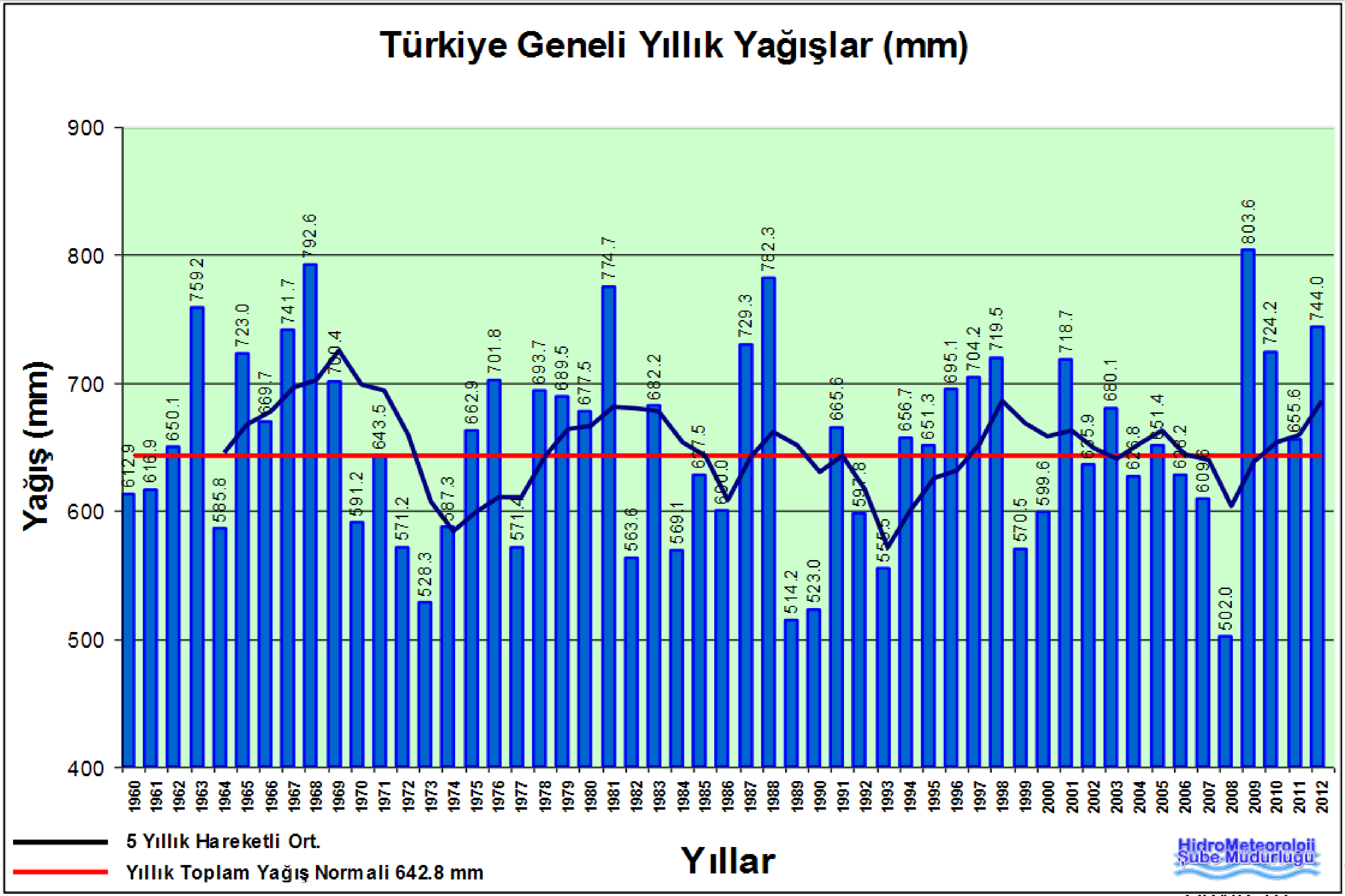


Example Normal Distributions

- Here we will examine some interesting **real data** whose values are distributed normally.

- For each example, histogram of the data is fitted to a Gaussian Function.

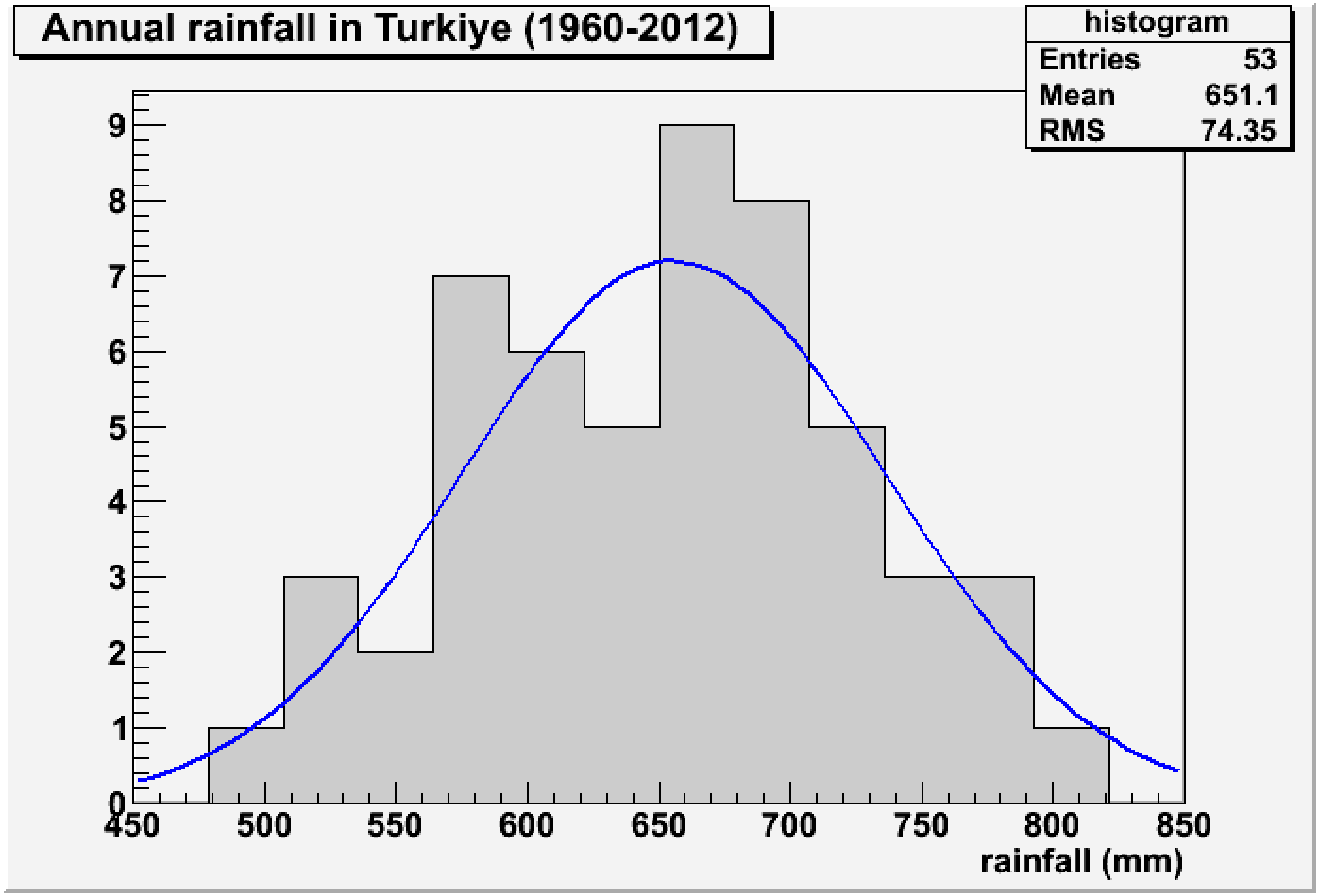
Annual Rainfall (1960-2012)



Annual Rainfall (1960-2012)

