A Statistician's view of a physics experiment:

$P(\text{data}|\text{hyp})$ is always known.

It is the function (or algorithm) which describes the experiment.
It gives the probability of observing data when the laws of physics are given by the hypothesis hyp. It describes the forward process (hyp→data).

forward process (hyp→data) occurs in real expt. (hyp true but unknown)
forward process (hyp→data) occurs in simulation (hyp known but untrue)

Data are random:	Repeat the experiment, the data will be different.
Hypothesis is NOT random:	The mass of the Higgs is always the same, even if it is unknown.

Let's look at three examples of a $P(\text{data}|\text{hyp})$.

Three examples of a $P(\text{data}|\text{hyp})$.

Ex 1. A counting experiment (proton decay):

$$P(\text{data}|\text{hyp}) = P(n|\mu) = e^{-\mu}\mu^n/n!$$

where n is the observed number of decays (an integer), and
μ is the expected number = decay rate $\times$ number of protons $\times$ time.

Ex 2. Measuring a particle mass from an invariant mass distribution:

$$f(X\,|\,\mu) = N(\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right]$$

where X are the measured mass values, μ is the true value, and σ is the width of the Gaussian resolution. Now the data X are continuous, so f is a probability density function (pdf).

Ex 3. In a big experiment, $P(\text{data}|\text{hyp})$ is given by the Monte Carlo.

The backward process (data→hyp) is called Statistics

There are two ways of inverting the forward reasoning to do statistics:

The Bayesian way, and the Frequentist way.

For both ways, the methods will depend on the kind of hypothesis involved:

1. Measuring a parameter: Point Estimation
2. Finding the error on the above: Interval Estimation
3. Comparing two hypotheses: Hypothesis Testing
4. Testing one hypothesis: Goodness-of-fit Testing (Frequentist only)
5. Making Decisions: Decision Theory (Bayesian only)

The Written Material for these lectures

The slides will be available.

There is also a book: *Statistical Methods in Experimental Physics* (second edition, 2006), by F. James.

This is the second edition of *Statistical Methods in Experimental Physics* (1971), by Eadie, Drijard, James, Roos and Sadoulet.

These lectures are largely based on (the second edition of) the book.

I also recommend another book, a more practical guide:
Data Analysis in High Energy Physics, edited by Olaf Behnke and others.
There are many authors, including Lorenzo Moneta.

Probability

All statistical methods are based on calculations of probability.
We will define three different kinds of probability:

- Mathematical probability is an abstract concept which obeys the Kolmogorov axioms, and is defined by those axioms alone.
  It cannot be measured.

We will need a specific operational definition that allows us to measure probability. There are two such definitions we will use:

- Frequentist probability is defined as the *limiting frequency* of favourable outcomes in a large number of identical experiments.
- Bayesian probability is defined as the *degree of belief* in a favourable outcome of a single experiment.

Mathematical Probability

The Theory of Probability was properly formalized as a branch of mathematics only in 1930 by Kolmogorov.

Let the set $X_i \in \Omega$ be exclusive events:
(one and only one X_i can occur).

Then $P(X_i)$, is a probability if it satisfies the Kolmogorov axioms:

- (a) $P(X_i) \geq 0$ for all i.
- (b) $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$
- (c) $\sum_\Omega P(X_i) = 1$.

Mathematical Probability 2

Some basic properties that follow directly from the axioms:

- If A is certain (or sure), then $P(A) = 1$.
- If A cannot happen, then $P(A) = 0$.

However, surprisingly:

- If $P(A) = 1$, then A is not necessarily certain.
  It may be **almost certain**.

Example: Let R be a real number drawn randomly between zero and one. Then $P(R \neq 1/2) = 1$, but it is only **almost certain**, since R could be $1/2$, with probability zero!

But mathematical probability is an abstract concept. It cannot be measured. We need a probability with an operational definition.
There are two such definitions of probability: Frequentist and Bayesian.

Frequentist Probability

First defined by John Venn in 1866.

