Statistical Methods for (Astro)particle Physics and Cosmology

Lecture 1: Introduction, statistical tests

https://www.lip.pt/events/2019/data-science/

School on Data Science in (Astro)particle Physics and Cosmology
Braga, 25-27 March, 2019

Glen Cowan
Physics Department
Royal Holloway, University of London
g.cowan@rhul.ac.uk
www.pp.rhul.ac.uk/~cowan

Outline

→ Lecture 1:

Introduction

Statistical tests, relation to Machine Learning

p-values

Lecture 2:

Parameter estimation

Methods of Maximum Likelihood and Least Squares

Bayesian parameter estimation

Lecture 3:

Interval estimation (limits)

Confidence intervals, asymptotic methods

Experimental sensitivity

Some statistics books, papers, etc.

G. Cowan, *Statistical Data Analysis*, Clarendon, Oxford, 1998

R.J. Barlow, *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences*, Wiley, 1989

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

Luca Lista, *Statistical Methods for Data Analysis in Particle Physics*, Springer, 2017.

L. Lyons, *Statistics for Nuclear and Particle Physics*, CUP, 1986

F. James., *Statistical and Computational Methods in Experimental Physics*, 2nd ed., World Scientific, 2006

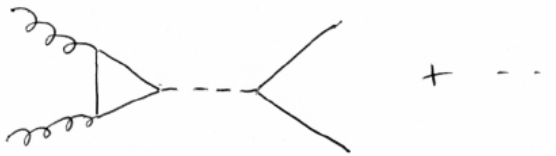
S. Brandt, *Statistical and Computational Methods in Data Analysis*, Springer, New York, 1998 (with program library on CD)

M. Tanabashi et al. (PDG), Phys. Rev. D 98, 030001 (2018); see also **`pdg.lbl.gov`** sections on probability, statistics, Monte Carlo

Theory ↔ Statistics ↔ Experiment

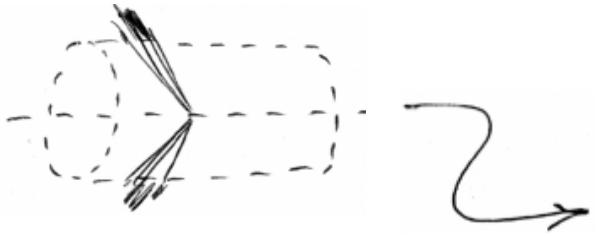
Theory (model, hypothesis):

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + \cdots$$

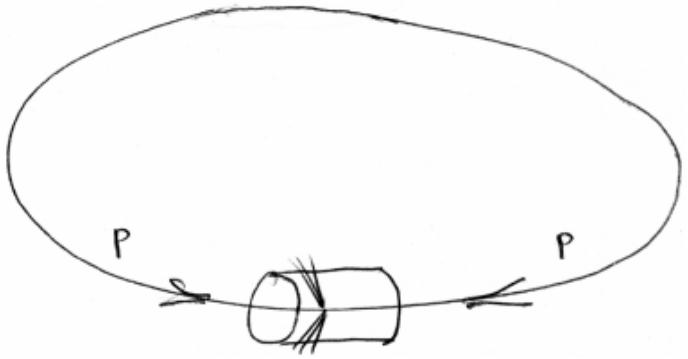


$$\sigma = \frac{G_F \alpha_s^2 m_H^2}{288 \sqrt{2\pi}} \times$$

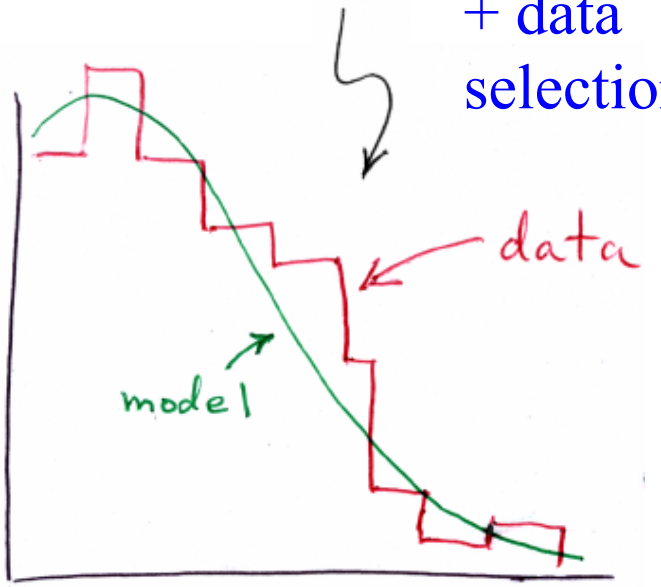
+ simulation of detector and cuts



Experiment:



+ data selection



Data analysis in particle physics

Observe events (e.g., pp collisions) and for each, measure a set of characteristics:

particle momenta, number of muons, energy of jets,...

Compare observed distributions of these characteristics to predictions of theory. From this, we want to:

Estimate the free parameters of the theory: $m_H = 125.4$

Quantify the uncertainty in the estimates: ± 0.4 GeV

Assess how well a given theory stands in agreement with the observed data: 0^+ good, 2^+ bad

To do this we need a clear definition of PROBABILITY

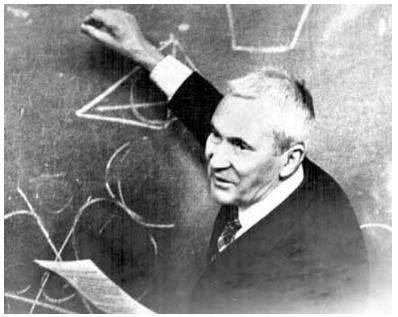
A definition of probability

Consider a set S with subsets A, B, ...

For all $A \subset S, P(A) \geq 0$

$$P(S) = 1$$

If $A \cap B = \emptyset, P(A \cup B) = P(A) + P(B)$



Kolmogorov axioms (1933)

Also define conditional probability of A given B:

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

Subsets A, B independent if: $P(A \cap B) = P(A)P(B)$

If A, B independent, $P(A|B) = \dfrac{P(A)P(B)}{P(B)} = P(A)$

Interpretation of probability

I. Relative frequency

A, B, ... are outcomes of a repeatable experiment

$$P(A) = \lim_{n\to\infty} \frac{\text{times outcome is } A}{n}$$

cf. quantum mechanics, particle scattering, radioactive decay...

II. Subjective probability

A, B, ... are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- Both interpretations consistent with Kolmogorov axioms.
- In particle physics frequency interpretation often most useful, but subjective probability can provide more natural treatment of non-repeatable phenomena:

systematic uncertainties, probability that Higgs boson exists,...

Bayes' theorem

From the definition of conditional probability we have,

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

but $P(A \cap B) = P(B \cap A)$, so

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

Bayes' theorem

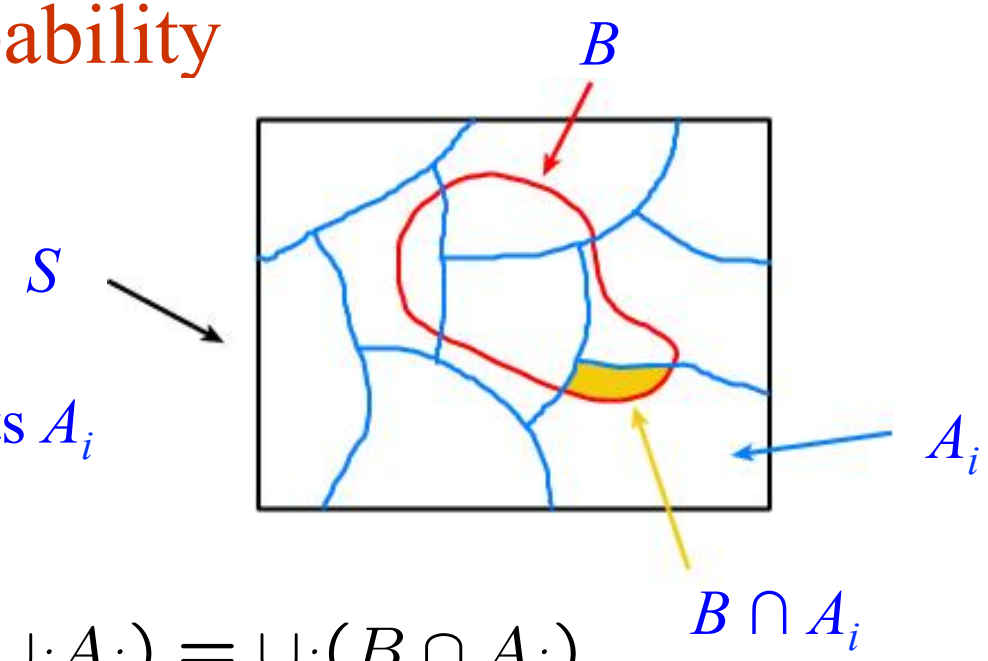
First published (posthumously) by the Reverend Thomas Bayes (1702−1761)

An essay towards solving a problem in the doctrine of chances, Philos. Trans. R. Soc. **53** (1763) 370; reprinted in Biometrika, **45** (1958) 293.

The law of total probability

Consider a subset B of the sample space S,

divided into disjoint subsets A_i such that $\cup_i A_i = S$,



$$\rightarrow \; B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i),$$

$$\rightarrow \; P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$\rightarrow \; P(B) = \sum_i P(B|A_i) P(A_i)$$ law of total probability

Bayes' theorem becomes

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

An example using Bayes' theorem

Suppose the probability (for anyone) to have a disease D is:

$$P(\mathrm{D}) = 0.001$$
$$P(\text{no D}) = 0.999$$

← prior probabilities, i.e., before any test carried out

Consider a test for the disease: result is + or −

$$P(+|\mathrm{D}) = 0.98$$
$$P(-|\mathrm{D}) = 0.02$$

← probabilities to (in)correctly identify a person with the disease

$$P(+|\text{no D}) = 0.03$$
$$P(-|\text{no D}) = 0.97$$

← probabilities to (in)correctly identify a healthy person

Suppose your result is +. How worried should you be?

Bayes' theorem example (cont.)

The probability to have the disease given a + result is

$$p(\mathrm{D}|+) = \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\text{no D})P(\text{no D})}$$

$$= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999}$$

$$= 0.032 \quad \leftarrow \text{posterior probability}$$

i.e. you're probably OK!

Your viewpoint: my degree of belief that I have the disease is 3.2%.

Your doctor's viewpoint: 3.2% of people like this have the disease.

Frequentist Statistics − general philosophy

In frequentist statistics, probabilities are associated only with the data, i.e., outcomes of repeatable observations (shorthand: $\vec{x}$).

Probability = limiting frequency

Probabilities such as

P (Higgs boson exists),
P ($0.117 < \alpha_s < 0.121$),

etc. are either 0 or 1, but we don't know which.

The tools of frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

A hypothesis is is preferred if the data are found in a region of high predicted probability (i.e., where an alternative hypothesis predicts lower probability).

Bayesian Statistics – general philosophy

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No general prescription for priors (subjective!)

