Statistical Methods for Particle Physics

Lecture 1: intro, parameter estimation, tests

`http://indico.ihep.ac.cn/event/4902/`

iSTEP 2015
Shandong University, Jinan
August 11-19, 2015

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

Outline

→ Lecture 1: Introduction and review of fundamentals

- Probability, random variables, pdfs
- Parameter estimation, maximum likelihood
- Statistical tests for discovery and limits

Lecture 2: Multivariate methods

- Neyman-Pearson lemma
- Fisher discriminant, neural networks
- Boosted decision trees

Lecture 3: Systematic uncertainties and further topics

- Nuisance parameters (Bayesian and frequentist)
- Experimental sensitivity
- The look-elsewhere effect

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.

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)

K.A. Olive et al. (Particle Data Group), *Review of Particle Physics*, Chin. Phys. C, 38, 090001 (2014).; see also **`pdg.lbl.gov`** sections on probability, statistics, Monte Carlo

More statistics books（中文）

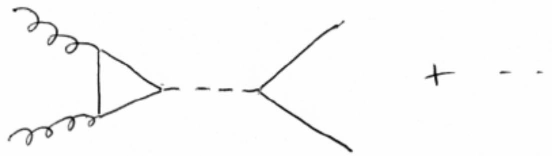
朱永生，实验物理中的概率和统计（第二版），科学出版社，北京，2006。

朱永生（编著），实验数据多元统计分析，科学出版社，北京，2009。

Theory ↔ Statistics ↔ Experiment

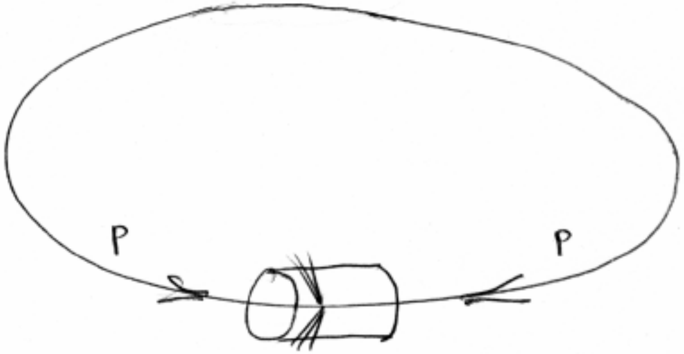
Theory (model, hypothesis):

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



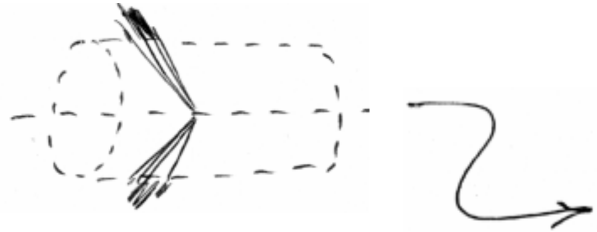
$$\sigma = \frac{G_F \, \alpha_s^2 \, m_H^2}{288 \sqrt{2\pi}} \times \sim$$

Experiment:



+ data selection

+ simulation of detector and cuts



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:

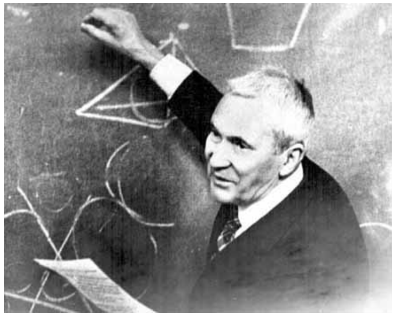
Estimate the free parameters of the theory: $m_H = 125.4$

Quantify the uncertainty in the estimates: $\pm\ 0.4$ 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, $P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$