Mean : <x> = 651.10 mm

Std. Dev. : σ = 74.35 mm

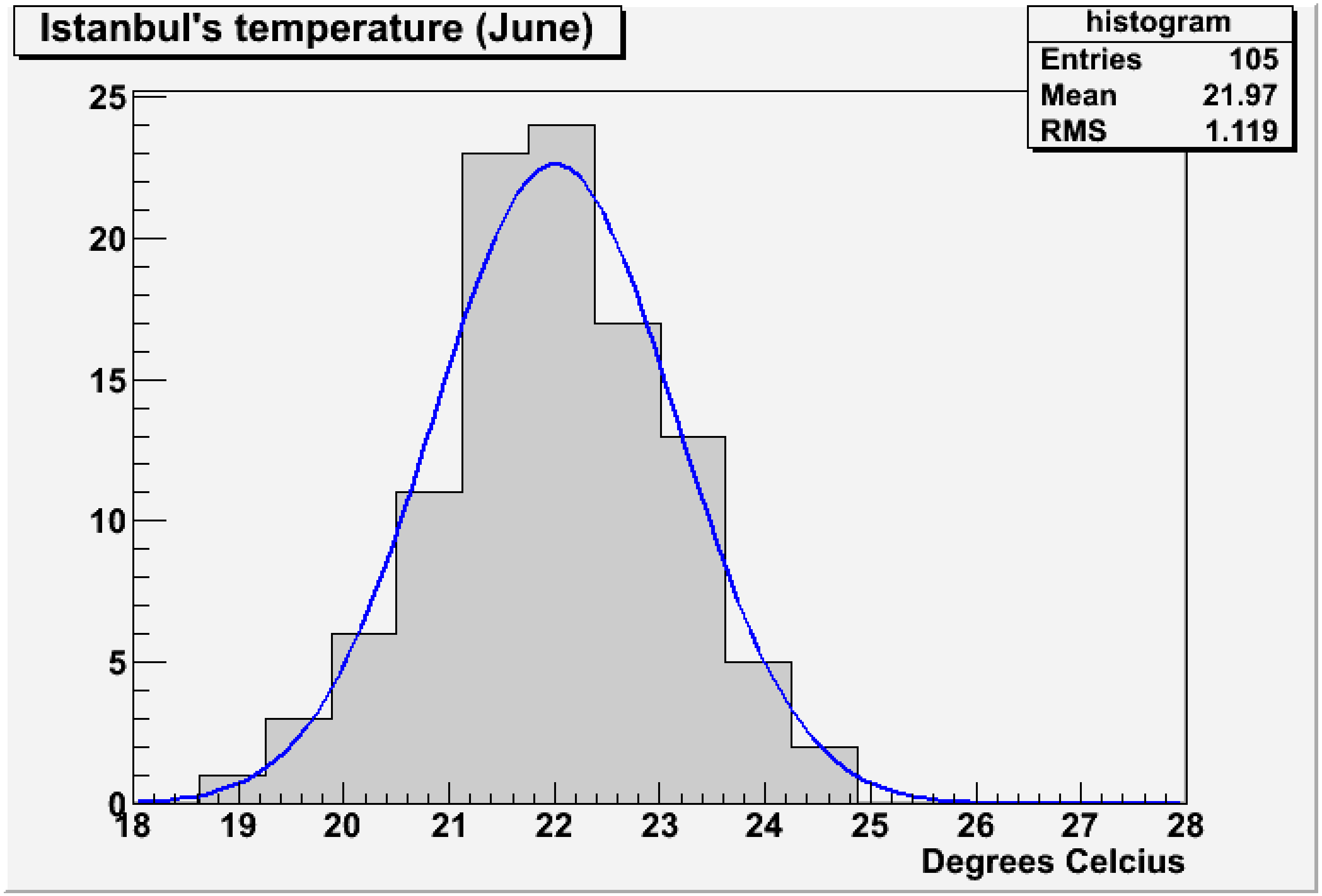


Data: http://www.mgm.gov.tr

Air temperature in Istanbul for the last 105 years.

Mean temperature: $\langle x \rangle = 21.97\ ^oC$

Std. Dev. : $\sigma = 1.12\ ^oC$

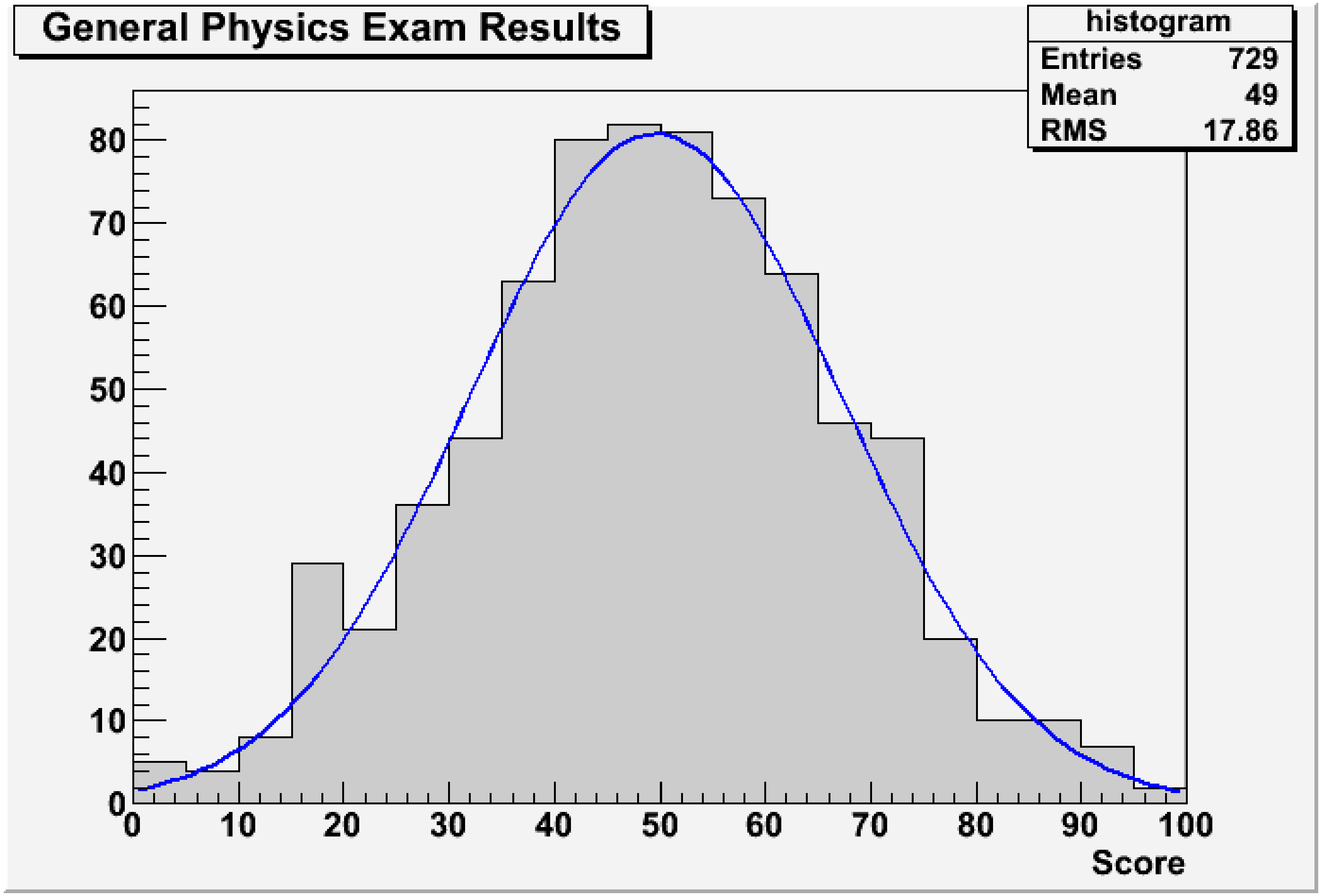


Data: http://data.giss.nasa.gov/tmp/gistemp/STATIONS/tmp_649170620000_14_0/station.txt

“EP106 General Physics II” Course exam results (2010)

Mean score: $\langle x \rangle = 49.0$

Std. Dev. : $\sigma = 17.9$

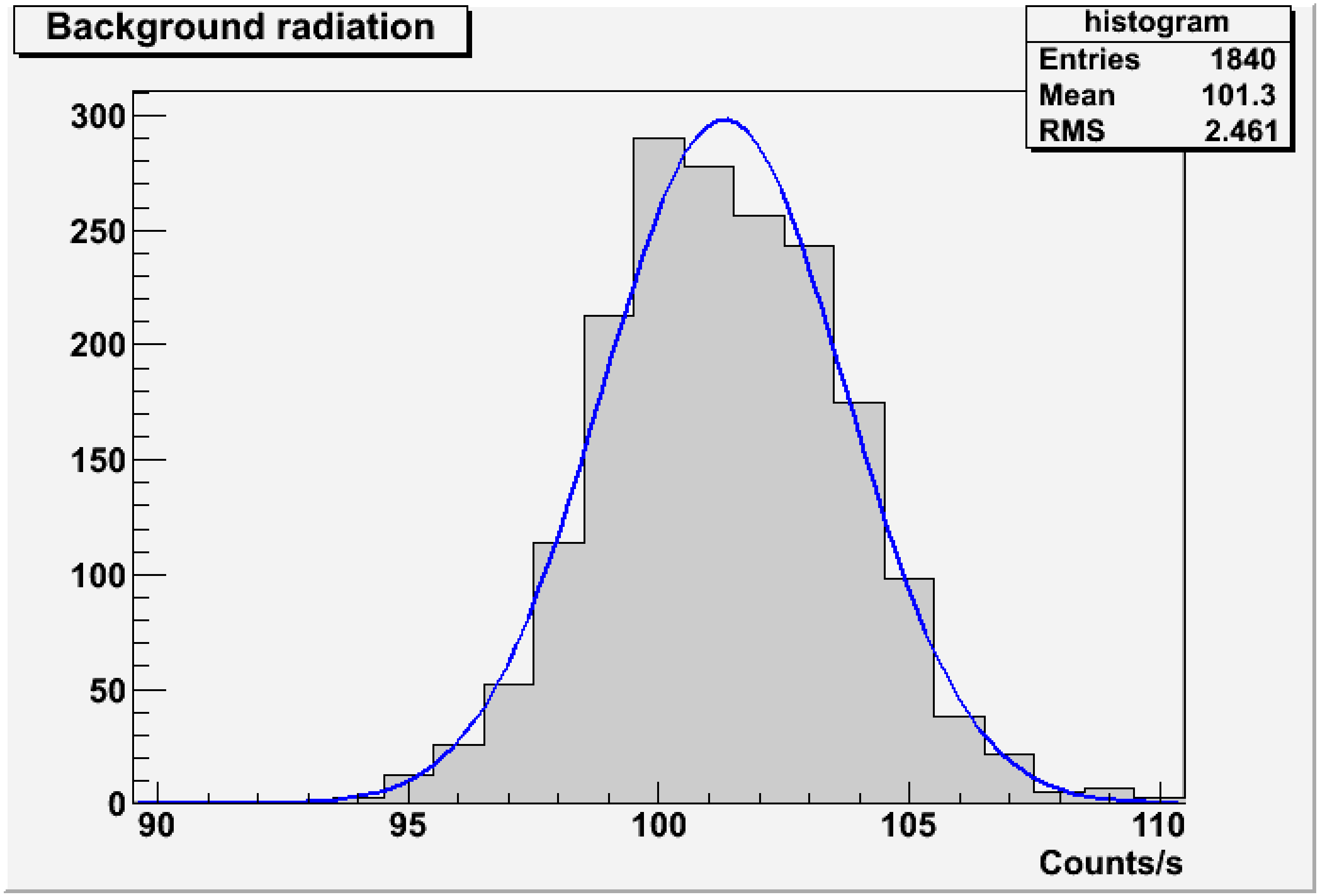


Data: http://www1.gantep.edu.tr/~physics/ep106/exam-statistics.php

Background Radiation in Gaziantep (2013)