Random variables and probability density functions

A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

Suppose outcome of experiment is continuous value x

$$P(x \text{ found in } [x, x + dx]) = f(x)\, dx$$

→ $f(x)$ = probability density function (pdf)

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$ x must be somewhere

Or for discrete outcome x_i with e.g. i = 1, 2, ... we have

$$P(x_i) = p_i$$ probability mass function

$$\sum_i P(x_i) = 1$$ x must take on one of its possible values

Other types of probability densities

Outcome of experiment characterized by several values, e.g. an n-component vector, $(x_1, \ldots x_n)$

→ joint pdf $f(x_1, \ldots, x_n)$

Sometimes we want only pdf of some (or one) of the components

→ marginal pdf $f_1(x_1) = \int \cdots \int f(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$

x_1, x_2 independent if $f(x_1, x_2) = f_1(x_1) f_2(x_2)$

Sometimes we want to consider some components as constant

→ conditional pdf $g(x_1|x_2) = \dfrac{f(x_1, x_2)}{f_2(x_2)}$

Expectation values

Consider continuous r.v. x with pdf $f(x)$.

Define expectation (mean) value as $E[x] = \int x\, f(x)\, dx$

Notation (often): $E[x] = \mu$ ~ "centre of gravity" of pdf.

For a function $y(x)$ with pdf $g(y)$,

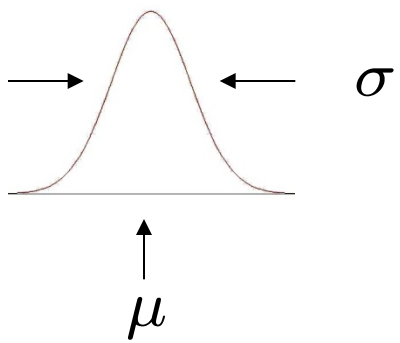
$$E[y] = \int y\, g(y)\, dy = \int y(x) f(x)\, dx \qquad \text{(equivalent)}$$

Variance: $V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2]$

Notation: $V[x] = \sigma^2$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

σ ~ width of pdf, same units as x.



Covariance and correlation

Define covariance cov[x,y] (also use matrix notation V_{xy}) as

$$\mathrm{cov}[x,y] = E[xy] - \mu_x\mu_y = E[(x-\mu_x)(y-\mu_y)]$$

Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\mathrm{cov}[x,y]}{\sigma_x\sigma_y}$$

If x, y, independent, i.e., $f(x,y) = f_x(x)f_y(y)$, then

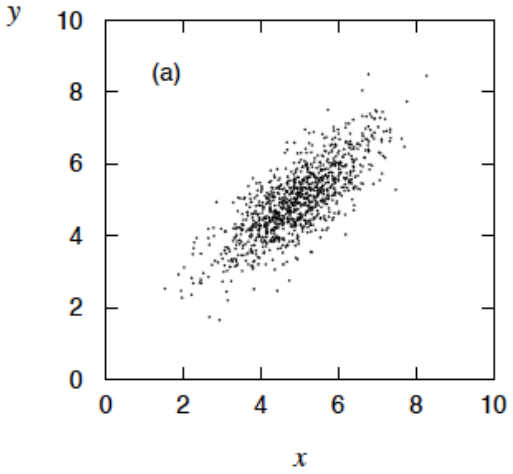
$$E[xy] = \int\int xy\, f(x,y)\, dxdy = \mu_x\mu_y$$

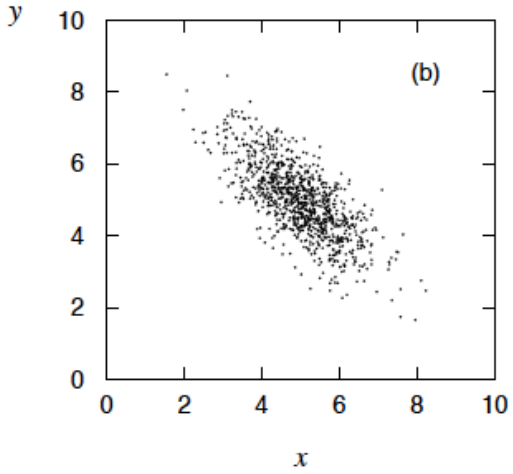
$\rightarrow$ $\mathrm{cov}[x,y] = 0$ x and y, 'uncorrelated'

N.B. converse not always true.

Correlation (cont.)

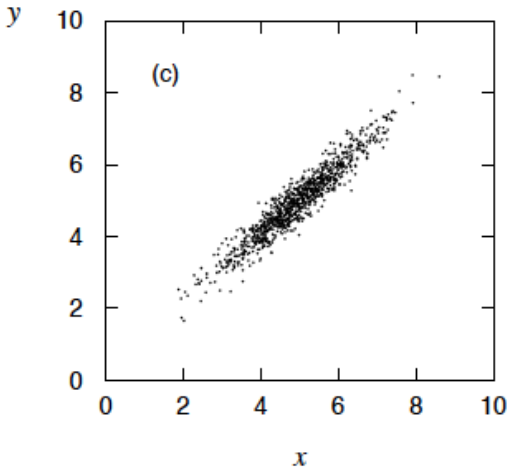
$\rho = 0.75$

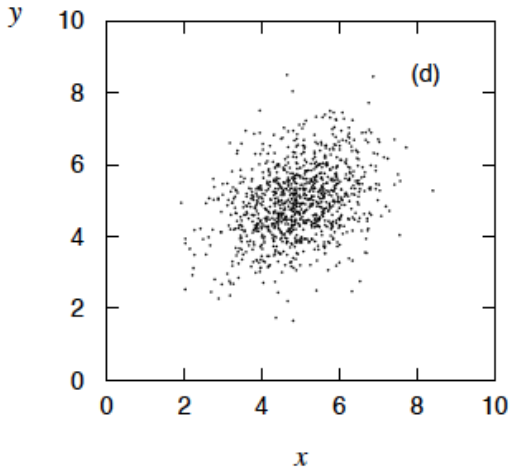




$\rho = -0.75$

$\rho = 0.95$





$\rho = 0.25$

Hypotheses

A hypothesis H specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x.

x could be uni-/multivariate, continuous or discrete.

E.g. write $x \sim P(x|H)$.

x could represent e.g. observation of a single particle,

a single event, or an entire "experiment".

Possible values of x form the sample space S (or "data space").

Simple (or "point") hypothesis: $P(x|H)$ completely specified.

Composite hypothesis: H contains unspecified parameter(s).

The probability for x given H is also called the likelihood of the hypothesis, $L(H) = P(x|H)$.

Often label hypothesis by continuous parameter(s) θ,

$\rightarrow$ likelihood function $L(\theta)$.

Frequentist hypothesis tests

Consider a hypothesis H_0 and alternative H_1.

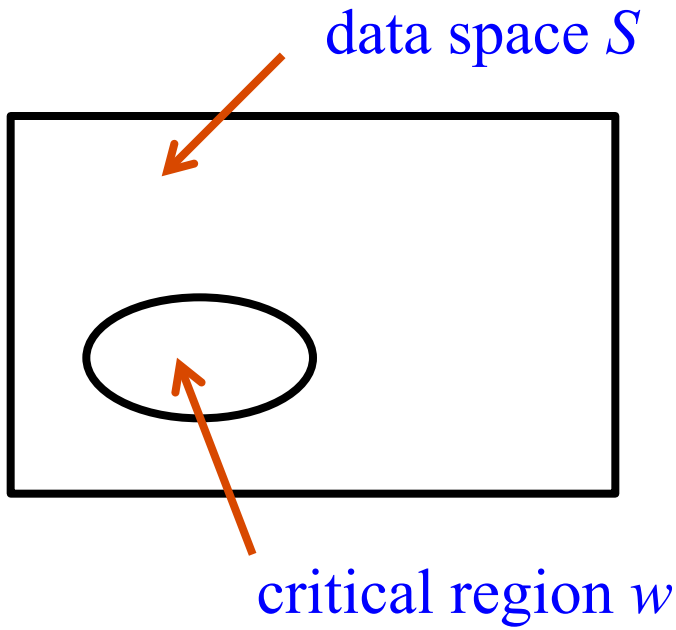
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \,|\, H_0) \leq \alpha$$

Need inequality if data are discrete.

α is called the size or significance level of the test.

If x is observed in the critical region, reject H_0.

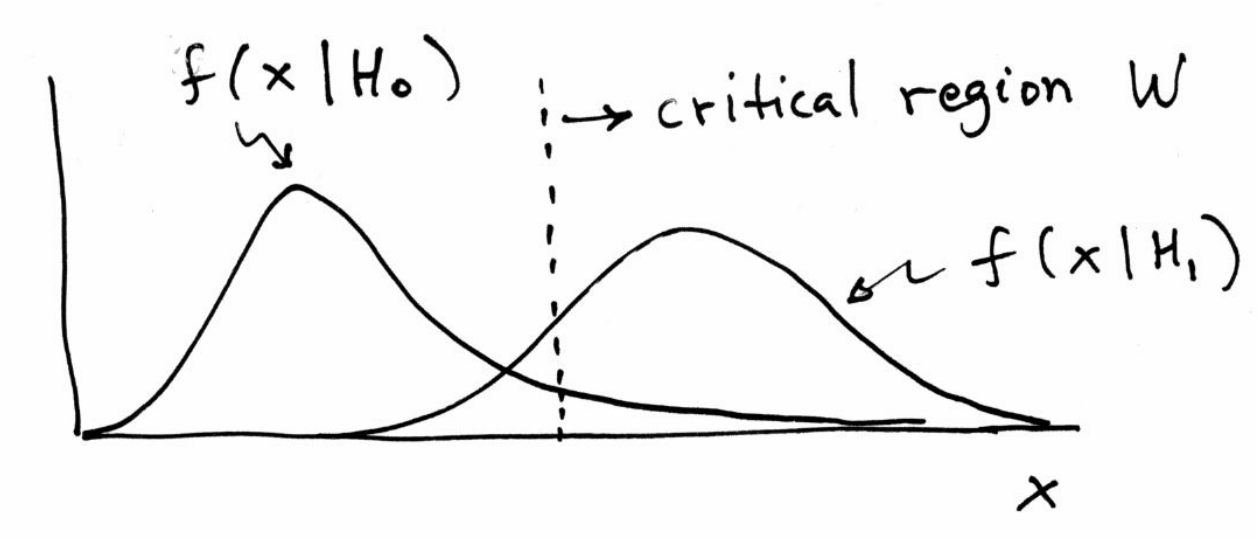


Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same significance level α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1.

Roughly speaking, place the critical region where there is a low probability to be found if H_0 is true, but high if H_1 is true:



Type-I, Type-II errors

Rejecting the hypothesis H_0 when it is true is a Type-I error.

The maximum probability for this is the size of the test:

$$P(x \in W \,|\, H_0) \leq \alpha$$

But we might also accept H_0 when it is false, and an alternative H_1 is true.