The frequentist probability of an event A is defined as the number of times A occurs, divided by the total number of trials N, in the limit of a large number of identical trials:

$$P(A) = lim_{N\to\infty} \frac{N(A)}{N}$$

where A occurs $N(A)$ times in N trials.

Frequentist probability is used in most scientific work, because it is *objective*. It can (in principle) be determined to any desired accuracy and is the same for all observers.

It is the probability of Quantum Mechanics.

Frequentist Probability 2

Objection to the definition of Frequentist probability:
It requires an infinite number of experiments.

Objection overruled: Many scientific concepts are defined as limits.

For example the electric field:

$$\vec{E} = \lim_{q \to 0} \frac{\vec{F}}{q}$$

where $\vec{F}$ is the force due to the field on the charge q.

Since the charge disturbs the field, it has to be infinitesimally small. In this case, the limit is not even possible because charge is quantized, but still the definition is perfectly valid.

Frequentist Probability 3

However, Frequentist Probability has an important limitation:

It can only be applied to repeatable phenomena.

Objection There are no repeatable phenomena, since
no experiment can be repeated exactly,
for example the age of the universe will be greater the second time.
Objection overruled: We will not apply it to phenomena that depend critically on the age of the universe.

Conclusion: Most scientific work involves repeatable phenomena, and frequentist probability is well defined for this work.
But we need in addition a more general kind of probability if we want to apply it to non-repeatable phenomena.

Bayesian Probability

For phenomena that are not repeatable, we need a more general definition.
(for example, the probability that it will rain **tomorrow**,
or that your detector will break down over the weekend)

Bayesian Probability is defined as the degree of belief that A will happen.

It depends not only on the phenomenon itself, but also on the state of knowledge and beliefs of the observer, and it will in general change with time as the observer gains more knowledge.

We cannot verify if the Bayesian probability is "correct" by observing how often something happens.

The operational definition of Bayesian Probability is based on
the coherent bet of de Finetti. (around 1930)
There are problems, such as whether the value of money is linear.

Some Properties of (any) Probability

$P(A \text{ or } B)$ means A or B or both
$P(A \text{ and } B)$ means both A and B

From the Venn diagram, we see that:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Conditional probability: $P(A|B)$ means the probability that A is true, given that B is true.

If A and B are independent, then $P(A|B) = P(A)$.

Example of conditional probability: HB is a human being.
A: HB is pregnant.
B: HB is a woman.

Then: $P(A|B) \approx 1\%$
$P(B|A) = 1.$

Bayes' Theorem

Rev. Thomas Bayes, published posthumously in 1746.

Bayes' Theorem says that the probability of both A and B being true simultaneously can be written:

$$P(A \text{ and } B) = P(A|B)P(B) = P(B|A)P(A)$$

which implies:

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

which can be written:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|\text{not}B)P(\text{not}B)}$$

An Application of Bayes' Theorem

Suppose we have a particle ID detector designed to identify kaons, such that if a kaon hits the detector, the probability that it will produce a positive pulse (T^+) is 0.9 :

$$P(T^+ | \mathrm{K}) = 0.9 \quad [90\% \text{ acceptance}]$$

and 1% if some other particle goes through:

$$P(T^+ | \text{not K}) = 0.01 \quad [1\% \text{ background}]$$

Now a particle gives a positive pulse. What is the probability that it is a K? The answer by Bayes' Theorem:

$$P(\mathrm{K}|T^+) = \frac{P(T^+ | \mathrm{K})P(\mathrm{K})}{P(T^+ | \mathrm{K})P(\mathrm{K}) + P(T^+ | \text{not K})P(\text{not K})}$$

Bayes Prior Probability

So the answer depends on the Prior Probability of the particle being a K, that is, the proportion of K in the beam.
Let us consider two possibilities:
$P(\mathrm{K})$ is 1% , and $P(\mathrm{K}) = 10^{-6}$
We would get the following probabilities:

K in beam	$\mathrm{K} = 1\%$	$\mathrm{K} = 10^{-6}$
$P(\mathrm{K}\|T^+)$	0.48	10^{-4}
$P(\mathrm{K}\|T^-)$	0.001	10^{-7}

We have learned that:
- Prior Probability is very important.
- Bayes' Theorem is useful in non-Bayesian analysis.
- This detector is not very useful if $P(\mathrm{K})$ is small.