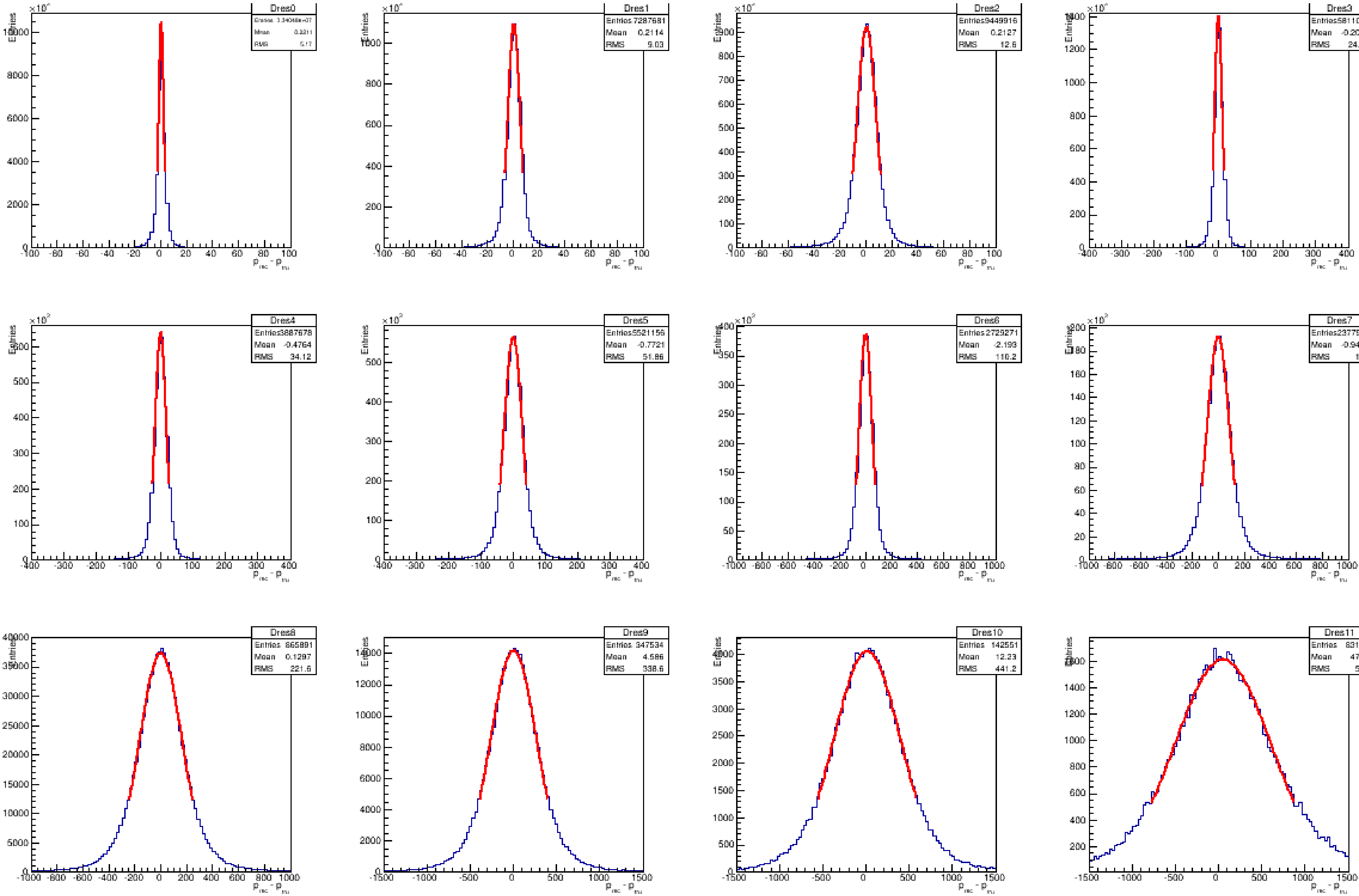
Mean : $\langle x \rangle = 101.3$ counts / sec

Std. Dev. : $\sigma = 2.5$ counts / sec



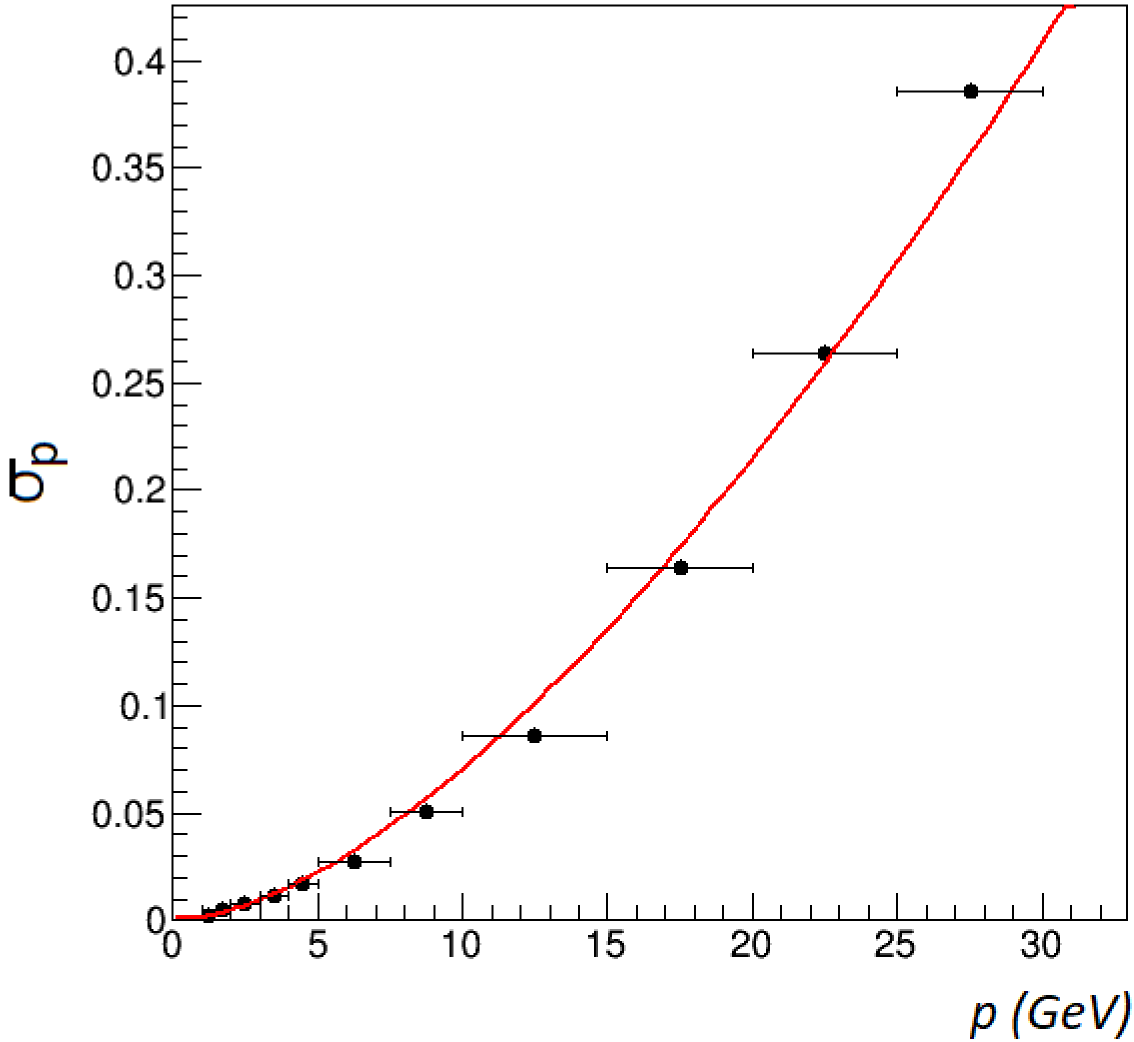
Data is obtained by: Research Assistant Sadık Zuhur (University of Gazaintep)

Tracker Resolution of ALEPH Detector (p – p_{true} distributions)



Tracker Resolution of ALEPH Detector

(σ_p = resolution = width of Gaussian)