This is called a Type-II error, and occurs with probability

$$P(x \in S - W \,|\, H_1) = \beta$$

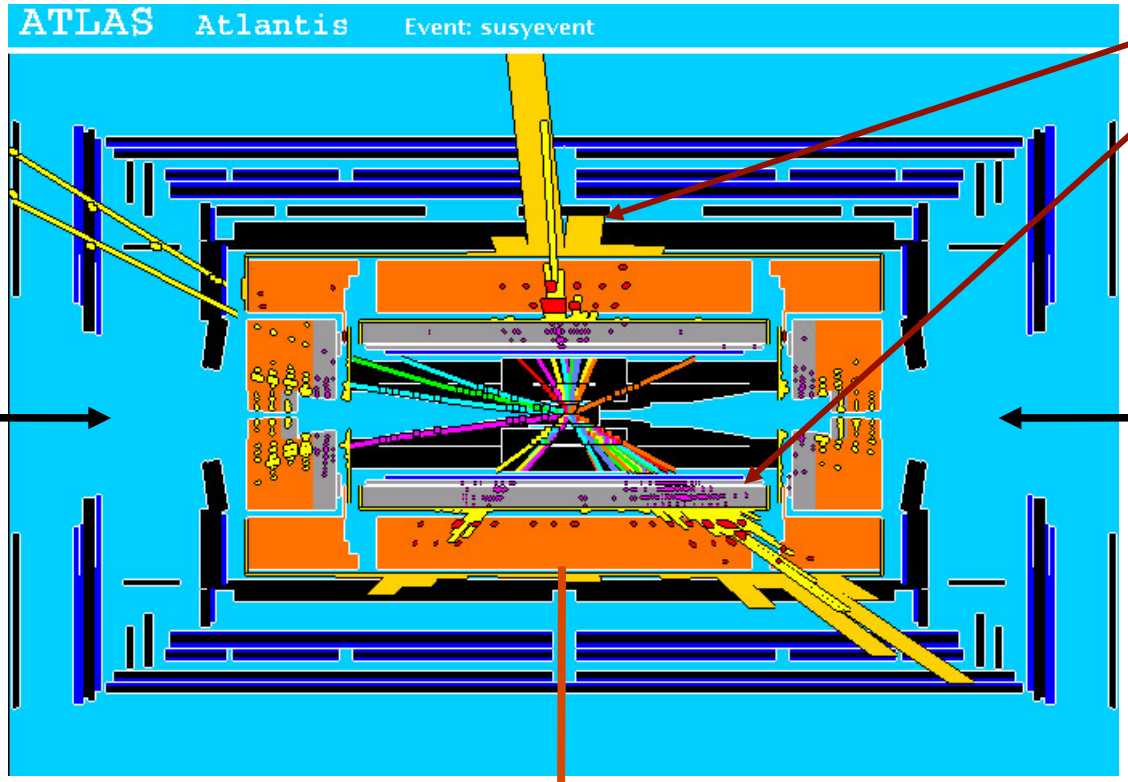
One minus this is called the power of the test with respect to the alternative H_1:

$$\text{Power} = 1 - \beta$$

A simulated SUSY event

high p_T muons

high p_T jets of hadrons

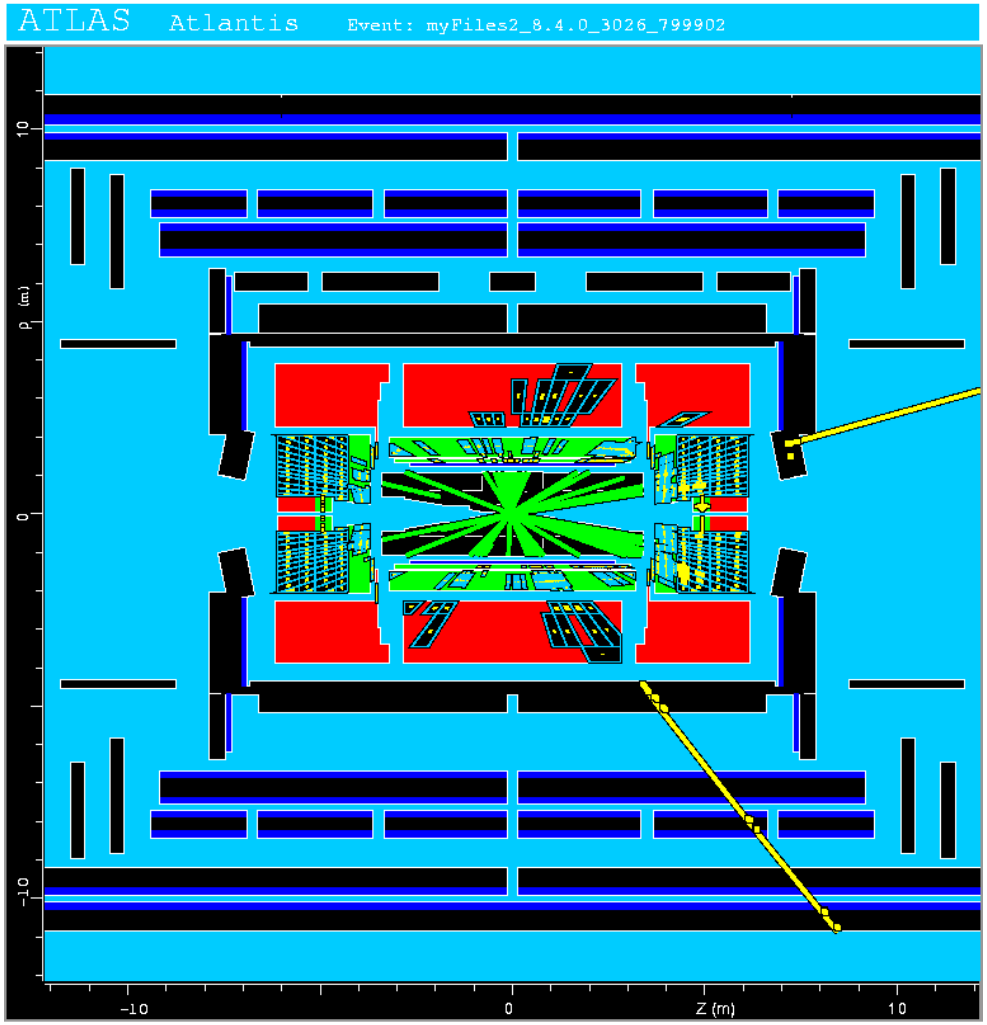


p

p

missing transverse energy

Background events



This event from Standard Model ttbar production also has high p_T jets and muons, and some missing transverse energy.

→ can easily mimic a SUSY event.

Physics context of a statistical test

1) Event Selection: Data space = measured properties of individual event.

The event types corresponding to the different hypotheses are known to exist, e.g., separation of different particle types (electron vs muon) or known event types (ttbar vs QCD multijet).

E.g. test H_0 : event is background vs. H_1 : event is signal.
Use selected events for further study.

2) Search for New Physics: Data space = properties of a sample of events.

The null hypothesis is

H_0 : all events correspond to background (e.g. Standard Model),

and the alternative is

H_1 : events include a type whose existence is not yet established (signal plus background)

The optimal statistical test for a search is closely related to that used for event selection.

Statistical tests for event selection

Suppose the result of a measurement for an individual event is a collection of numbers $\vec{x} = (x_1, \ldots, x_n)$

x_1 = number of muons,

x_2 = mean p_T of jets,

x_3 = missing energy, ...

$\vec{x}$ follows some n-dimensional joint pdf, which depends on the type of event produced, i.e., was it

$$pp \to t\bar{t}\ , \quad pp \to \tilde{g}\tilde{g}\ , \ldots$$

For each reaction we consider we will have a hypothesis for the pdf of $\vec{x}$, e.g., $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, etc.

E.g. call H_0 the background hypothesis (the event type we want to reject); H_1 is signal hypothesis (the type we want).

Selecting events

Suppose two kinds of events, corresponding to hypotheses H_0 (background) and H_1 (signal) and we want to select those of signal type.

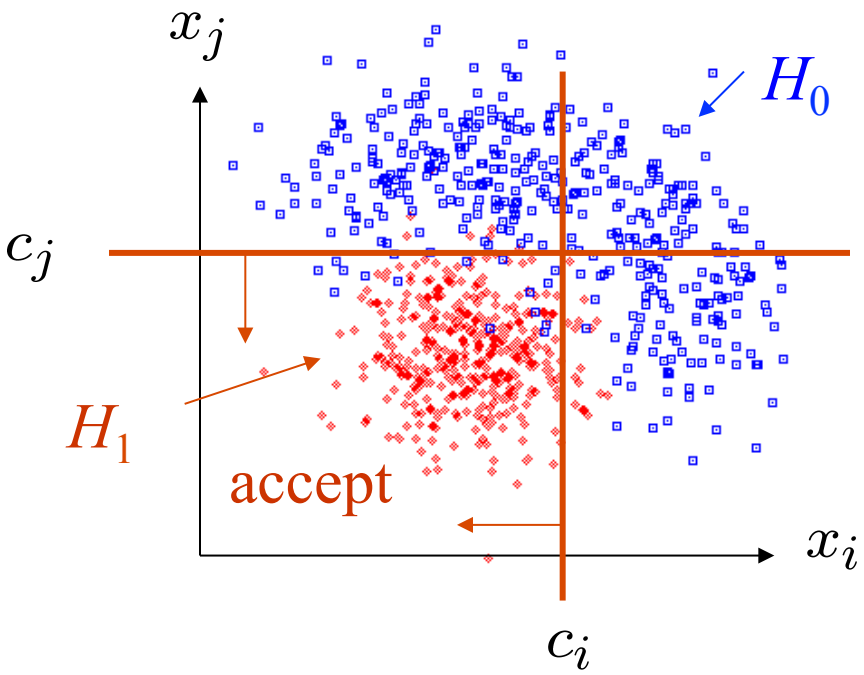
Formally do this by constructing a test of H_0 (background). If the H_0 is rejected, the event is "accepted" as candidate signal.

What is the best critical region ("decision boundary") for this?