Other Fundamental Concepts

The **Hypothesis** is what we want to test, verify, measure, decide.

Examples: H: The standard model is correct.
H: The mass of the proton is m_p (continuous range of hypotheses
H: The tau neutrino is massless.

A **Random Variable** is data which can take on different values, unpredictable except in probability:

$$P(\text{data}|\text{hypothesis})$$

is assumed known, provided any unknowns in the hypothesis are given some assumed values. Example: for a Poisson process

$$P(N|\mu) = \frac{e^{-\mu}\mu^N}{N!}$$

where the possible values of the data N are discrete, and μ is the parameter of interest (the hypothesis).

Other Fundamental Concepts: pdf

When the data are continuous, the probability P becomes a Probability Density Function, or pdf, as in:

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

where μ is the true value of the quantity being measured, x is the measured value, and σ is a parameter, the width of the Gaussian.

In the above example, μ is the parameter of interest (what we are trying to measure), and σ, if it is not known, is a **Nuisance parameter**: an unknown whose value does not interest us, but is unfortunately necessary for the calculation of $P(\text{data}|\text{hypothesis})$.

The Likelihood Function

If in $P(\text{data}|\text{hypothesis})$, we put in the values of the data observed in the experiment, and consider the resulting function as a function of the unknown parameter(s), it becomes

$$P(\text{data}|\text{hypothesis})|_{\text{data obs.}} = \mathcal{L}(\text{hypothesis})$$

$\mathcal{L}$ is called the **Likelihood Function**.

R. A. Fisher, the first person to use it, knew that it was not a probability, so he called it a **likelihood**. It will turn out to have some important properties.

How Likelihoods and pdfs transform

Suppose we wish to transform variables, either the data $X \rightarrow Y(X)$ or the parameters $\theta \rightarrow \tau(\theta)$.

- For a likelihood function, the function values remain invariant, and one simply substitutes the transformed parameter values: $\mathcal{L}(\tau) = \mathcal{L}(\tau(\theta))$.
- However, for a pdf, the invariant is the **integrated probability** between corresponding points, so one must in addition multiply by the Jacobian of the transformation $X \rightarrow Y(X)$:

$$Pdf(X) = J(X, Y)Pdf(Y)$$

where the Jacobian J is just $\partial X/\partial Y$ in one dimension, and is the determinant of the matrix of partial derivatives of X_i with respect to Y_i in many dimensions.

So the peaks and valleys in a likelihood are invariant,
whereas the shape of a pdf can be transformed into anything.

Probability Distributions: Expectation

Let X represent continuous data, and X_i discrete data, and in both cases the entire space of the data is Ω.
Let the X be distributed with pdf $f(X)$ if continuous, and with probability $f(X_i)$ if discrete.
Then we define the expectation of a function $g(X)$ as:

$E(g) = \int_\Omega g(X) f(X) dX$ X continuous

$E(g) = \sum_\Omega g(X_i) f(X_i)$ X discrete

The *expectation* E is a *linear operator*
$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)]\,.$

The expectation of X is called the mean μ: $\mu = \int_\Omega X f(X) dX$

Remember that f is always normalized: $\int_\Omega f(X) dX = \sum_\Omega f(X_i) = 1$

Probability Distributions: Variance

The expectation of the function $(X-\mu)^2$ is called the *variance* $V(X)$ of the density $f(X)$:

$$\begin{aligned} V(X) &= \sigma^2 = E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E(X^2) - \mu^2 \\ &= \int (X-\mu)^2 f(X) dX \,. \end{aligned}$$

The square root of the variance is the standard deviation σ.

The expectation and the variance do not always exist (Cauchy, Landau dist.).