Standard Normal Curve

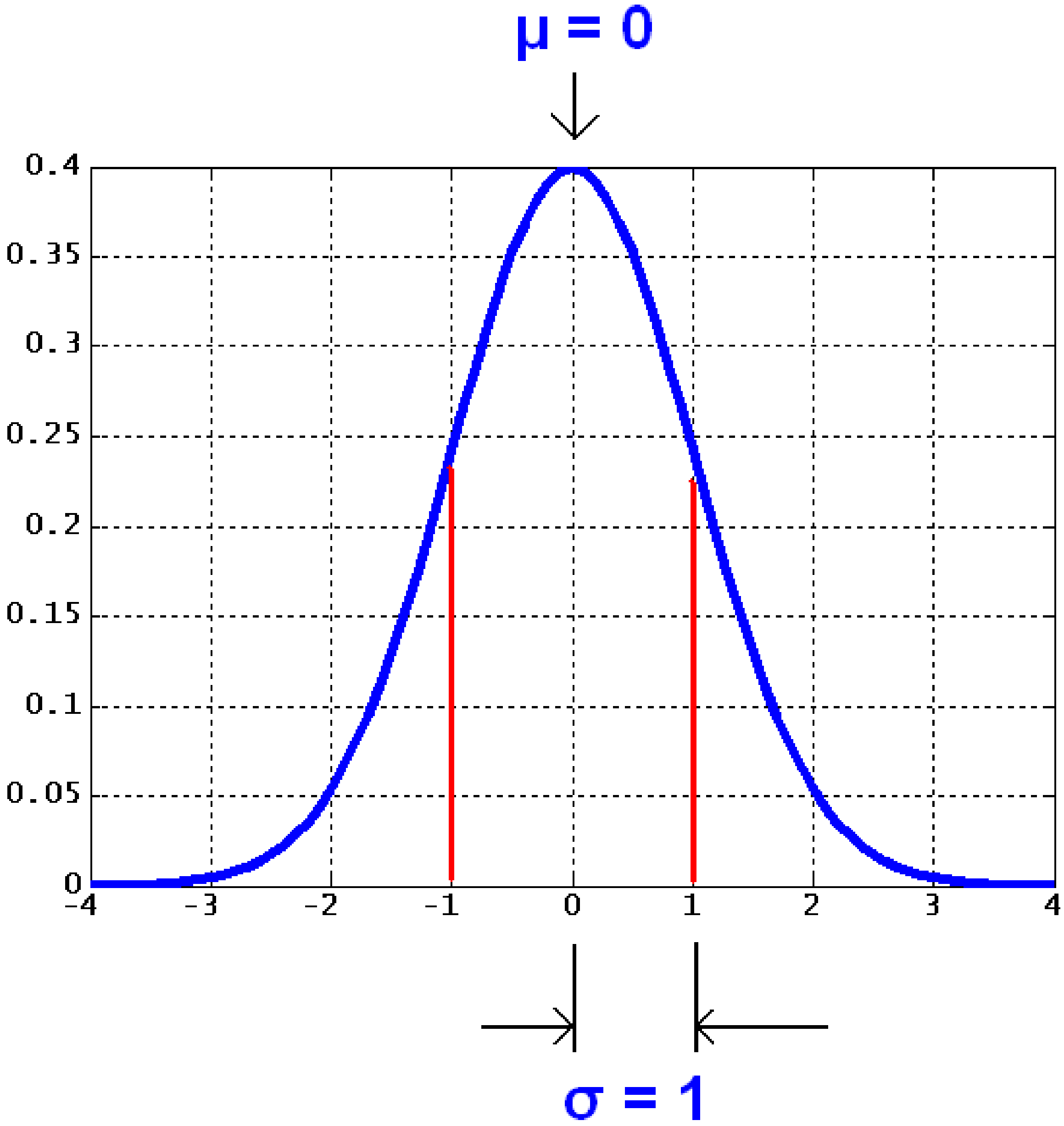
The normal distribution function for

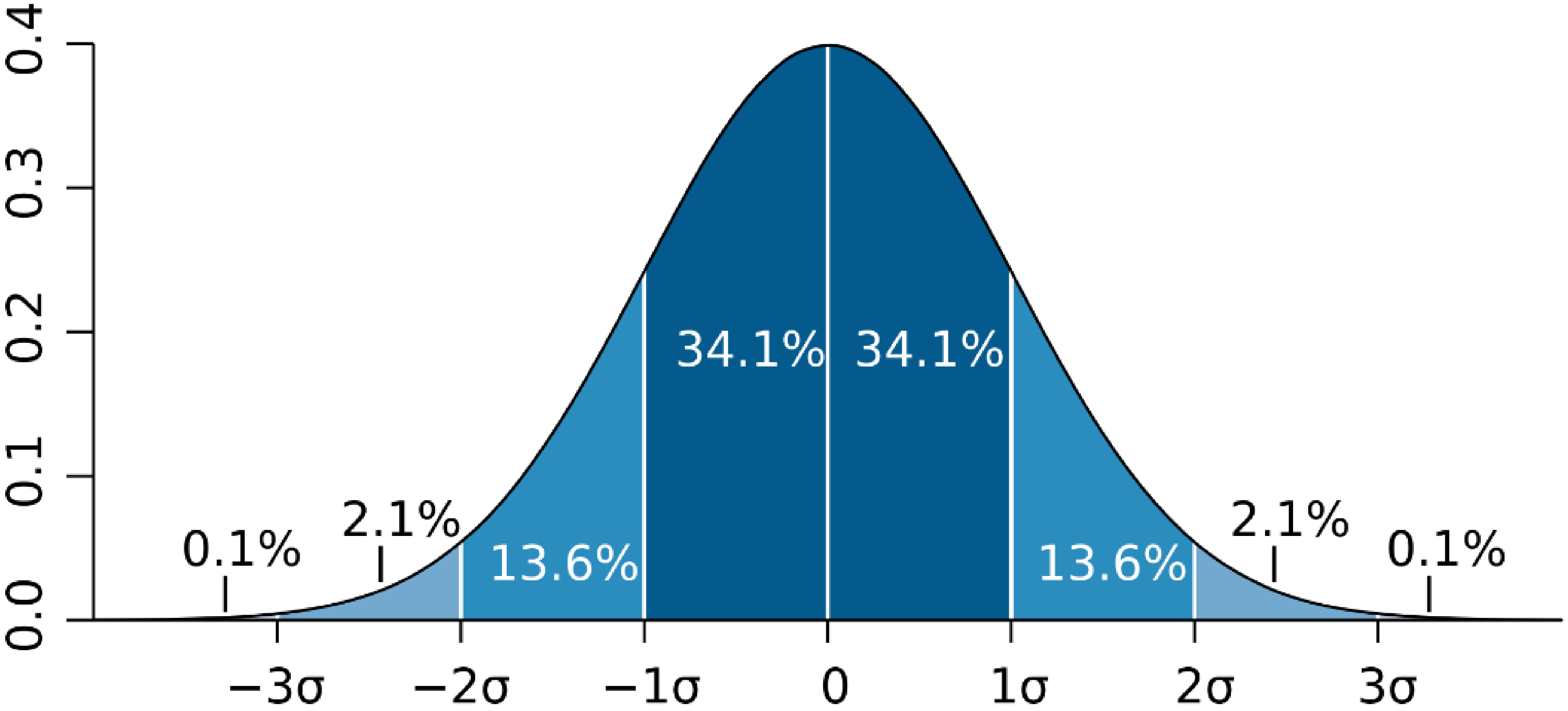
$$\boxed{\mu = 0 \text{ and } \sigma = 1}$$

is called the **standard normal distribution function**.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \approx 0.4 \exp(-x^2/2)$$

Standard Normal Curve



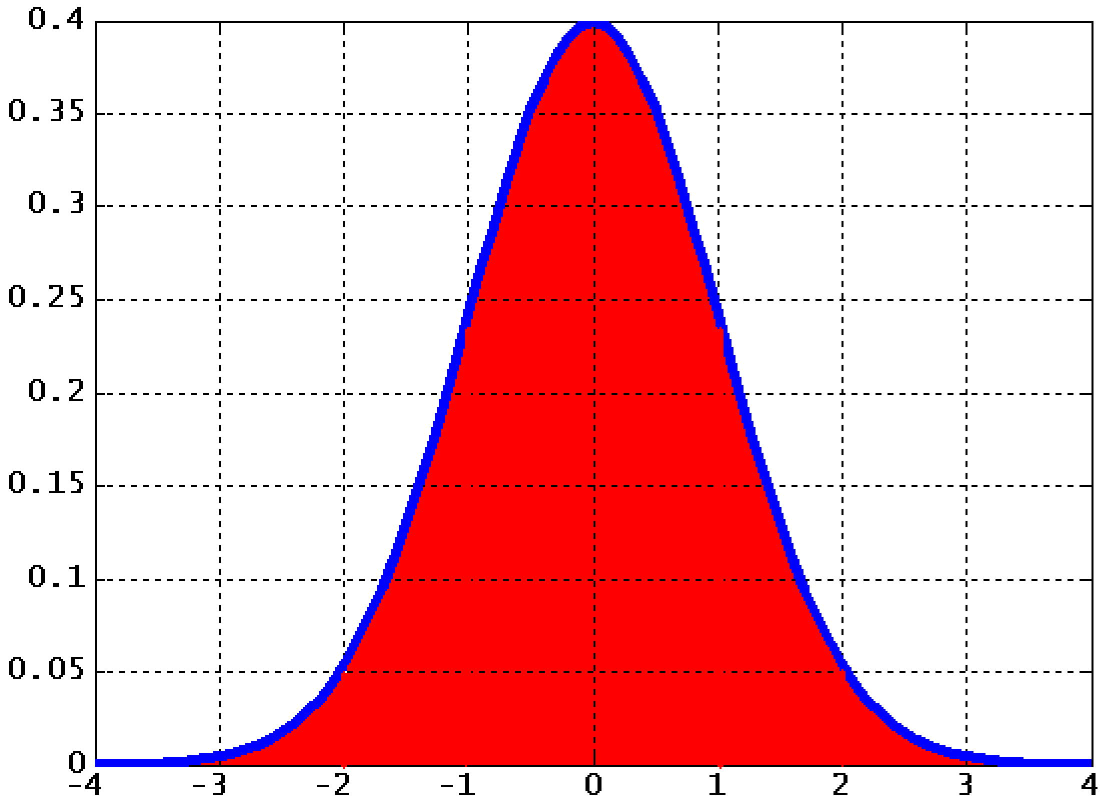

0.4
0.3
0.2
0.1
0.0
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ

Area Under the Curve

Total area under the standard normal curve is 1.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