Perhaps select events with 'cuts':

$$x_i < c_i$$

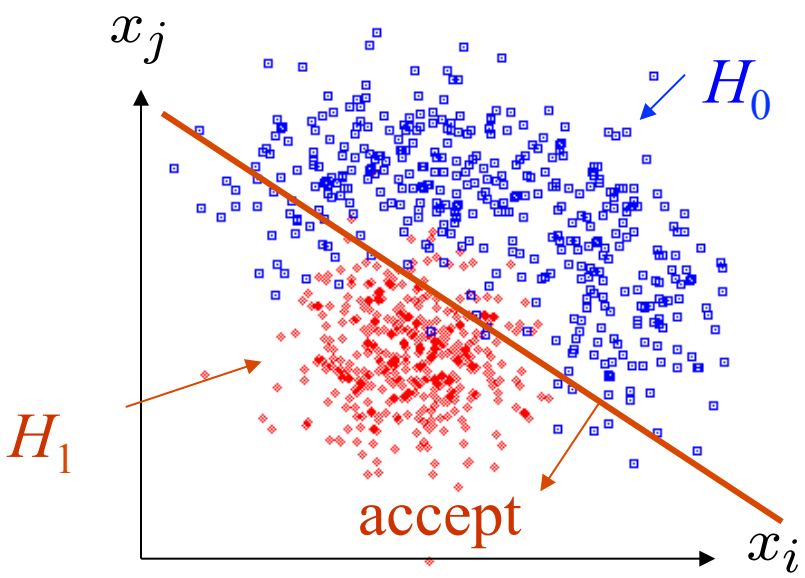
$$x_j < c_j$$



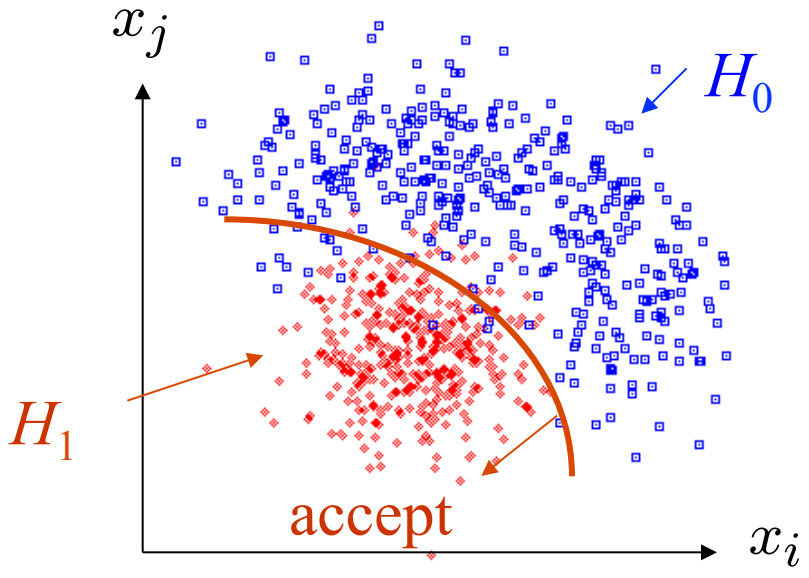
Other ways to select events

Or maybe use some other sort of boundary:

linear



or nonlinear



How can we do this in an 'optimal' way?

Test statistics

The boundary of the critical region for an n-dimensional data space $\boldsymbol{x} = (x_1,..., x_n)$ can be defined by an equation of the form

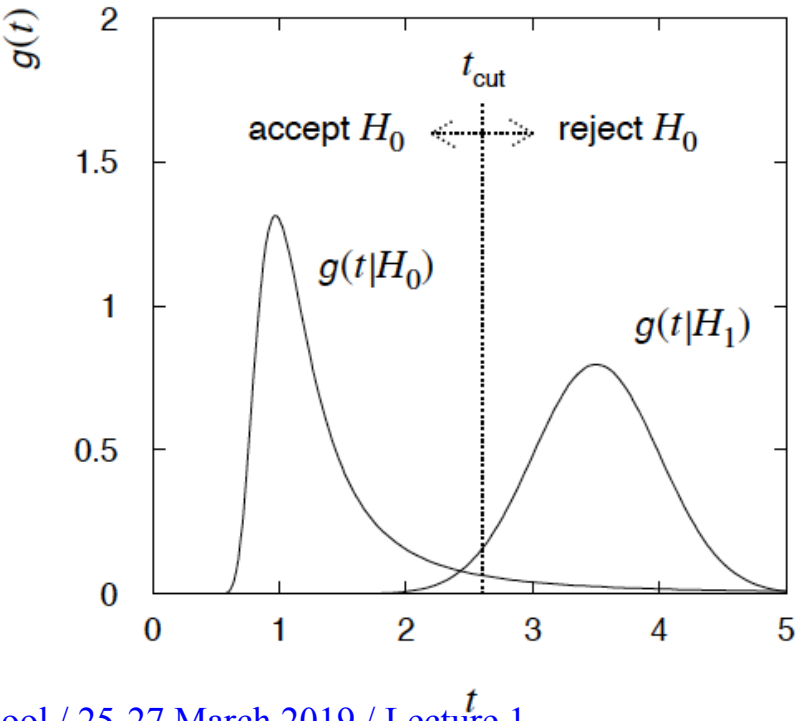
$$t(x_1, \ldots, x_n) = t_{\text{cut}}$$

where $t(x_1,\ldots, x_n)$ is a scalar test statistic.

We can work out the pdfs $g(t|H_0),\ g(t|H_1),\ \ldots$

Decision boundary is now a single 'cut' on t, defining the critical region.

So for an n-dimensional problem we have a corresponding 1-d problem.



Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way'?

Neyman-Pearson lemma states:

To get the highest power for a given significance level in a test of H_0, (background) versus H_1, (signal) the critical region should have

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > c$$

inside the region, and $\leq c$ outside, where c is a constant chosen to give a test of the desired size.

Equivalently, optimal scalar test statistic is

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)}$$

N.B. any monotonic function of this is leads to the same test.

Efficiencies, purity

Let H_0 = b (event is background, H_1 = s (event is signal).

For each event test b. If b rejected, "accept" as candidate signal.

background efficiency = $\varepsilon_b = P(\boldsymbol{x} \in W \,|\, \mathrm{b}) = \alpha$

signal efficiency = ε_s = power = $P(\boldsymbol{x} \in W \,|\, \mathrm{s}) = 1 - \beta$

To find purity of candidate signal sample, use Bayes' theorem:

Here W is signal region

ε_s

prior probability

$$P(\mathrm{s}|\mathbf{x} \in W) = \frac{P(\mathbf{x} \in W|\mathrm{s})P(\mathrm{s})}{P(\mathbf{x} \in W|\mathrm{s})P(\mathrm{s}) + P(\mathbf{x} \in W|\mathrm{b})P(\mathrm{b})}$$

ε_b

posterior probability = signal purity

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(\boldsymbol{x}|\mathrm{s})$, $f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$

Instead we may have Monte Carlo models for signal and background processes, so we can produce simulated data:

$$\text{generate } \boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{s}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$$

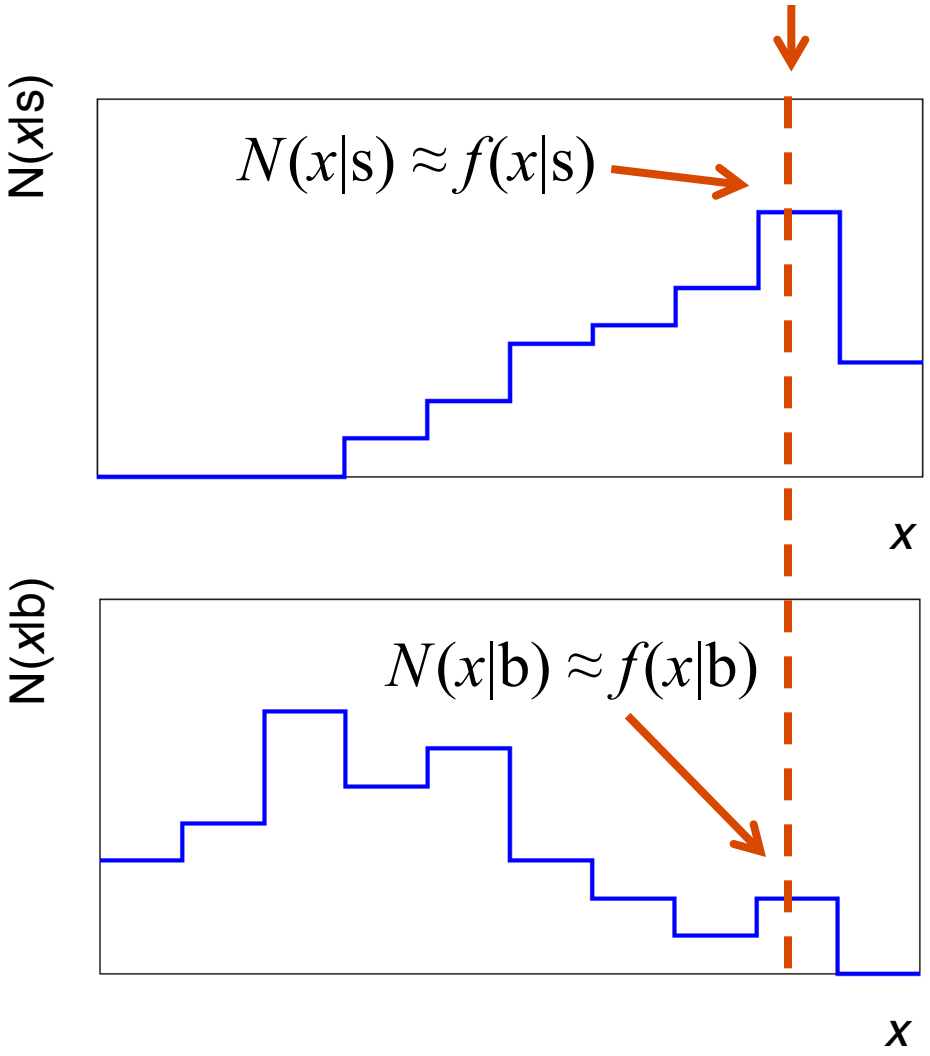
$$\text{generate } \boldsymbol{x} \sim f(\boldsymbol{x}|\mathrm{b}) \quad \rightarrow \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$$

This gives samples of "training data" with events of known type.

Can be expensive (1 fully simulated LHC event ~ 1 CPU minute).

Approximate LR from histograms

Want $t(x) = f(x|\mathrm{s})/f(x|\mathrm{b})$ for x here



One possibility is to generate MC data and construct histograms for both signal and background.

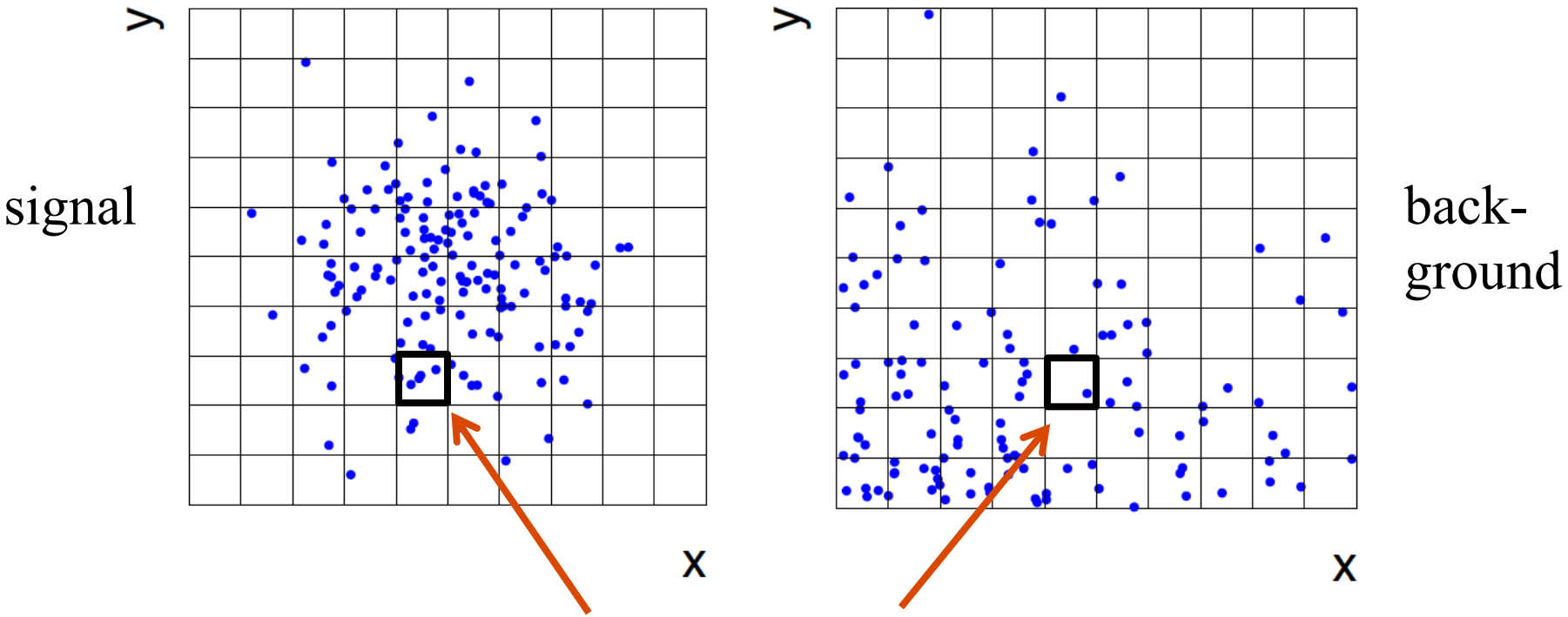
Use (normalized) histogram values to approximate LR:

$$t(x) \approx \frac{N(x|s)}{N(x|b)}$$

Can work well for single variable.