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

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

Bayes' theorem

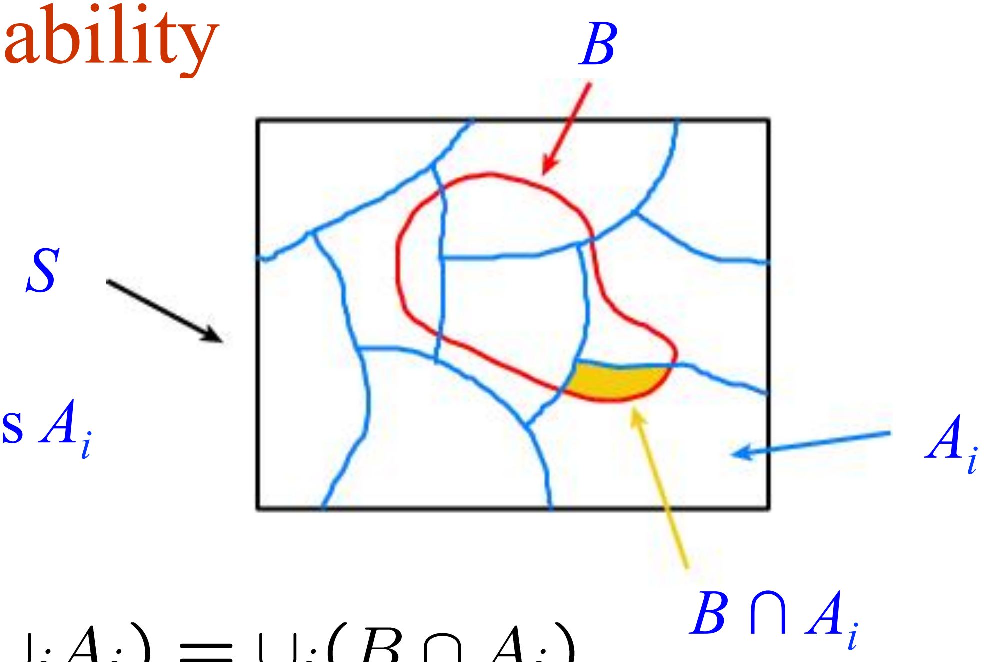
First published (posthumously) by the Reverend Thomas Bayes (1702−1761)

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 $\cup_i A_i = S$,



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

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

$$\rightarrow \quad P(B) = \textstyle\sum_i P(B|A_i) P(A_i)$$

law of total probability

Bayes' theorem becomes

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

An example using Bayes' theorem

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

$$\begin{aligned} P(\mathrm{D}) &= 0.001 \\ P(\text{no D}) &= 0.999 \end{aligned}$$

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

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

$$\begin{aligned} P(+|\mathrm{D}) &= 0.98 \\ P(-|\mathrm{D}) &= 0.02 \end{aligned}$$

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

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

← 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

$$
\begin{aligned}
p(\mathrm{D}|+) &= \frac{P(+|\mathrm{D})P(\mathrm{D})}{P(+|\mathrm{D})P(\mathrm{D}) + P(+|\text{no D})P(\text{no D})} \\
&= \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.03 \times 0.999} \\
&= 0.032 \qquad \leftarrow \text{posterior probability}
\end{aligned}
$$

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. $i = 1, 2, \ldots$ 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 $g(y)$,

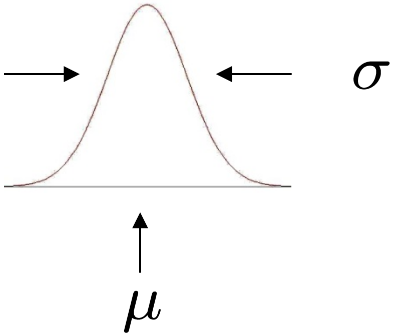
$$E[y] = \int y\, g(y)\, dy = \int y(x) f(x)\, dx$$ (equivalent)