Probability Distributions: Variance of a Sum

From the definition, the Variance of the sum of random variables is

$$V(\alpha X + Y) = \alpha^2 V(X) + V(Y) + 2\alpha \text{ cov}(X, Y)$$

where

$$\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$$

and we define the correlation in terms of the covariance

$$\text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation coefficient ρ satisfies $-1 \leq \rho \leq 1$.

If X and Y are independent, they are also uncorrelated and $\rho = 0$.

Probability Distributions: Variance of a Ratio

Suppose X and Y are independently distributed, with density functions $f(X)$ and $g(Y)$, respectively. Let $E(X) = \mu_X$ and $V(X) = \sigma_X^2$, and similarly for Y. We wish to consider the distribution of the random variable $U = X/Y$.

Approximate variance formula:

$$V\left(\frac{X}{Y}\right) \simeq \left(\frac{\mu_X}{\mu_Y}\right)^2 \left[\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2} - \frac{2\rho_{XY}}{\mu_X}\frac{\sigma_X\sigma_Y}{\mu_Y}\right],$$

The above "rule of thumb" is well known, but may be very wrong.
In particular, if f and g are Gaussian distributions,

$V(X/Y)$ is infinite.

See Cousins and James, Am J. Physics, **74**,159 (Feb 2006)

Discrete Distributions: Binomial

Consider the case where there are two possible outcomes
(which we may call success and failure)
for each of N trials.

The binomial distribution gives the probability of finding
exactly r successes in N trials,
when the probability of success in each single trial is a constant, p.

The distribution of
the number of events in a single bin of a histogram
is binomial
(if the bin contents are independent,
but the total number of events N is fixed).

Discrete Distributions: Binomial

Variable	r, positive integer $\leq N$.
Parameters	N, positive integer.
	p, $0 \leq p \leq 1$.
Probability function	$P(r) = \binom{N}{r} p^r (1-p)^{N-r}$,
	$r = 0, 1, \ldots, N$.
Expected value	$E(r) = Np$.
Variance	$V(r) = Np(1-p)$.
Probability generating function	$G(Z) = [pZ + (1-p)]^N$.

Discrete Distributions: Multinomial

The generalization of the binomial distribution to the case of more than two possible outcomes of an experiment is called the multinomial distribution.

It gives the probability of exactly r_i outcomes of type i in N independent trials, where the probability of outcome i in a single trial is p_i, $i = 1, 2, \ldots, k$.

Note that, as with the binomial distribution, the total number of trials, $N = \sum_{i=1}^{k} r_i$ is fixed.

Discrete Distributions: Multinomial

Variable $r_i, i = 1, 2, \ldots, k$, positive integers $\leq N$.
Parameters N, positive integer.
k, positive integer.
$p_1 \geq 0, p_2 \geq 0, \ldots, p_k \geq 0, \quad \sum_{i=1}^{k} p_i = 1$

Probability function

$$P(r_1, r_2, \ldots, r_k) = \frac{N!}{r_1! r_2! \cdots r_k!} p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k} .$$

Expected values: $E(r_i) = Np_i$.
Variances: $V(r_i) = Np_i(1 - p_i)$.
Covariances: $\mathrm{cov}(r_i, r_j) = -Np_i p_j, \quad i \neq j$.
Probability generating function

$$G(Z_2, \ldots, Z_k) = (p_1 + p_2 Z_2 + \cdots + p_k Z_k)^N .$$

Discrete Distributions: Poisson

The Poisson distribution gives the probability of finding exactly r events in a given length of time, if the events occur independently, at a constant rate.

It is a limiting case of the binomial distribution for $p \to 0$ and $N \to \infty$, when $Np = \mu$, a finite constant.

As $\mu \to \infty$, the Poisson distribution converges to the Normal distribution.

If events occur randomly and independently in time, so that the number of events occuring in a fixed time interval is Poisson-distributed, then the time between two successive events is exponentially distributed.