Approximate LR from 2D-histograms

Suppose problem has 2 variables. Try using 2-D histograms:



Approximate pdfs using $N(x,y|\mathrm{s})$, $N(x,y|\mathrm{b})$ in corresponding cells.

But if we want M bins for each variable, then in n-dimensions we have M^n cells; can't generate enough training data to populate.

→ Histogram method usually not usable for $n > 1$ dimension.

Strategies for multivariate analysis

Neyman-Pearson lemma gives optimal answer, but cannot be used directly, because we usually don't have $f(\boldsymbol{x}|\mathrm{s})$, $f(\boldsymbol{x}|\mathrm{b})$.

Histogram method with M bins for n variables requires that we estimate M^n parameters (the values of the pdfs in each cell), so this is rarely practical.

A compromise solution is to assume a certain functional form for the test statistic $t(\boldsymbol{x})$ with fewer parameters; determine them (using MC) to give best separation between signal and background.

Alternatively, try to estimate the probability densities $f(\boldsymbol{x}|\mathrm{s})$ and $f(\boldsymbol{x}|\mathrm{b})$ (with something better than histograms) and use the estimated pdfs to construct an approximate likelihood ratio.

Multivariate methods (→ Machine Learning)

Many new (and some old) methods:

- Fisher discriminant
- (Deep) neural networks
- Kernel density methods
- Support Vector Machines
- Decision trees
  - Boosting
  - Bagging

Much more on this in the lectures by Tommaso Dorigo

Resources on multivariate methods

C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006

T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, 2009

R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd ed., Wiley, 2001

A. Webb, Statistical Pattern Recognition, 2nd ed., Wiley, 2002.

Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014.

朱永生（编著），实验数据多元统计分析，科学出版社，北京，2009。

Testing significance / goodness-of-fit

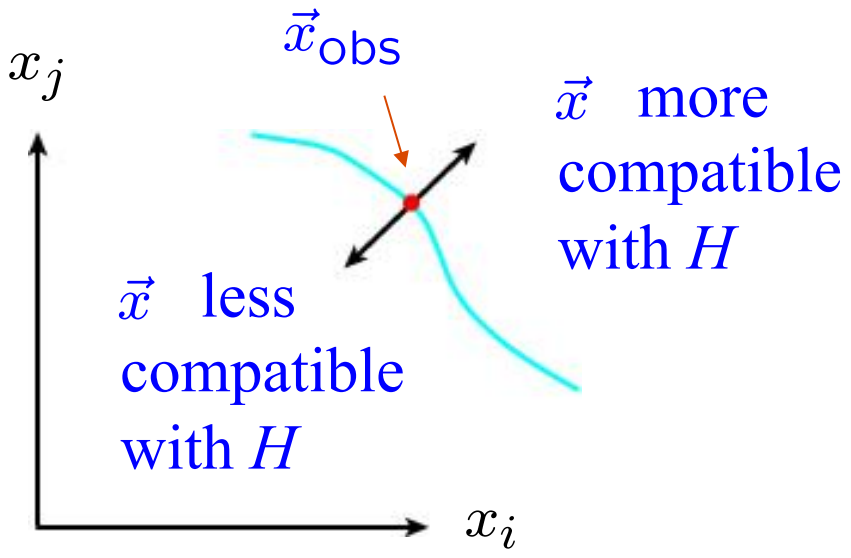
Suppose hypothesis H predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_1, \ldots, x_n)$.

We observe a single point in this space: $\vec{x}_{\text{obs}}$

What can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point $\boldsymbol{x}_{\text{obs}}$.

Note – "less compatible with H" means "more compatible with some alternative H' ".



p-values

Express 'goodness-of-fit' by giving the *p*-value for *H*:

p = probability, under assumption of *H*, to observe data with equal or lesser compatibility with *H* relative to the data we got.

This is not the probability that *H* is true!

In frequentist statistics we don't talk about *P*(*H*) (unless *H* represents a repeatable observation). In Bayesian statistics we do; use Bayes' theorem to obtain

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

where $\pi(H)$ is the prior probability for *H*.

For now stick with the frequentist approach;
result is *p*-value, regrettably easy to misinterpret as *P*(*H*).

Distribution of the p-value

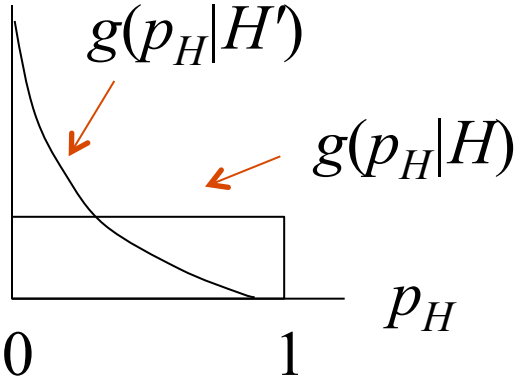
The p-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p-value of H is found from a test statistic $t(\boldsymbol{x})$ as

$$p_H = \int_t^\infty f(t'|H)dt'$$

The pdf of p_H under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \le p_H \le 1)$$

In general for continuous data, under assumption of H, $p_H \sim$ Uniform[0,1] and is concentrated toward zero for Some class of relevant alternatives.



Using a p-value to define test of H_0

One can show the distribution of the p-value of H, under assumption of H, is uniform in [0,1].

So the probability to find the p-value of H_0, p_0, less than α is

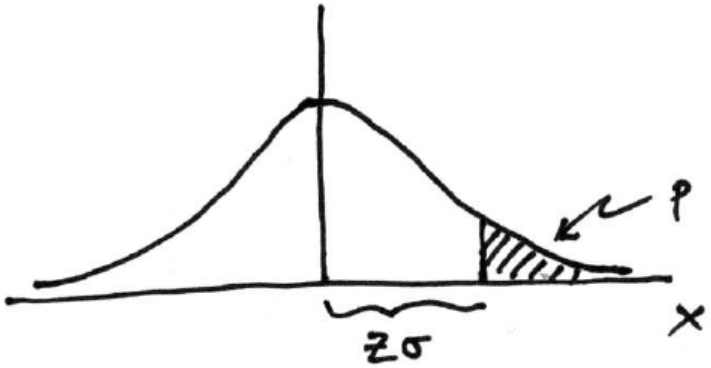
$$P(p_0 \leq \alpha | H_0) = \alpha$$

We can define the critical region of a test of H_0 with size α as the set of data space where $p_0 \leq \alpha$.

Formally the p-value relates only to H_0, but the resulting test will have a given power with respect to a given alternative H_1.

Significance from p-value

Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

E.g. $Z = 5$ (a "5 sigma effect") corresponds to $p = 2.9 \times 10^{-7}$.

The Poisson counting experiment

Suppose we do a counting experiment and observe n events.

Events could be from *signal* process or from *background* – we only count the total number.

Poisson model:

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

s = mean (i.e., expected) # of signal events

b = mean # of background events

Goal is to make inference about s, e.g.,

test $s = 0$ (rejecting $H_0 \approx$ "discovery of signal process")

test all non-zero s (values not rejected = confidence interval)

In both cases need to ask what is relevant alternative hypothesis.

Poisson counting experiment: discovery p-value

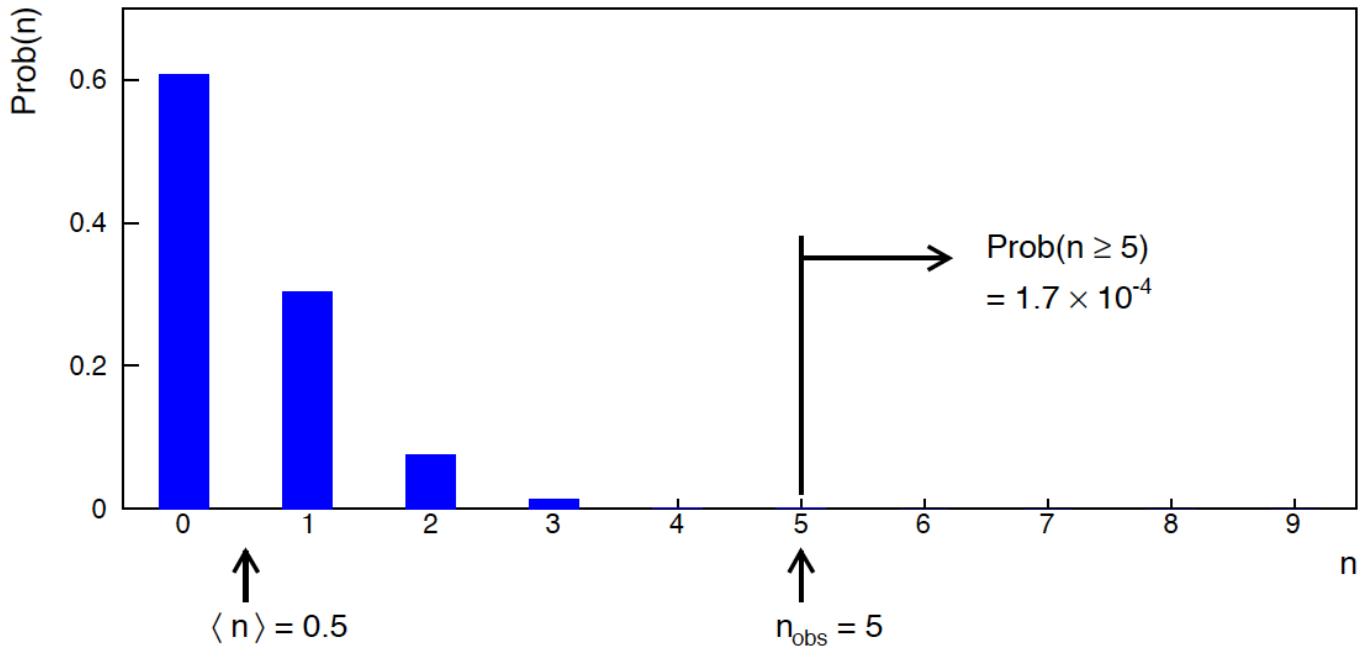
Suppose $b = 0.5$ (known), and we observe $n_{\text{obs}} = 5$.