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

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

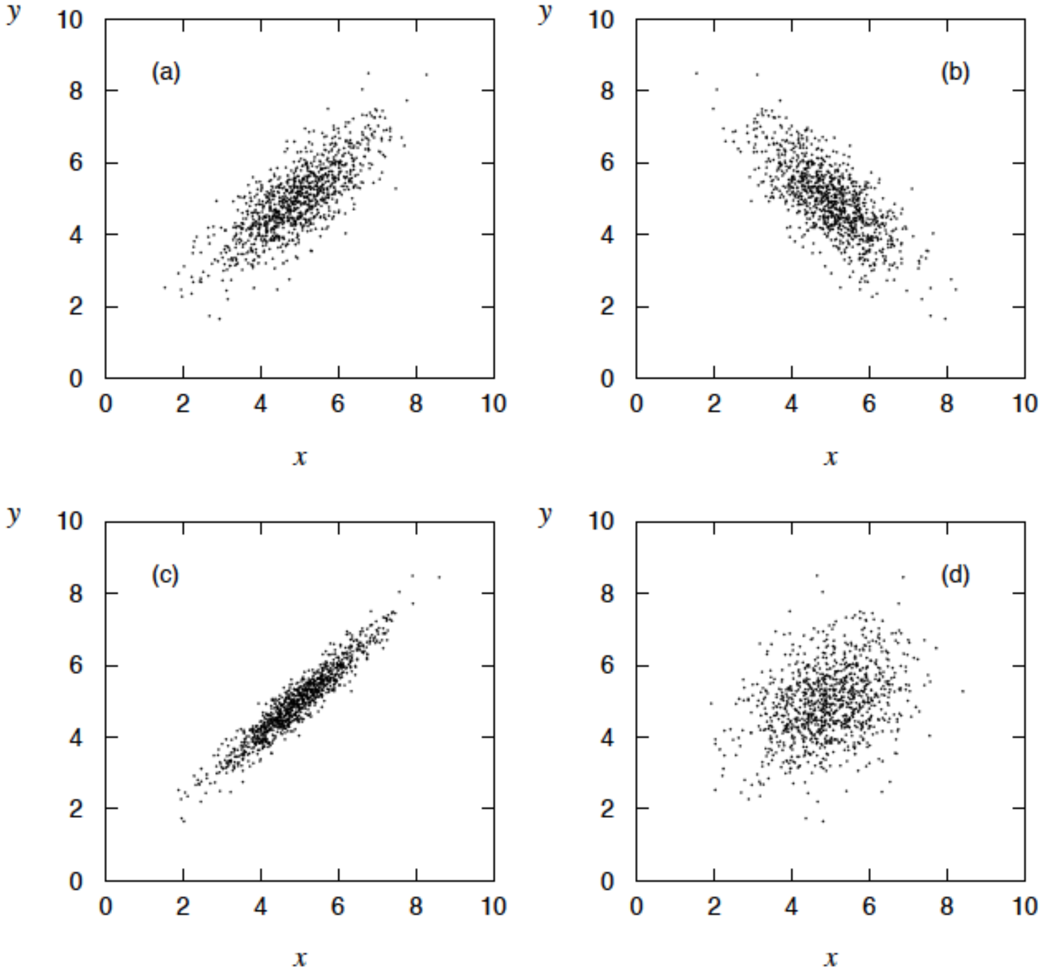
If x, y, independent, i.e., $f(x,y) = f_x(x)f_y(y)$, then

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

→ $\mathrm{cov}[x,y] = 0$ x and y, ‘uncorrelated’

N.B. converse not always true.

Correlation (cont.)



$\rho = 0.75$

$\rho = -0.75$

$\rho = 0.95$

$\rho = 0.25$

Review of frequentist parameter estimation

Suppose we have a pdf characterized by one or more parameters:

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$$

random variable parameter

Suppose we have a sample of observed values: $\vec{x} = (x_1, \ldots, x_n)$

We want to find some function of the data to estimate the parameter(s):