Discrete Distributions: Poisson

Variable	r, positive integer.
Parameter	μ, positive real number.
Probability function	$P(r) = \frac{\mu^r e^{-\mu}}{r!}$
Expected value	$E(r) = \mu\,.$
Variance	$V(r) = \mu\,.$
Skewness	$\gamma_1 = \frac{1}{\sqrt{\mu}}\,.$
Kurtosis	$\gamma_2 = \frac{1}{\mu}\,.$
Probability generating function	$G(Z) = e^{\mu(Z-1)}$

Continuous Distributions: Normal (Gaussian)

The most important theoretical distribution in statistics is the Normal probability density function, or Gaussian, usually abbreviated $N(\mu, \sigma^2)$. Its cumulative distribution is called the Normal probability integral or error function. One may find in the literature several variations of the definition of the error function.
Note that the half-width of the pdf at half-height is not σ, but 1.176σ.
The probability content of various intervals is given below:

$$P\left(-1.00 \leq \frac{X-\mu}{\sigma} \leq 1.00\right) = 0.683$$

$$P\left(-1.64 \leq \frac{X-\mu}{\sigma} \leq 1.64\right) = 0.90$$

$$P\left(-1.96 \leq \frac{X-\mu}{\sigma} \leq 1.96\right) = 0.95$$

Continuous Distributions: Normal (Gaussian)

Parameters μ, real.

σ, positive real number.

Probability density function

$$f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(X-\mu)^2}{\sigma^2}\right]$$

Cumulative distribution

$$F(X) = \phi\left(\frac{X-\mu}{\sigma}\right) \qquad \text{where} \qquad \phi(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Z} e^{-\frac{1}{2}x^2}\,dx\,.$$

Expected value $E(X) = \mu\,.$

Variance $V(X) = \sigma^2\,.$

Characteristic function

$$\phi(t) = \exp\left[it\mu - \frac{1}{2}t^2\sigma^2\right].$$

Normal Distribution in many variables

Variables $\mathbf{X}$, k-dimensional real vector.

Parameters $\boldsymbol{\mu}$, k-dimensional real vector.

$\underset{\sim}{V}$, $k \times k$ matrix, positive semi-definite.

Probability density function

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{k/2} |\ \underset{\sim}{V}\ |^{\frac{1}{2}}} \exp\left[-\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}}\ \underset{\sim}{V}^{-1}\ (\mathbf{X} - \boldsymbol{\mu}) \right].$$

Expected values $E(\mathbf{X}) = \boldsymbol{\mu}\,.$

Covariances $\mathrm{cov}(\mathbf{X}) = \underset{\sim}{V}$

$$V(X_i) = V_{ii}$$

$$\mathrm{cov}(X_i X_j) = V_{ij}\,, \quad \text{the } (i,j)^{\text{th}} \text{ element of } \underset{\sim}{V}\ .$$

Characteristic function: $\phi(\mathbf{t}) = \exp\left[i\mathbf{t}\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\ \underset{\sim}{V}\ \mathbf{t} \right]$

Continuous Distributions: Chi-square

Suppose that $X_1, \ldots, X_N$ are independent, *standard Normal* variables, $N(0,1)$.
Then the sum of squares

$$X^2_{(N)} = \sum_{i=1}^{N} X_i^2$$

is said to have a *chi-square distribution* $\chi^2(N)$, with N degrees of freedom.

The pdf of the *chi-square distribution* was first derived by Karl Pearson when he published his famous Chi-square Test (1900).

Continuous Distributions: Chi-square

Variable X, positive real number.