Should we claim evidence for a new discovery?

Give p-value for hypothesis $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$

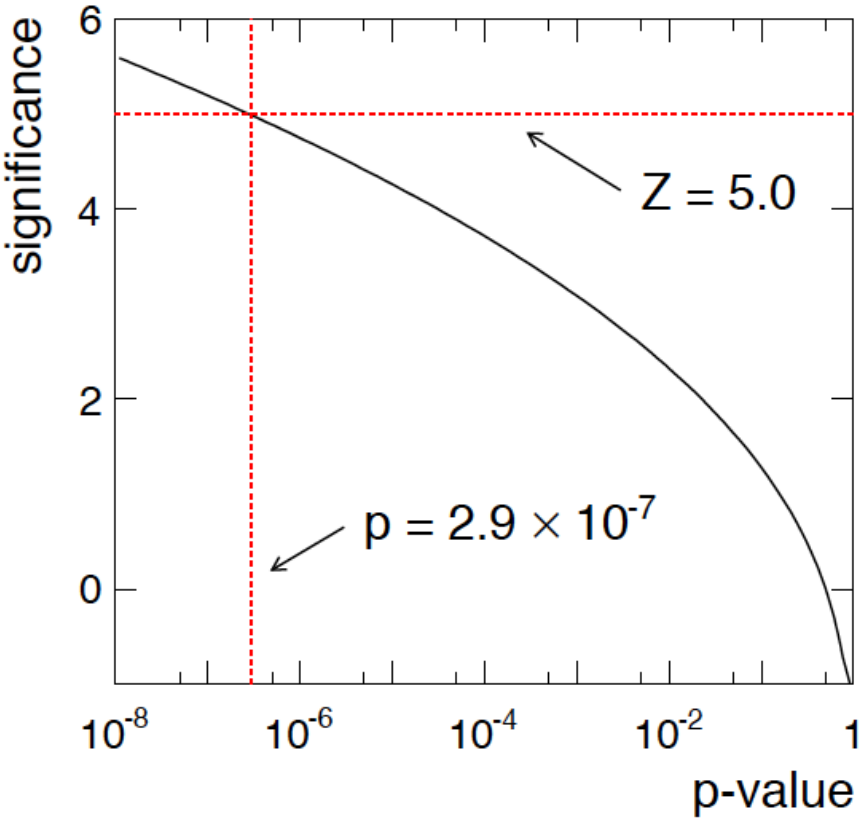
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$



Poisson counting experiment: discovery significance

Equivalent significance for $p = 1.7 \times 10^{-4}$: $Z = \Phi^{-1}(1 - p) = 3.6$

Often claim discovery if $Z > 5$ ($p < 2.9 \times 10^{-7}$, i.e., a "5-sigma effect")



In fact this tradition should be revisited: *p*-value intended to quantify probability of a signal-like fluctuation assuming background only; not intended to cover, e.g., hidden systematics, plausibility signal model, compatibility of data with signal, "look-elsewhere effect" (~multiple testing), etc.

Extra slides

Some distributions

Distribution/pdf	Example use in HEP
Binomial	Branching ratio
Multinomial	Histogram with fixed N
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student's t	Resolution function with adjustable tails

Binomial distribution

Consider N independent experiments (Bernoulli trials):

outcome of each is 'success' or 'failure',

probability of success on any given trial is p.

Define discrete r.v. n = number of successes ($0 \leq n \leq N$).

Probability of a specific outcome (in order), e.g. 'ssfsf' is

$$pp(1-p)p(1-p) = p^n(1-p)^{N-n}$$

But order not important; there are $\dfrac{N!}{n!(N-n)!}$

ways (permutations) to get n successes in N trials, total probability for n is sum of probabilities for each permutation.

Binomial distribution (2)

The binomial distribution is therefore

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

random variable

parameters

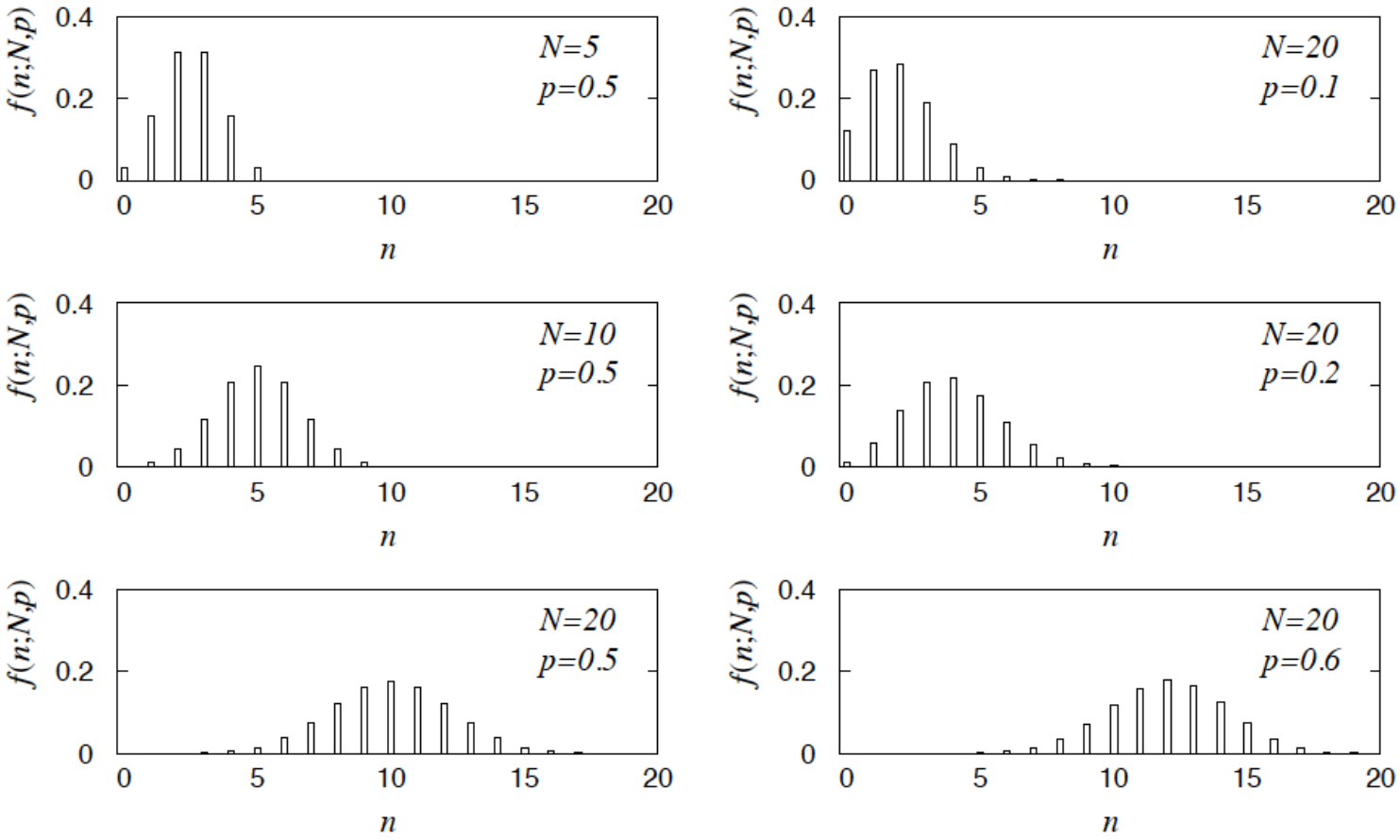
For the expectation value and variance we find:

$$E[n] = \sum_{n=0}^{N} n f(n; N, p) = Np$$

$$V[n] = E[n^2] - (E[n])^2 = Np(1-p)$$

Binomial distribution (3)

Binomial distribution for several values of the parameters:



Example: observe N decays of $W^{\pm}$, the number n of which are $W \to \mu\nu$ is a binomial r.v., p = branching ratio.

Multinomial distribution

Like binomial but now m outcomes instead of two, probabilities are

$$\vec{p} = (p_1, \ldots, p_m) , \quad \text{with} \quad \sum_{i=1}^{m} p_i = 1 .$$

For N trials we want the probability to obtain:

n_1 of outcome 1,
n_2 of outcome 2,
$\vdots$
n_m of outcome m.

This is the multinomial distribution for $\vec{n} = (n_1, \ldots, n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

Multinomial distribution (2)

Now consider outcome i as 'success', all others as 'failure'.

→ all n_i individually binomial with parameters N, p_i

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i) \quad \text{for all } i$$

One can also find the covariance to be

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

Example: $\vec{n} = (n_1, \ldots, n_m)$ represents a histogram with m bins, N total entries, all entries independent.

Poisson distribution

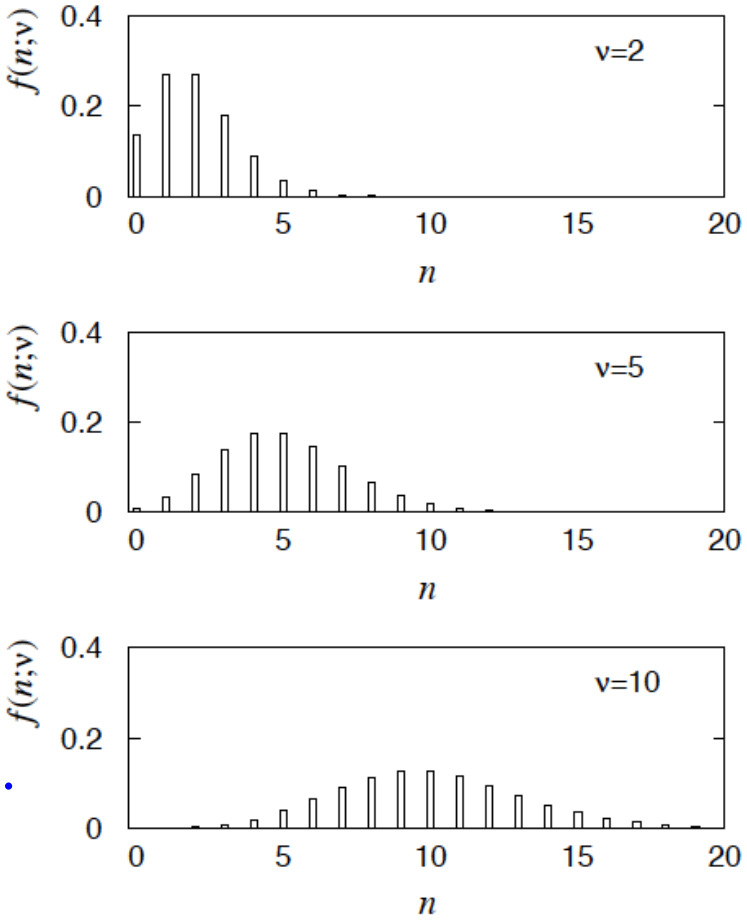
Consider binomial n in the limit

$$N \to \infty, \qquad p \to 0, \qquad E[n] = Np \to \nu \, .$$

→ n follows the Poisson distribution:

$$f(n;\nu) = \frac{\nu^n}{n!} e^{-\nu} \quad (n \geq 0)$$

$$E[n] = \nu \, , \quad V[n] = \nu \, .$$



Example: number of scattering events n with cross section σ found for a fixed integrated luminosity, with $\nu = \sigma \int L \, dt$.