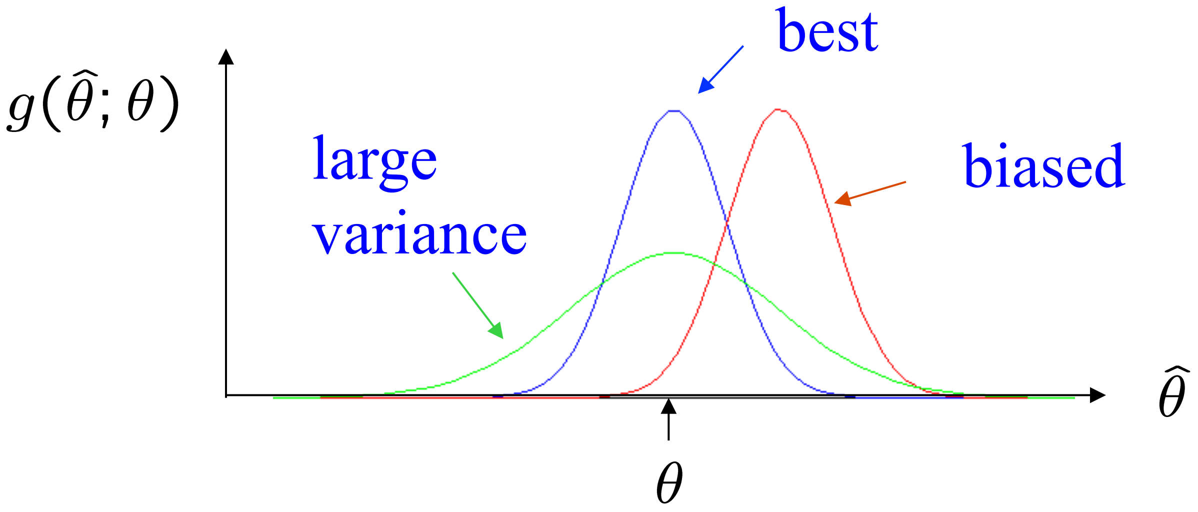
$$\boxed{\widehat{\theta}(\vec{x})}$$

← estimator written with a hat

Sometimes we say ‘estimator’ for the function of $x_1, \ldots, x_n$;
‘estimate’ for the value of the estimator with a particular data set.

Properties of estimators

If we were to repeat the entire measurement, the estimates from each would follow a pdf:



We want small (or zero) bias (systematic error): $b = E[\hat{\theta}] - \theta$

→ average of repeated measurements should tend to true value.

And we want a small variance (statistical error): $V[\hat{\theta}]$

→ small bias & variance are in general conflicting criteria

Distribution, likelihood, model

Suppose the outcome of a measurement is x. (e.g., a number of events, a histogram, or some larger set of numbers).

The probability density (or mass) function or 'distribution' of x, which may depend on parameters θ, is:

$P(x|\theta)$ (Independent variable is x; θ is a constant.)

If we evaluate $P(x|\theta)$ with the observed data and regard it as a function of the parameter(s), then this is the likelihood:

$L(\theta) = P(x|\theta)$ (Data x fixed; treat L as function of θ.)

We will use the term 'model' to refer to the full function $P(x|\theta)$ that contains the dependence both on x and θ.

Bayesian use of the term 'likelihood'

We can write Bayes theorem as

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta)\pi(\theta)\,d\theta}$$

where $L(x|\theta)$ is the likelihood. It is the probability for x given θ, evaluated with the observed x, and viewed as a function of θ.

Bayes' theorem only needs $L(x|\theta)$ evaluated with a given data set (the 'likelihood principle').

For frequentist methods, in general one needs the full model.

For some approximate frequentist methods, the likelihood is enough.

The likelihood function for i.i.d.*. data

* i.i.d. = independent and identically distributed

Consider n independent observations of x: $x_1, \ldots, x_n$, where x follows $f(x;\, \theta)$. The joint pdf for the whole data sample is:

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

In this case the likelihood function is

$$L(\vec{\theta}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}) \qquad (x_i \text{ constant})$$

Maximum likelihood

The most important frequentist method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(x|\theta)$

The resulting estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance: $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$

In general they may have a nonzero bias: $b = E[\hat{\theta}] - \theta$

Under conditions usually satisfied in practice, bias of ML estimators is zero in the large sample limit, and the variance is as small as possible for unbiased estimators.