Parameter N, positive integer ["degrees of freedom"]

Probability density $$f(X) = \frac{\frac{1}{2}\left(\frac{X}{2}\right)^{(N/2)-1} e^{-X/2}}{\Gamma\left(\frac{N}{2}\right)}\,.$$

Expected value $$E(X) = N\,.$$

Variance $$V(X) = 2N\,.$$

Characteristic function: $\phi(t) = (1-2it)^{-N/2}$

Continuous Distributions: Chi-square

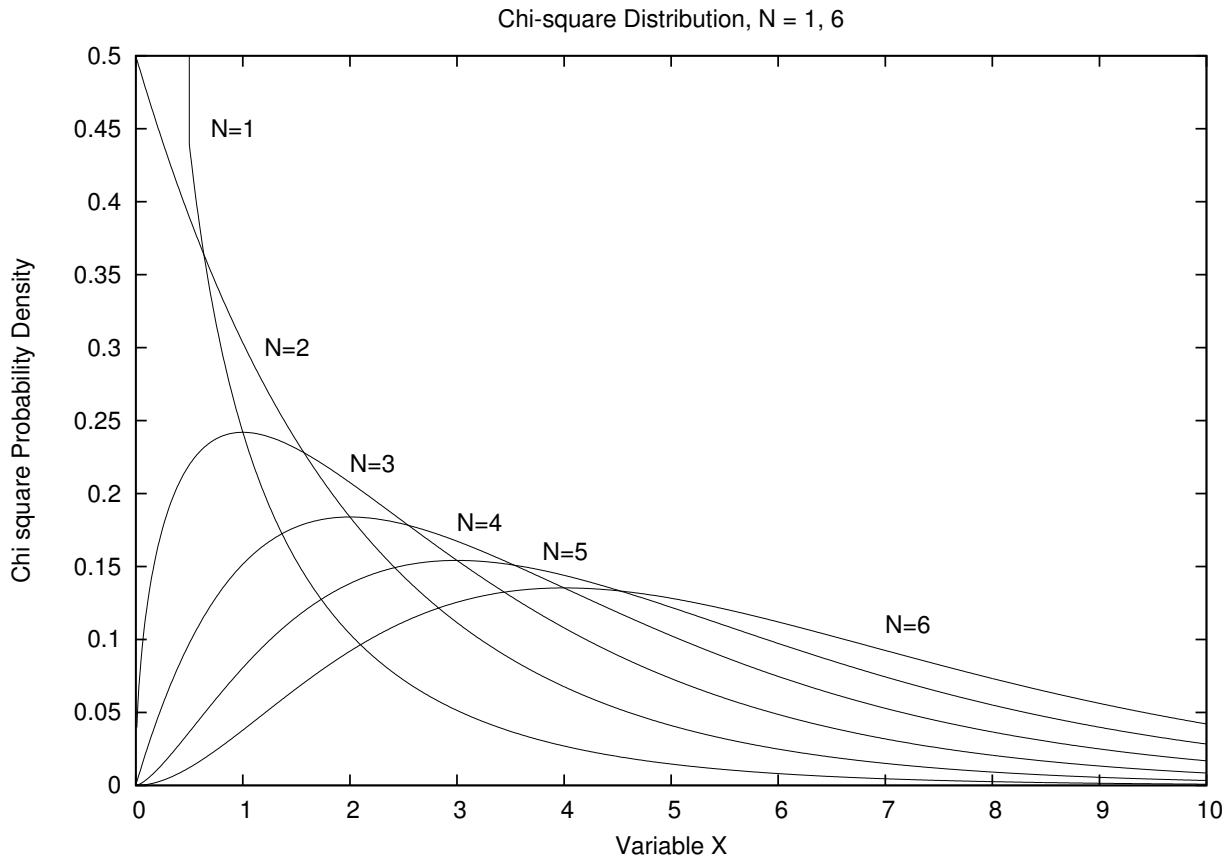
An important relationship exists between the cumulative Poisson distribution and the cumulative χ^2 distribution:

$$P(r \leq N_0 \,|\, \mu) = 1 - P[\chi^2(2N_0 + 2) < 2\mu]\,,$$

$$\text{or} \quad P(r > N_0 \,|\, \mu) = P[\chi^2(2N_0 + 2) < 2\mu]\,,$$

$$\text{or} \quad P(r \geq N_0 \,|\, \mu) = P[\chi^2(2N_0) < 2\mu]\,.$$

Continuous Distributions: Chi-square



Continuous Distributions: Cauchy or Breit-Wigner

Probability density function

$$f(X) = \frac{1}{\pi} \frac{1}{(1 + X^2)} .$$

Expected value: $E(X)$ is undefined.
The Variance, Skewness and Kurtosis are all divergent.
The Characteristic function is: $\phi(t) = e^{-|t|}$.