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

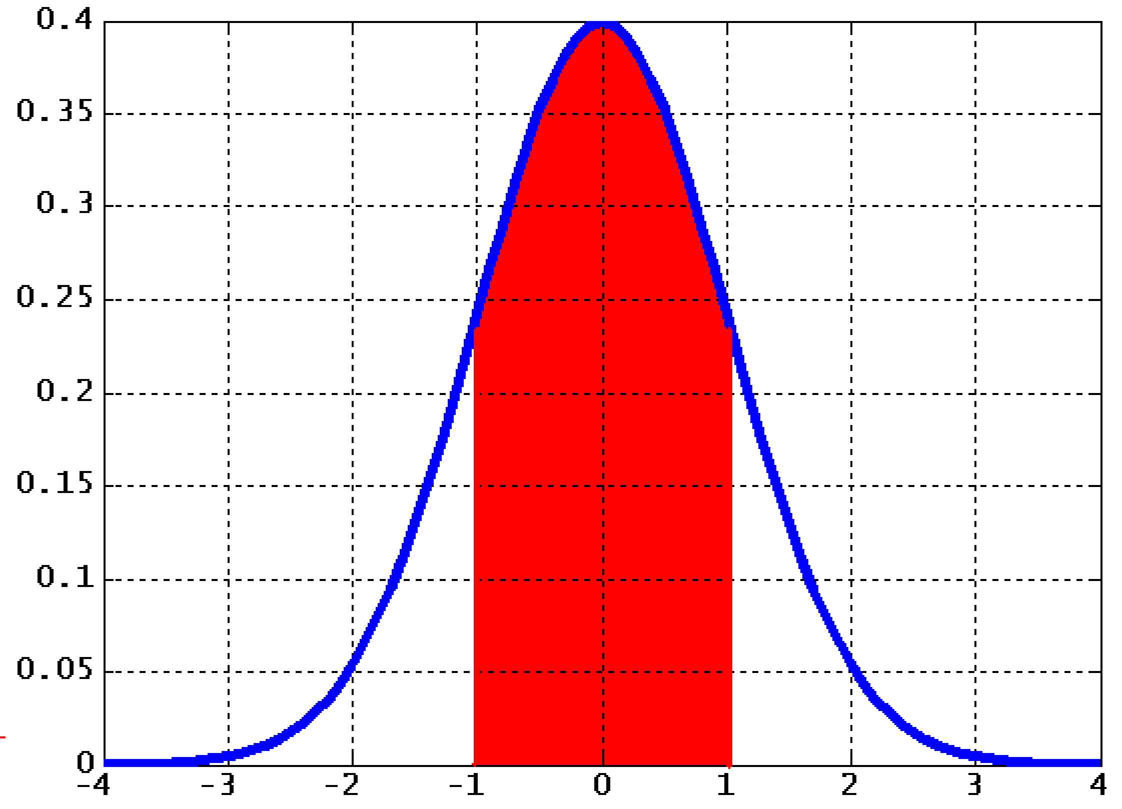


Area under the standard normal curve between [-1, 1] is:

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.6827$$

This corresponds

+- 1 sigma

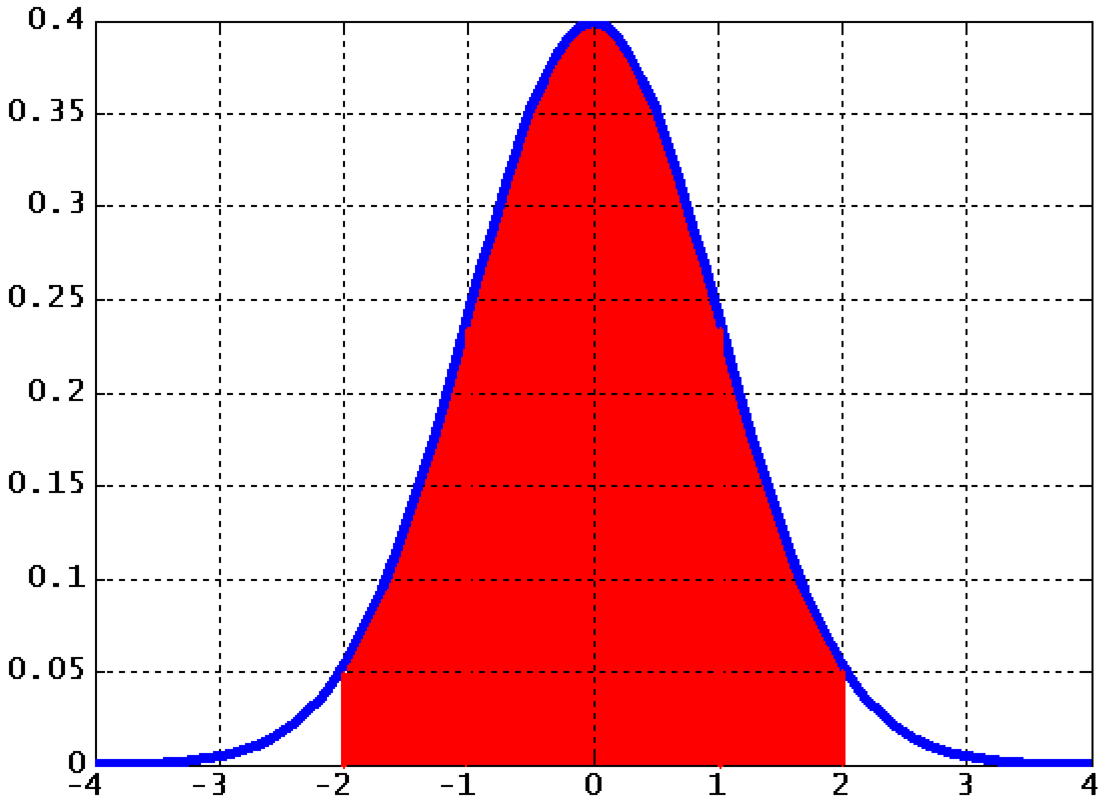


Area under the standard normal curve between [-2, 2] is:

$$\int_{-2}^{2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9545$$

This corresponds

+- 2 sigma

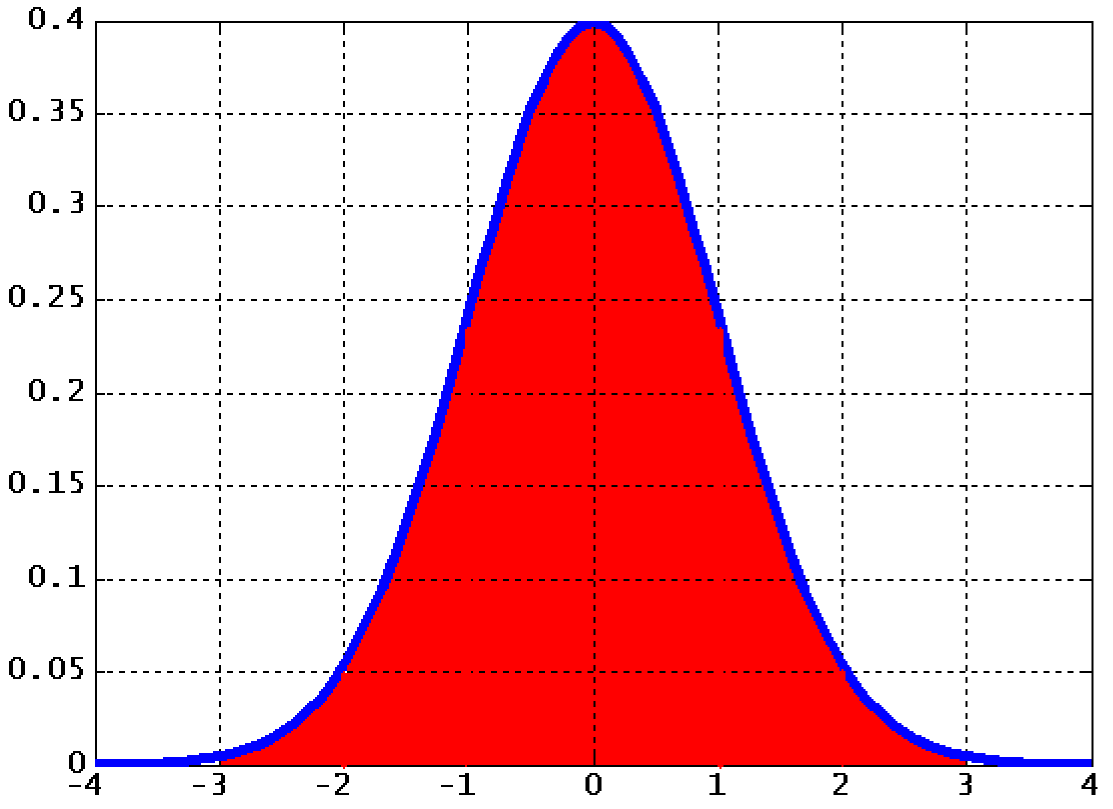


Area under the standard normal curve between [-3, 3] is:

$$\int_{-3}^{3} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.9973$$

This corresponds

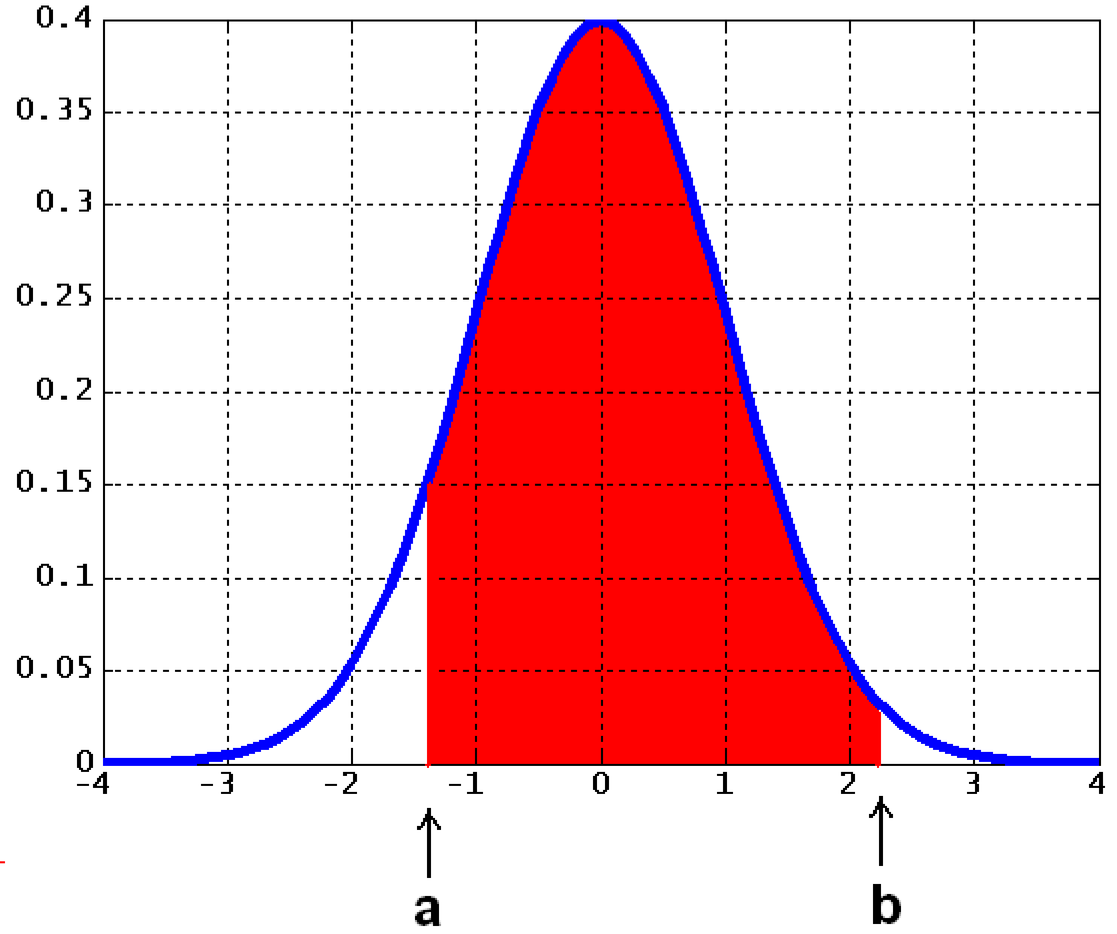
+- 3 sigma

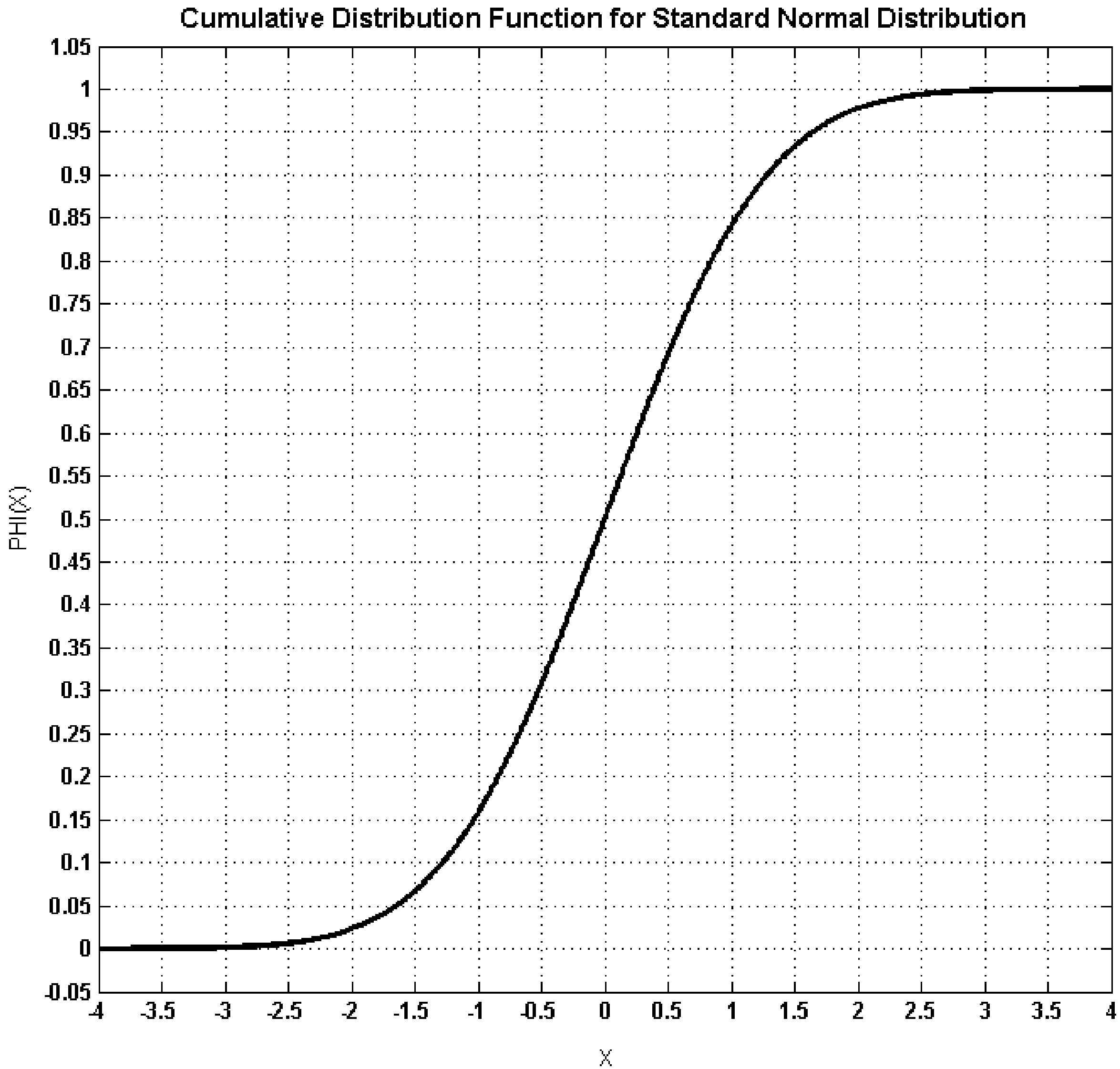


Area under the standard normal curve between [a, b] is:

$$\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \Phi(b) - \Phi(a)$$

The values of the function phi(x) can be taken from a table or from the figure on next page.

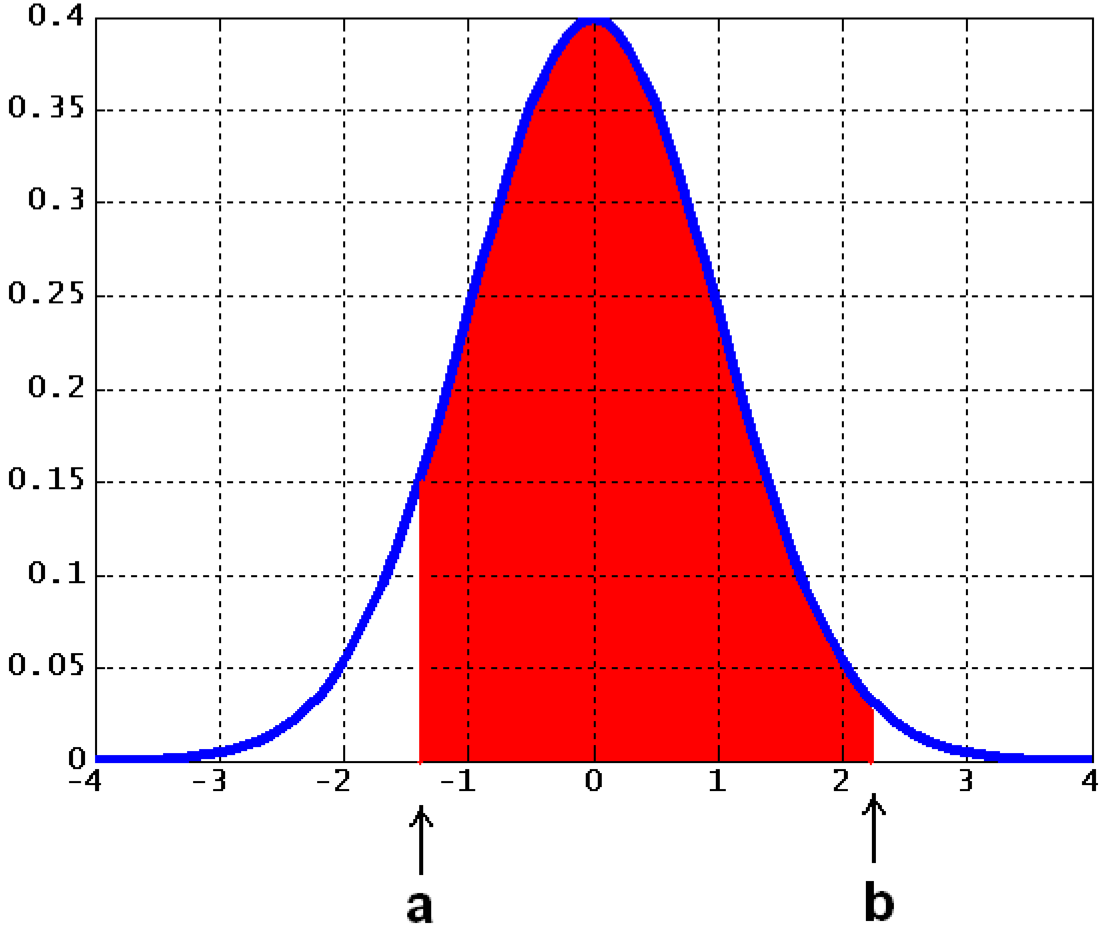




$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}}\, dz = \frac{1}{2}\left[\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right]$$