ML estimator may not in some cases be regarded as the optimal trade-off between these criteria (cf. regularized unfolding).

ML example: parameter of exponential pdf

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

and suppose we have i.i.d. data, $t_1, \ldots, t_n$

The likelihood function is $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

The value of τ for which $L(\tau)$ is maximum also gives the maximum value of its logarithm (the log-likelihood function):

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

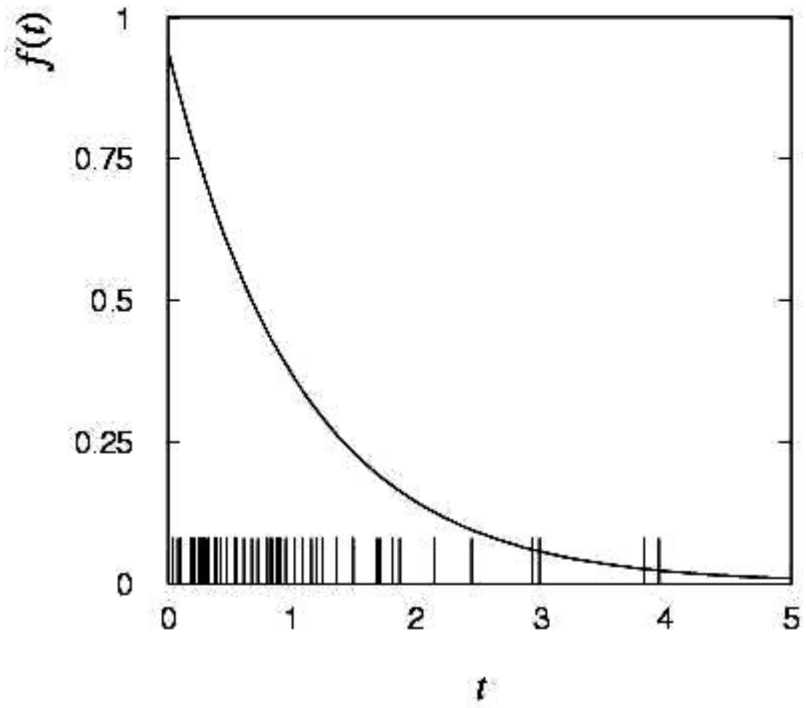
ML example: parameter of exponential pdf (2)

Find its maximum by setting $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$,

$$\rightarrow \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

Monte Carlo test:

generate 50 values
using $\tau = 1$:



We find the ML estimate:

$$\hat{\tau} = 1.062$$

ML example: parameter of exponential pdf (3)

For the exponential distribution one has for mean, variance:

$$E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$$

$$V[t] = \int_0^\infty (t - \tau)^2 \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$$

For the ML estimator $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n t_i$ we therefore find

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^n t_i\right] = \frac{1}{n}\sum_{i=1}^n E[t_i] = \tau \quad \longrightarrow \quad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^n t_i\right] = \frac{1}{n^2}\sum_{i=1}^n V[t_i] = \frac{\tau^2}{n} \quad \longrightarrow \quad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

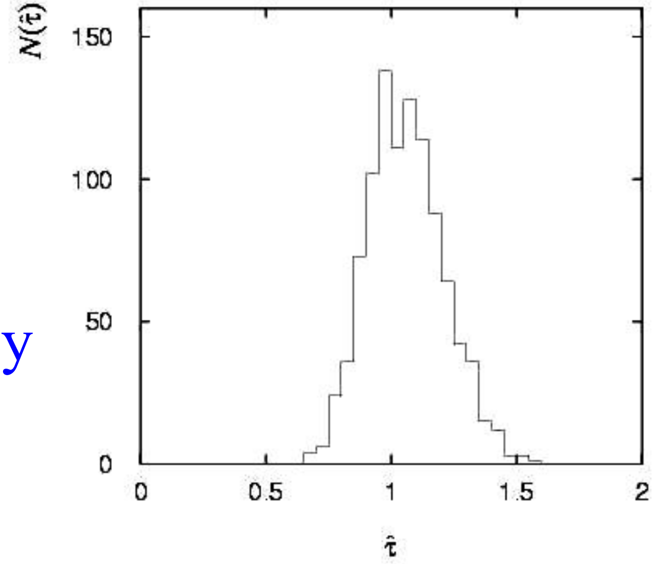
Variance of estimators: Monte Carlo method