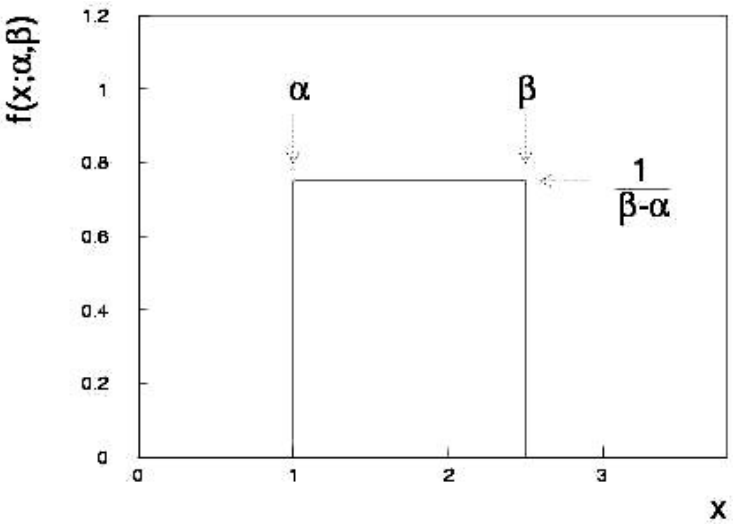
Uniform distribution

Consider a continuous r.v. x with $-\infty < x < \infty$. Uniform pdf is:

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

$$E[x] = \frac{1}{2}(\alpha+\beta)$$

$$V[x] = \frac{1}{12}(\beta-\alpha)^2$$



N.B. For any r.v. x with cumulative distribution $F(x)$, $y = F(x)$ is uniform in [0,1].

Example: for $\pi^0 \to \gamma\gamma$, E_γ is uniform in $[E_{\text{min}}, E_{\text{max}}]$, with

$$E_{\text{min}} = \frac{1}{2}E_\pi(1-\beta)\,, \qquad E_{\text{max}} = \frac{1}{2}E_\pi(1+\beta)$$

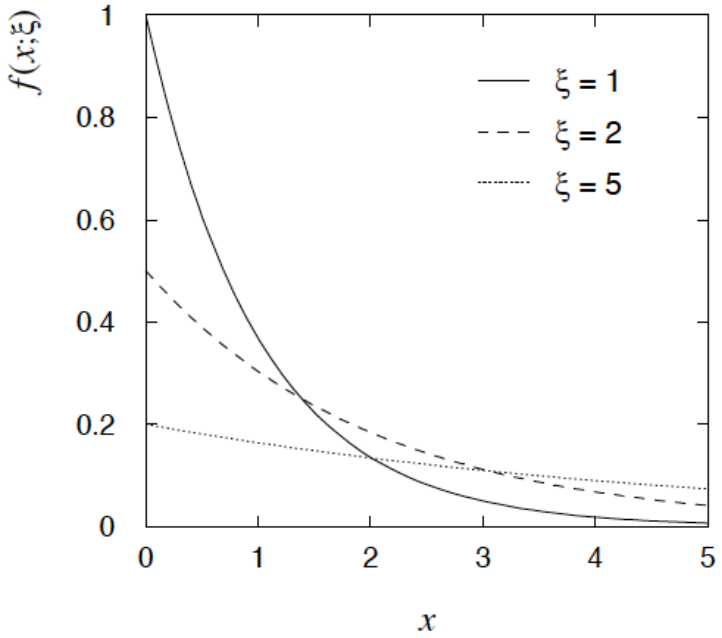
Exponential distribution

The exponential pdf for the continuous r.v. x is defined by:

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

$$E[x] = \xi$$

$$V[x] = \xi^2$$



Example: proper decay time t of an unstable particle

$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad (\tau = \text{mean lifetime})$$

Lack of memory (unique to exponential): $f(t - t_0 | t \geq t_0) = f(t)$

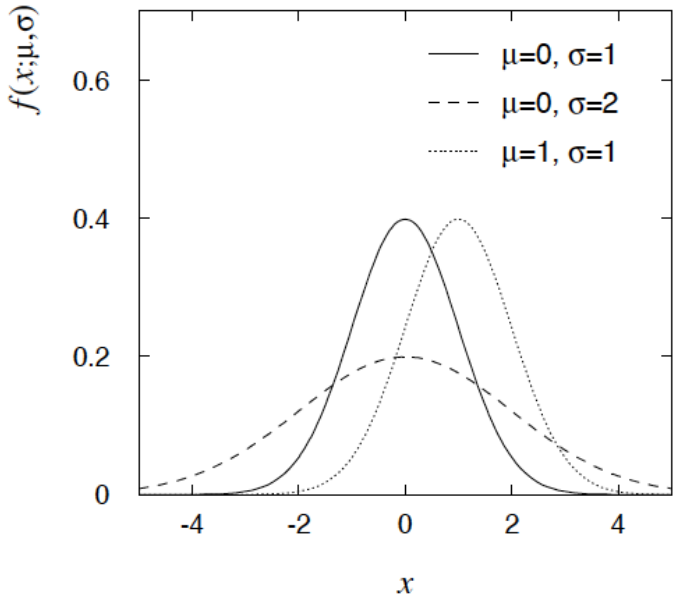
Gaussian distribution

The Gaussian (normal) pdf for a continuous r.v. x is defined by:

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

$E[x] = \mu$ (N.B. often μ, σ^2 denote mean, variance of any r.v., not only Gaussian.)

$V[x] = \sigma^2$



Special case: $\mu = 0$, $\sigma^2 = 1$ ('standard Gaussian'):

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^{x} \varphi(x')\,dx'$$

If $y \sim$ Gaussian with μ, σ^2, then $x = (y - \mu)/\sigma$ follows $\varphi(x)$.

Gaussian pdf and the Central Limit Theorem

The Gaussian pdf is so useful because almost any random variable that is a sum of a large number of small contributions follows it. This follows from the Central Limit Theorem:

For n independent r.v.s x_i with finite variances σ_i^2, otherwise arbitrary pdfs, consider the sum

$$y = \sum_{i=1}^{n} x_i$$

In the limit $n \to \infty$, y is a Gaussian r.v. with

$$E[y] = \sum_{i=1}^{n} \mu_i \qquad V[y] = \sum_{i=1}^{n} \sigma_i^2$$

Measurement errors are often the sum of many contributions, so frequently measured values can be treated as Gaussian r.v.s.

Central Limit Theorem (2)

The CLT can be proved using characteristic functions (Fourier transforms), see, e.g., SDA Chapter 10.

For finite n, the theorem is approximately valid to the extent that the fluctuation of the sum is not dominated by one (or few) terms.

Beware of measurement errors with non-Gaussian tails.

Good example: velocity component v_x of air molecules.

OK example: total deflection due to multiple Coulomb scattering. (Rare large angle deflections give non-Gaussian tail.)

Bad example: energy loss of charged particle traversing thin gas layer. (Rare collisions make up large fraction of energy loss, cf. Landau pdf.)

Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector $\vec{x} = (x_1, \ldots, x_n)$:

$$f(\vec{x}; \vec{\mu}, V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T V^{-1} (\vec{x} - \vec{\mu})\right]$$

$\vec{x}$, $\vec{\mu}$ are column vectors, $\vec{x}^T$, $\vec{\mu}^T$ are transpose (row) vectors,

$$E[x_i] = \mu_i, \ , \quad \text{cov}[x_i, x_j] = V_{ij} \ .$$

For $n = 2$ this is

$$f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

where $\rho = \text{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient.

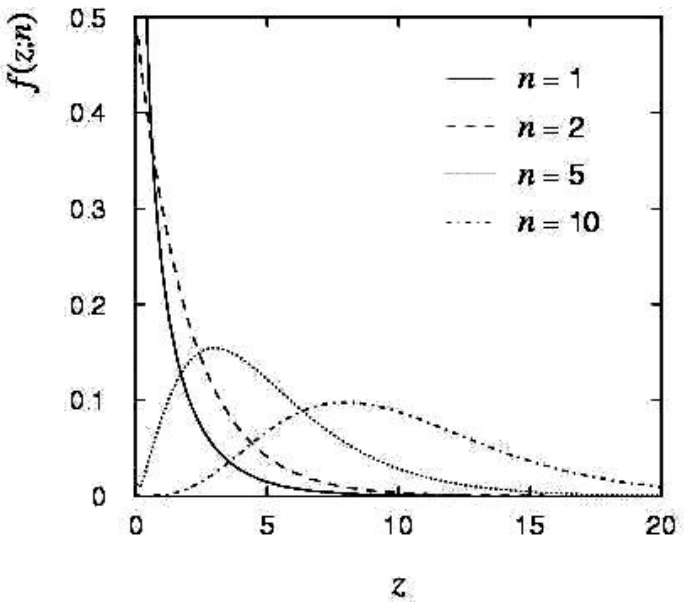
Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z ($z \geq 0$) is defined by

$$f(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

$n = 1, 2, \ldots =$ number of 'degrees of freedom' (dof)

$$E[z] = n\,, \qquad V[z] = 2n\,.$$



For independent Gaussian x_i, $i = 1, \ldots, n$, means μ_i, variances σ_i^2,

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$ follows χ^2 pdf with n dof.

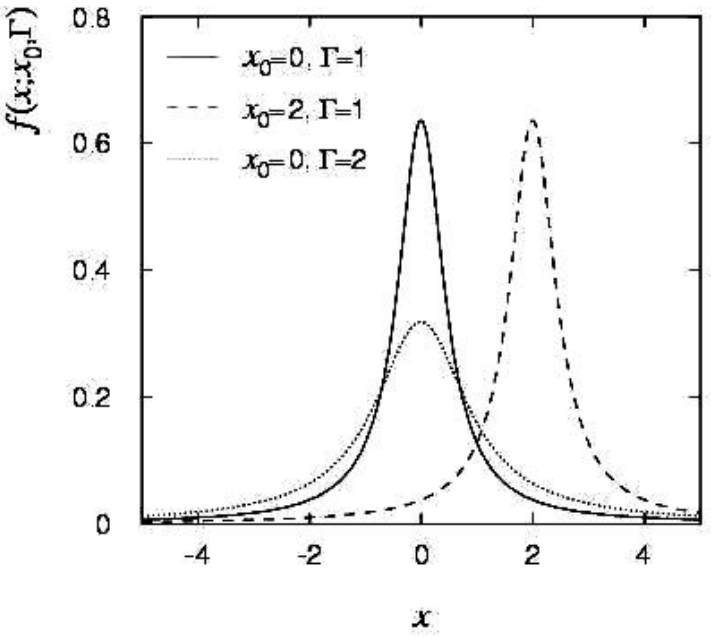
Example: goodness-of-fit test variable especially in conjunction with method of least squares.