Example 4

$$\frac{1}{\sqrt{2\pi}}\int_{-1.2}^{2.3} e^{-x^2/2}dx = \Phi(2.3) - \Phi(-1.2) = 0.99 - 0.12 = 0.87$$

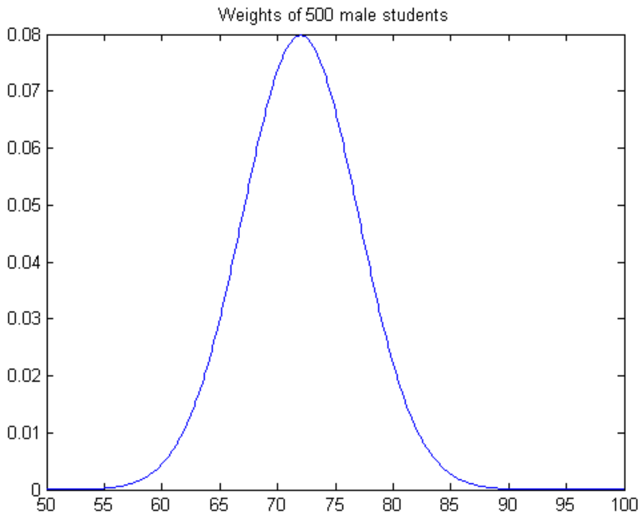


Example 5

Mean weight of 500 male students at a certain university is 72 kg and the standard deviation is 5 kg. Assuming that the weights are normally distributed, find how many students weigh:

(a) between 66 and 75 kg (Answer: 305)

(b) more than 80 kg (Answer: 27)



(a) Convertion to standard normal values

a = (66-72)/5 = -1.2

b = (75-72)/5 = 0.6

$$p = \frac{1}{\sqrt{2\pi}} \int_{-1.2}^{0.6} e^{-x^2/2} dx = \Phi(0.6) - \Phi(-1.2) = 0.6107$$

$$N = (500)(0.6107) = 305$$

Why are Gaussians so useful?

Central limit theorem:

When independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

More specifically:

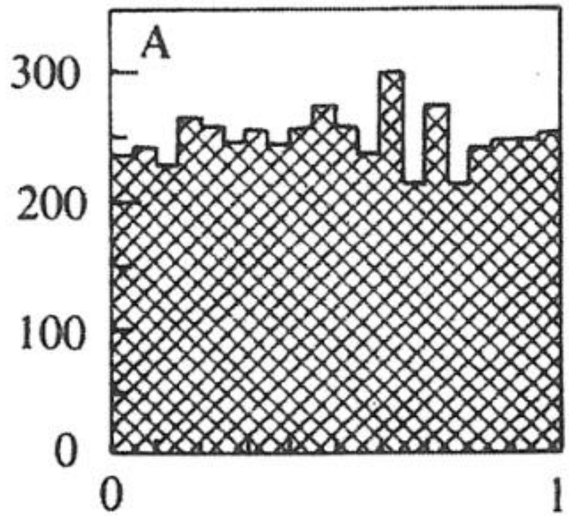
Consider *n* random variables with finite variance σ_i^2 and arbitrary pdfs:

$$y = \sum_{i=1}^{n} x_i \quad \xrightarrow{n \to \infty} \quad y \text{ follows Gaussian with } E[y] = \sum_{i=1}^{n} \mu_i, \; V[y] = \sum_{i=1}^{n} \sigma_i^2$$

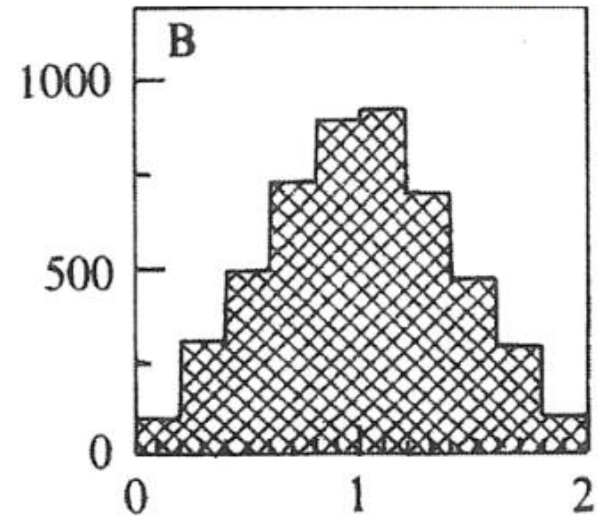
Measurement uncertainties are often the sum of many independent contributions. The underlying pdf for a measurement can therefore be assumed to be a Gaussian.

Sum or difference of two Gaussian random variables is again a Gaussian.

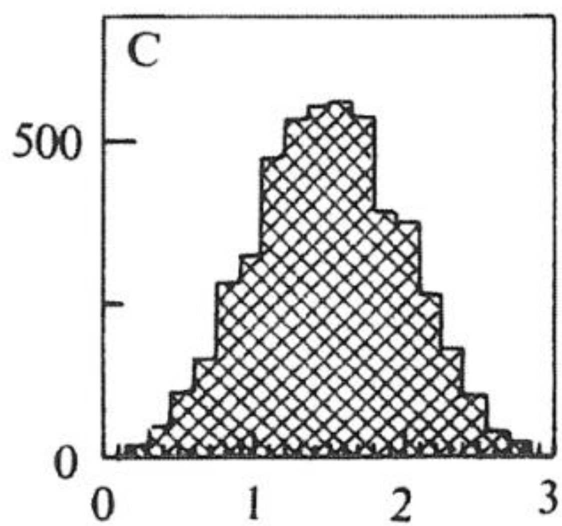
A: x taken from a uniform PD in [0,1], with $\mu=0.5$ and $\sigma^2=1/12$, $N=5000$



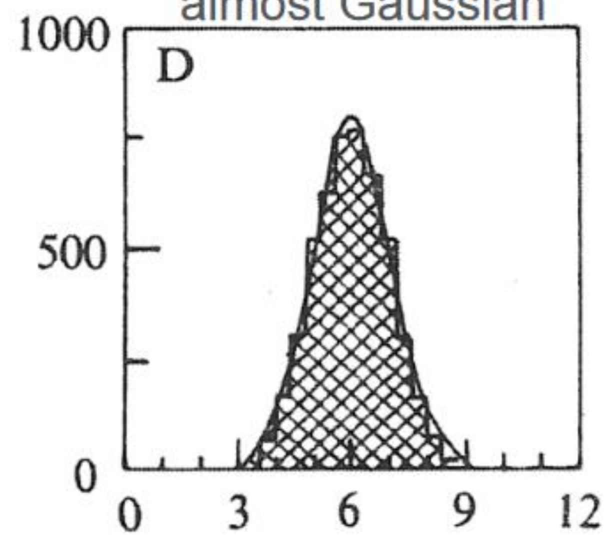
B: $X = x_1+x_2$ from A, $N=5000$, flat shoulders



C: $X = x_1+x_2+x_3$ from A, curved shoulders



D: $X=x_1+x_2+\ldots+x_{12}$ from A, almost Gaussian

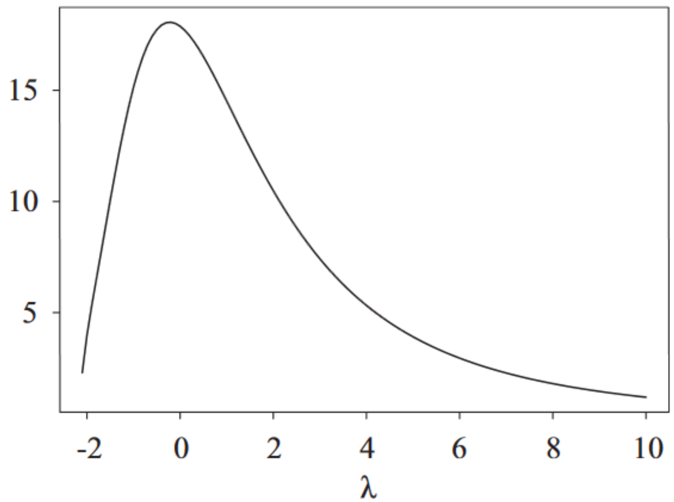


Landau Distribution

L. Landau, J. Phys. USSR 8 (1944) 201

W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. 30 (1980) 253.