Having estimated our parameter we now need to report its 'statistical error', i.e., how widely distributed would estimates be if we were to repeat the entire measurement many times.

One way to do this would be to simulate the entire experiment many times with a Monte Carlo program (use ML estimate for MC).

For exponential example, from sample variance of estimates we find:

$$\widehat{\sigma}_{\widehat{\tau}} = 0.151$$



Note distribution of estimates is roughly Gaussian − (almost) always true for ML in large sample limit.

Variance of estimators from information inequality

The information inequality (RCF) sets a lower bound on the variance of any estimator (not only ML):

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Minimum Variance Bound (MVB)

$(b = E[\hat{\theta}] - \theta)$

Often the bias b is small, and equality either holds exactly or is a good approximation (e.g. large data sample limit). Then,

$$V[\hat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Estimate this using the 2nd derivative of $\ln L$ at its maximum:

$$\widehat{V}[\hat{\theta}] = -\left.\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\right|_{\theta=\hat{\theta}}$$

Variance of estimators: graphical method

Expand ln L (θ) about its maximum:

$$\ln L(\theta) = \ln L(\widehat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta})^2 + \ldots$$

First term is ln $L_{\max}$, second term is zero, for third term use information inequality (assume equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \widehat{\theta})^2}{2\widehat{\sigma^2}_{\widehat{\theta}}}$$

i.e., $$\ln L(\widehat{\theta} \pm \widehat{\sigma}_{\widehat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

→ to get $\widehat{\sigma}_{\widehat{\theta}}$, change θ away from $\widehat{\theta}$ until ln L decreases by 1/2.

Example of variance by graphical method

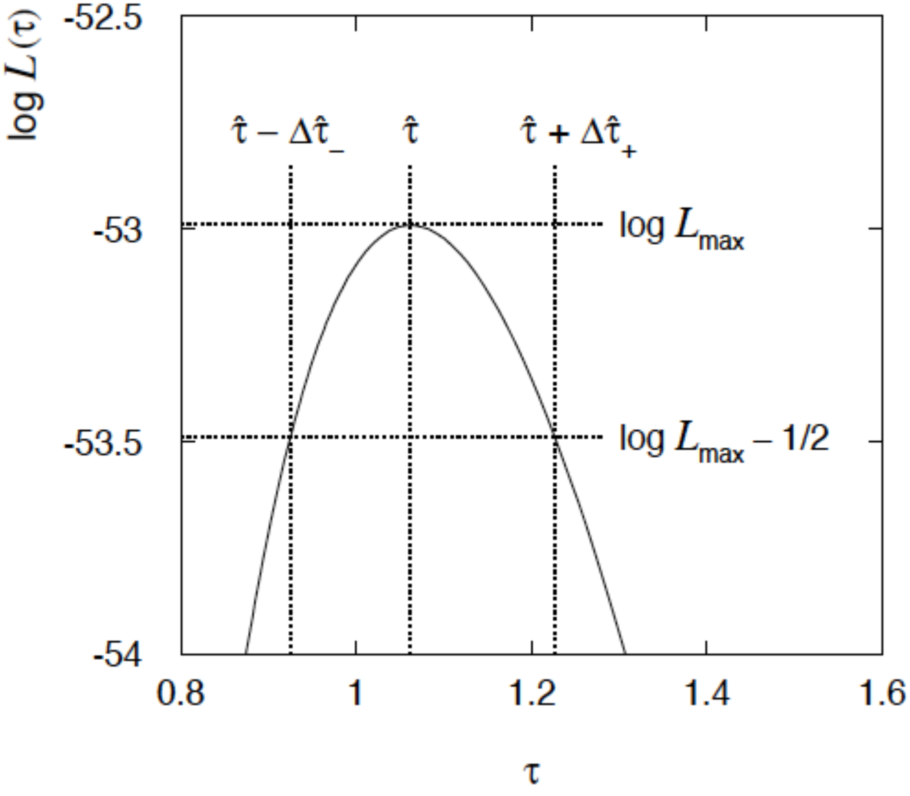
ML example with exponential:

$$\hat{\tau} = 1.062$$

$$\Delta\hat{\tau}_- = 0.137$$

$$\Delta\hat{\tau}_+ = 0.165$$

$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15$$



Not quite parabolic ln L since finite sample size ($n = 50$).

Information inequality for *n* parameters

Suppose we have estimated *n* parameters $\vec{\theta} = (\theta_1, \ldots, \theta_n)$.

The (inverse) minimum variance bound is given by the Fisher information matrix:

$$I_{ij} = E\left[-\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right] = -n \int f(x;\vec{\theta}) \frac{\partial^2 \ln f(x;\vec{\theta})}{\partial\theta_i \partial\theta_j} \, dx$$

The information inequality then states that $V - I^{-1}$ is a positive semi-definite matrix, where $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$. Therefore

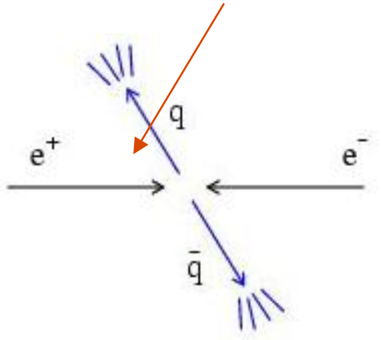
$$V[\hat{\theta}_i] \geq (I^{-1})_{ii}$$

Often use I^{-1} as an approximation for covariance matrix, estimate using e.g. matrix of 2nd derivatives at maximum of L.

Two-parameter example of ML

Consider a scattering angle distribution with $x = \cos\theta$,

$$f(x;\alpha,\beta) = \frac{1+\alpha x+\beta x^2}{2+2\beta/3}$$



Data: $x_1,\ldots, x_n$, $n = 2000$ events.

As test generate with MC using $\alpha = 0.5$, $\beta = 0.5$

From data compute log-likelihood:

$$\ln L(\alpha,\beta) = \sum_{i=1}^{n} \ln f(x_i;\alpha,\beta)$$

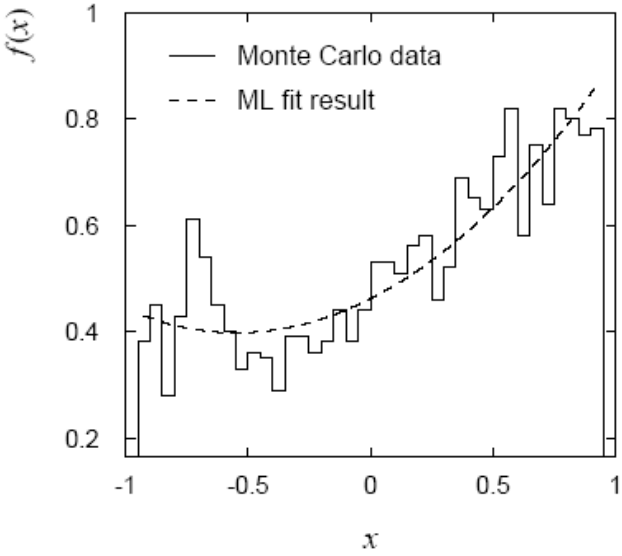
Maximize numerically (e.g., program MINUIT)

Example of ML: fit result

Finding maximum of ln $L(\alpha, \beta)$ numerically (**`MINUIT`**) gives

$$\hat{\alpha} = 0.508$$

$$\hat{\beta} = 0.47$$



N.B. Here no binning of data for fit, but can compare to histogram for goodness-of-fit (e.g. 'visual' or χ^2).