Cauchy (Breit-Wigner) distribution

The Breit-Wigner pdf for the continuous r.v. x is defined by

$$f(x;\Gamma,x_0) = \frac{1}{\pi}\frac{\Gamma/2}{\Gamma^2/4 + (x-x_0)^2}$$

($\Gamma = 2$, $x_0 = 0$ is the Cauchy pdf.)



$E[x]$ not well defined, $V[x] \to \infty$.

x_0 = mode (most probable value)

Γ = full width at half maximum

Example: mass of resonance particle, e.g. ρ, K^*, ϕ^0, ...

Γ = decay rate (inverse of mean lifetime)

Landau distribution

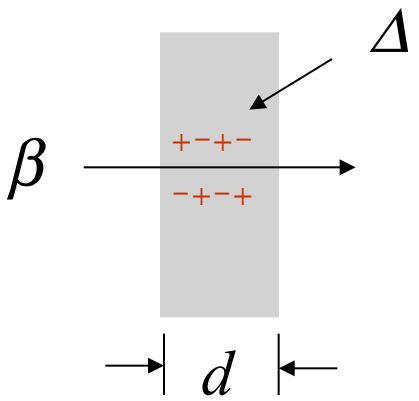
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau pdf:

$$f(\Delta;\beta) = \frac{1}{\xi}\phi(\lambda) \ ,$$

$$\phi(\lambda) = \frac{1}{\pi}\int_0^\infty \exp(-u\ln u - \lambda u)\sin\pi u\, du \ ,$$

$$\lambda = \frac{1}{\xi}\left[\Delta - \xi\left(\ln\frac{\xi}{\epsilon'} + 1 - \gamma_{\mathsf{E}}\right)\right] \ ,$$

$$\xi = \frac{2\pi N_{\mathsf{A}} e^4 z^2 \rho \sum Z}{m_{\mathsf{e}} c^2 \sum A}\frac{d}{\beta^2} \ , \qquad \epsilon' = \frac{I^2 \exp\beta^2}{2m_{\mathsf{e}}c^2\beta^2\gamma^2} \ .$$



L. Landau, J. Phys. USSR **8** (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

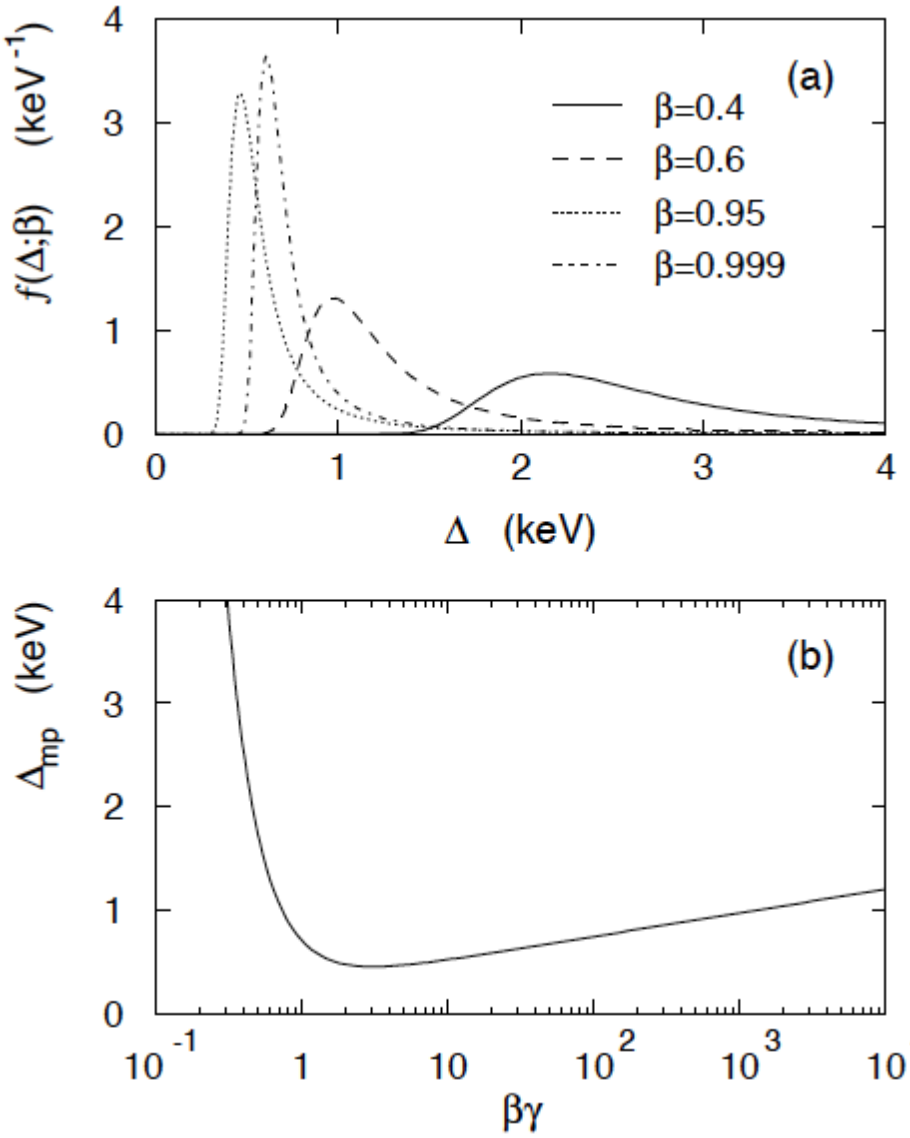
Landau distribution (2)

Long 'Landau tail'

→ all moments ∞

Mode (most probable value) sensitive to β,

→ particle i.d.



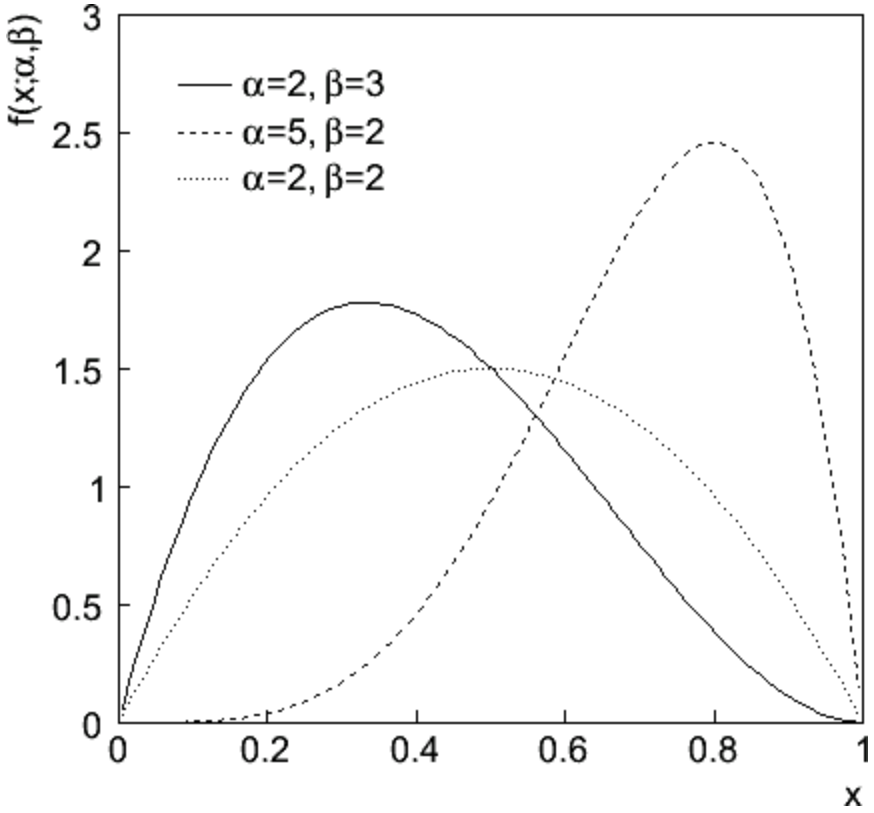
Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha+\beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Often used to represent pdf of continuous r.v. nonzero only between finite limits.



Gamma distribution

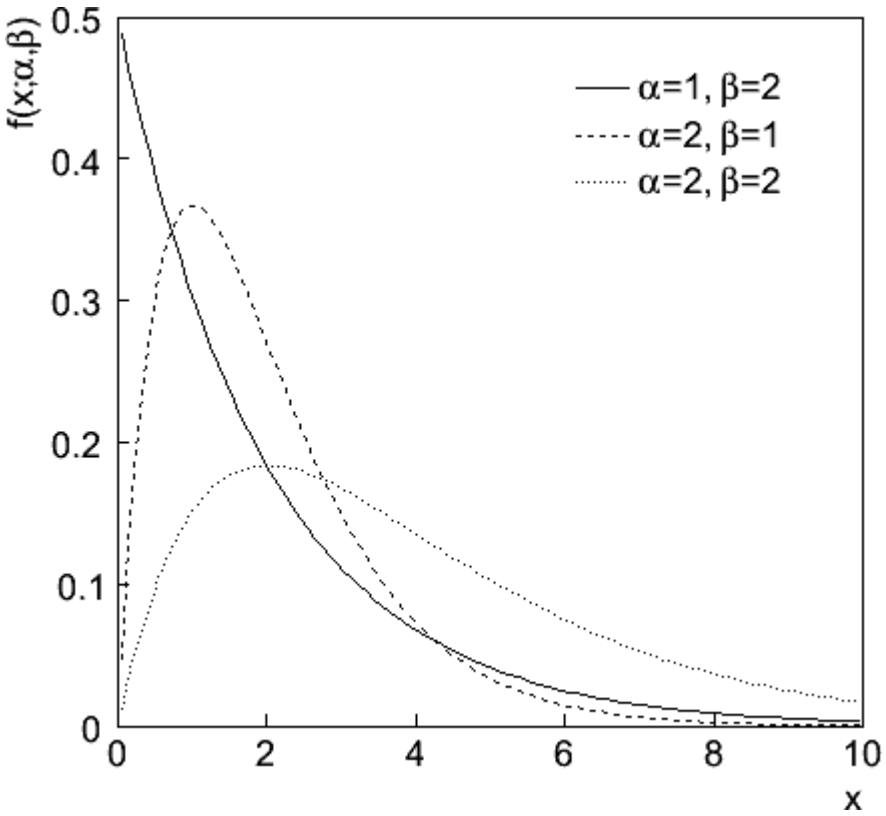
$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$

Often used to represent pdf of continuous r.v. nonzero only in $[0,\infty]$.

Also e.g. sum of *n* exponential r.v.s or time until *n*th event in Poisson process ~ Gamma



Student's *t* distribution

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

$$E[x] = 0 \quad (\nu > 1)$$

$$V[x] = \frac{\nu}{\nu-2} \quad (\nu > 2)$$

ν = number of degrees of freedom
(not necessarily integer)

$\nu = 1$ gives Cauchy,

$\nu \to \infty$ gives Gaussian.