$$f(x) = \frac{1}{\pi}\int_0^\infty e^{-u\ln u - xu}\sin(\pi u)\,du$$



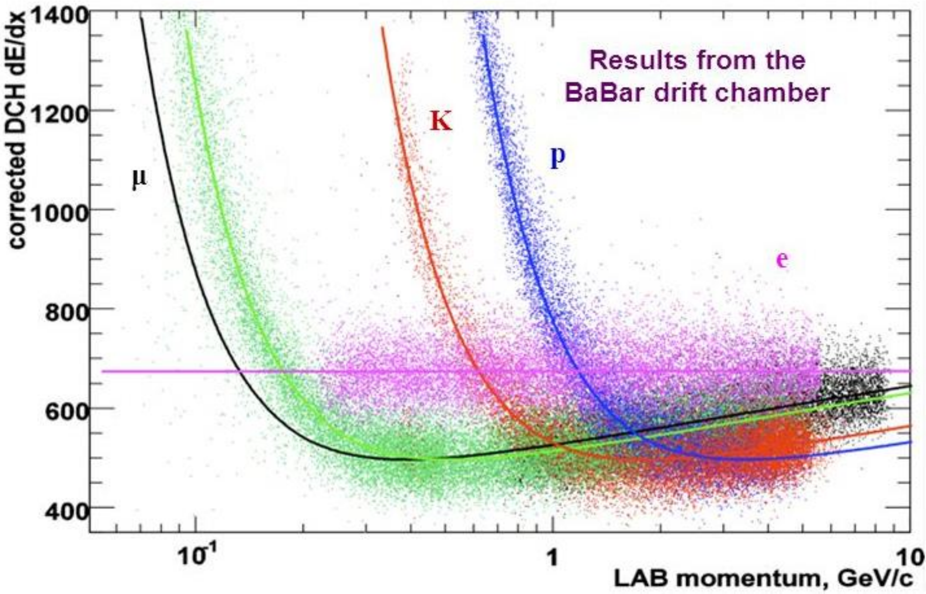
`TRandom::Landau(mu,sigma)`

Root generates random number following a Landau distribution with location parameter mu and scale parameter sigma (x-mu)/sigma. Note that mu is not the mpv and sigma is not the standard deviation of the distribution which is not defined.

For mu =0 and sigma=1, the mpv = -0.22278

Example:

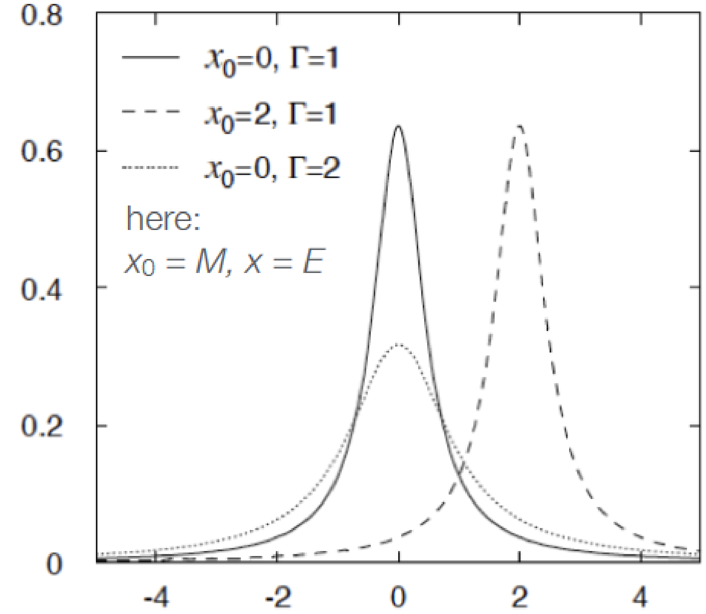
* Describes energy loss of charged particles in a thin layer of material. Tail with large energy loss due to occasional cration of delta rays.



Breit-Wigner Distribution

$$f(E; M, \Gamma) = \frac{1}{2\pi} \frac{\Gamma}{(E - M)^2 + (\Gamma/2)^2}$$

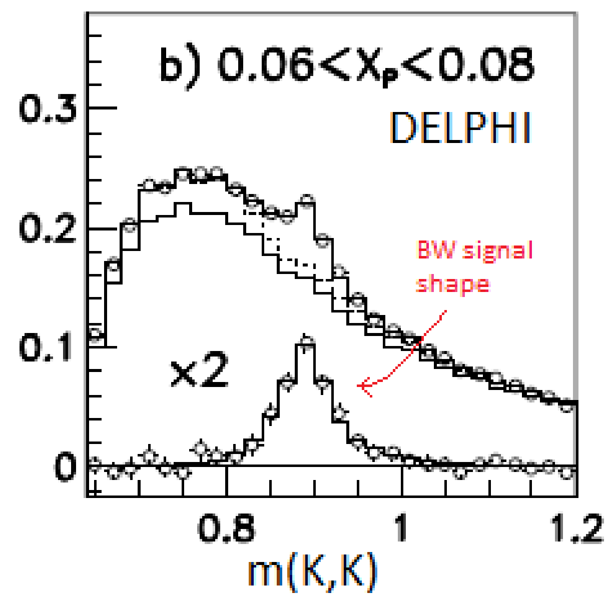
- *Breit-Wigner = Cauchy = Lorentzian*
- *Mean and std is not defined!*



Example: $\phi \rightarrow K^+K^-$ Decay

Production cross section of a resonance with mass M and width Γ (full width at half maximum)

$\phi \rightarrow K^+ K^-$ candidates



Student's t Distribution

$x_1, \ldots, x_n$ are selected from a normal distribution with mean μ & StdDev σ

Sample mean and estimate of the variance:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ → follows standard normal distr. (μ=0, σ=1)

$t := \dfrac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}}$ → follows Student's t distr. with n–1 degrees of freedom

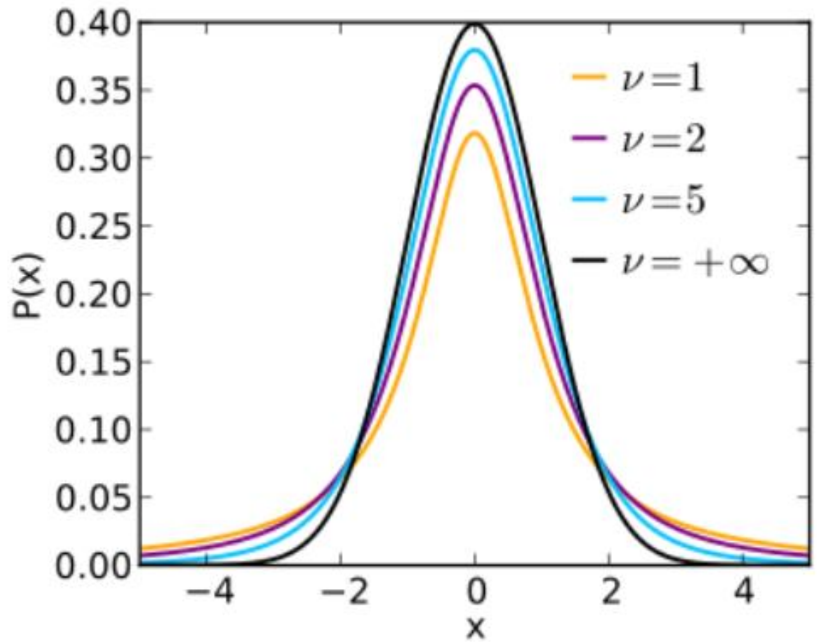
Student's t distribution:

$$f(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

With ν = n – 1 for n measurements; t-distribution can be used to construct a confidence interval for the true mean

$\nu = 1:\ $ Cauchy distr.

$\nu \to \infty:\ $ Gaussian

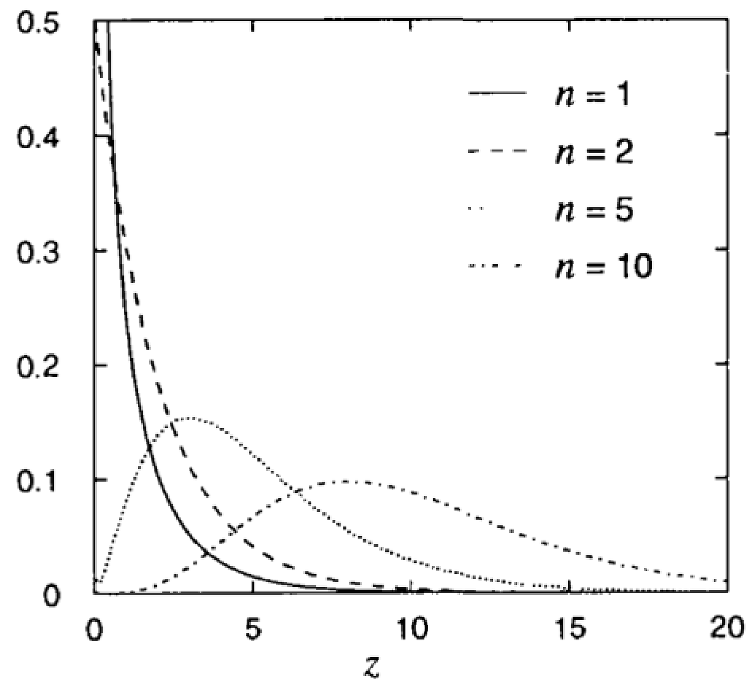


X^2 (chi-square) Distribution

Let $x_1, x_2, \ldots, x_n$ be n independent standard normal ($\mu = 0$, $\sigma = 1$) random variables. Then the sum of their squares

$$z = \sum_{i=1}^{n} x_i^2$$

follows a $\chi 2$ distribution with *n* dof (degrees of freedom).



χ^2 distribution:

$$f(z;n) = \frac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \qquad (z \geq 0)$$

$$\langle z \rangle = \mu_z = E[z] = n$$

$$\sigma^2 = V[z] = 2n$$

Example:

Quantifies goodness of fit:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - h(x_i)}{\sigma_i} \right)^2$$