The physically important Breit–Wigner distribution is a form of Cauchy, usually written as

$$f(X) = \frac{1}{\pi} \left(\frac{\Gamma}{\Gamma^2 + (X - X_0)^2} \right) .$$

The parameters X_0 and Γ represent location and scale parameters respectively, being the mode and half-width at half-height respectively. However it should be noted that the mean and moments of the distribution are still undefined.

Continuous Distributions: Uniform

Probability density function

$$f(X) = \frac{1}{b-a}, \quad a \le X \le b$$

$$f(X) = 0, \qquad \text{otherwise}$$

Expected value $$E(X) = \frac{a+b}{2}.$$

Variance $$V(X) = \frac{(b-a)^2}{12}.$$

Skewness $$\gamma_1 = 0.$$

Kurtosis $$\gamma_2 = -1.2.$$

Characteristic function

$$\phi(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

Continuous Distributions: Landau

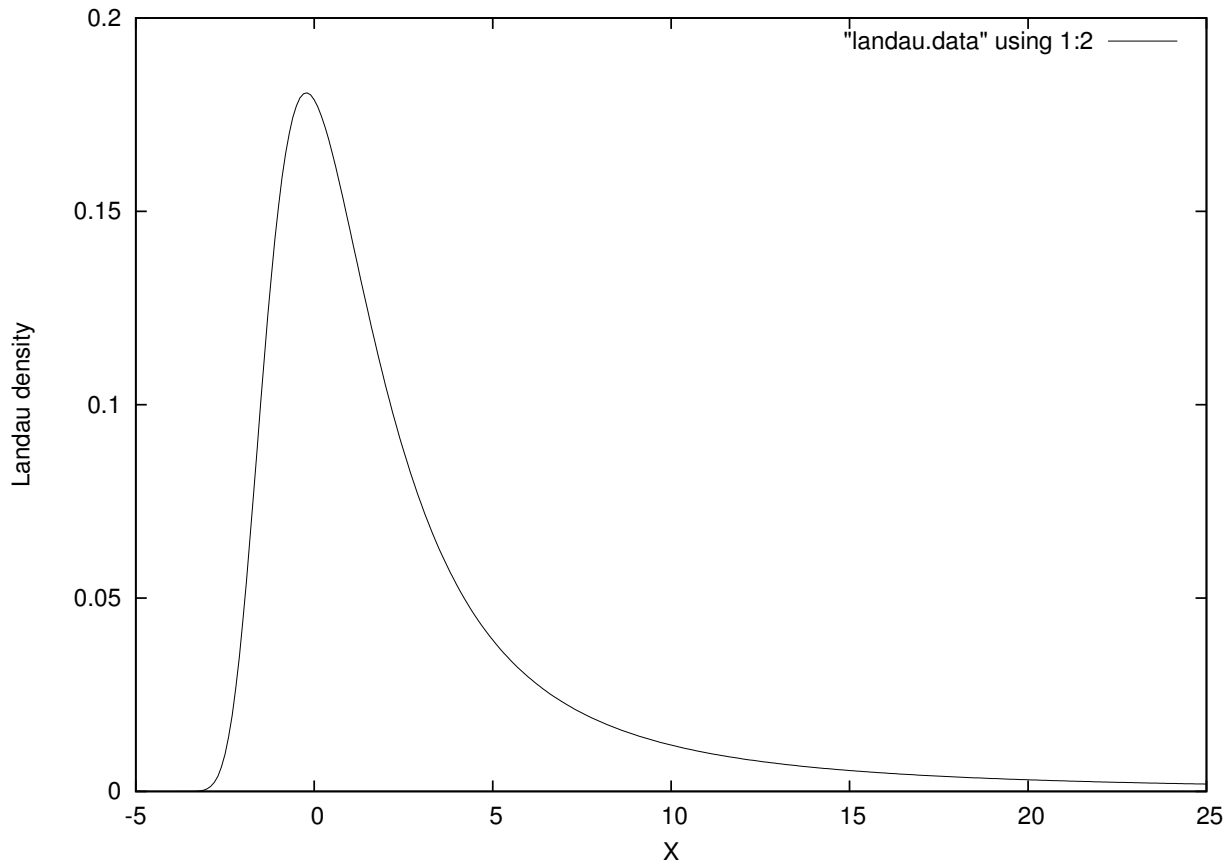
The Landau Density function:

$$\phi(x) = \frac{1}{\pi} \int_0^\infty y^{-y} \sin(\pi y) \exp(-xy) dy$$

where x is a linear function of the energy loss of a charged particle traversing a very thin layer of matter.

This density has an infinite tail going to zero so slowly (like the Cauchy) that the variance of x diverges and its expectation is undefined. Its properties are not easily expressible in closed form, but it is so important that many programs exist to handle it (notably, in ROOT, in the CERN Program Library, and in GSL).

The Landau density



Convergence

We say that the sequence $t_n, (n = 1, \ldots \infty)$ converges to T if:

The Usual Convergence:

You give me an $\epsilon > 0$,

and I'll give you an N
such that,for all $n > N$,
$|t_n - T| < \epsilon$.

Convergence

We say that the sequence $t_n, (n = 1, \ldots \infty)$ converges to T if:

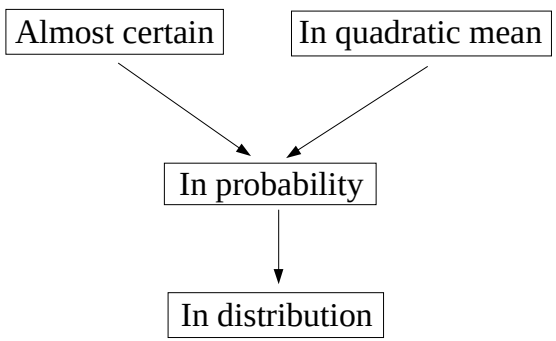
The Usual Convergence:

You give me an $\epsilon > 0$,

and I'll give you an N
such that,for all $n > N$,
$|t_n - T| < \epsilon$.

Convergence in Probability:

You give me an $\epsilon > 0$
and a $p < 1$,
and I'll give you an N
such that, for all $n > N$,
$P(|t_n - T| < \epsilon) > p$.



stronger
to
weaker
convergence

The Law of Large Numbers

Let $\{X_1, \ldots, X_N\}$ be a sequence of *independent* random variables, each having the same mean μ, and variances σ_1^2. Then the

Law of Large Numbers

says that the average converges to μ as $N \to \infty$:

$$\lim_{N\to\infty} \bar{X} \quad = \quad \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} X_i \quad = \quad \mu\,.$$

If the σ_i are all the same, one can show that the above converges in probability. Depending on the behaviour of the σ_i for large i, one can prove stronger laws of large numbers which differ from this weak law by having stronger types of convergence.
These stronger laws will not be of interest to us.

The Central Limit Theorem

Given independent random variables X_i, with mean μ_i and variance σ_i^2,

recall that the distribution of the sum $S = \sum X_i$
will have mean $\sum \mu_i$ and a variance $\sum \sigma_i^2$.

This holds for any distributions provided that the X_i are independent, the individual means and variances exist and do not increase too rapidly with i.

Under the same conditions, the Central Limit theorem states that, as $N \to \infty$, the sum converges (weakly) to a Gaussian $\mathcal{N}$:

$$\frac{S - \sum_{i=1}^{N} \mu_i}{\sqrt{\sum_{i=1}^{N} \sigma_i^2}} \to \mathcal{N}(0,1)\,.$$