(Co)variances from $(\widehat{V^{-1}})_{ij} = -\left.\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right|_{\vec{\theta}=\hat{\vec{\theta}}}$ (**`MINUIT`** routine **`HESSE`**)

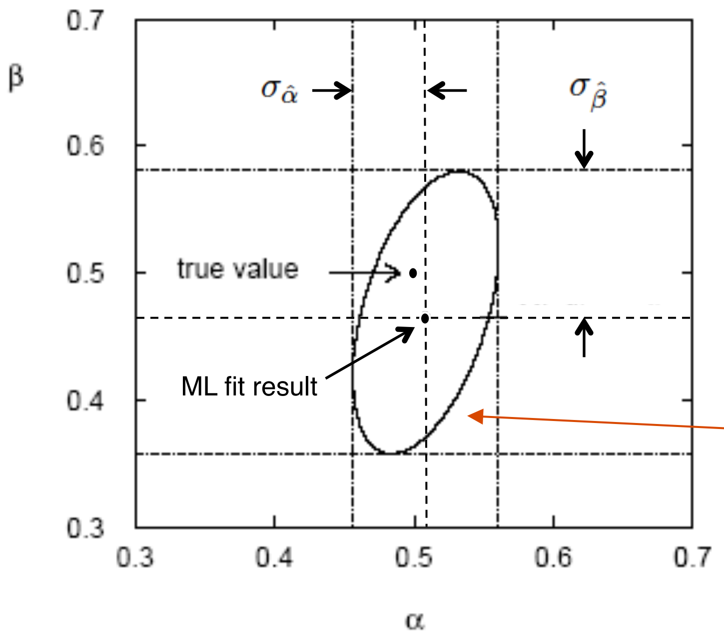
$$\hat{\sigma}_{\hat{\alpha}} = 0.052 \qquad \text{cov}[\hat{\alpha}, \hat{\beta}] = 0.0026$$

$$\hat{\sigma}_{\hat{\beta}} = 0.11 \qquad r = 0.46$$

Variance of ML estimators: graphical method

Often (e.g., large sample case) one can approximate the covariances using only the likelihood $L(\theta)$:

$$\widehat{V}_{ij}^{-1} \approx - \left. \frac{\partial^2 \ln L}{\partial \theta_i \, \partial \theta_j} \right|_{\theta = \hat{\theta}}$$



This translates into a simple graphical recipe:

$$\ln L(\alpha, \beta) = \ln L_{\max} - 1/2$$

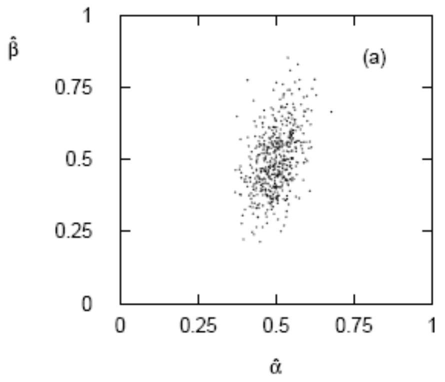
→ Tangent lines to contours give standard deviations.

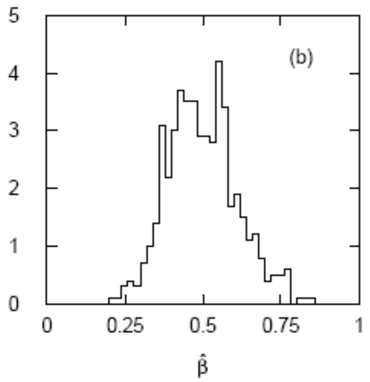
→ Angle of ellipse ϕ related to correlation: $\tan 2\phi = \frac{2\rho\sigma_{\widehat{\alpha}}\sigma_{\widehat{\beta}}}{\sigma_{\widehat{\alpha}}^2 - \sigma\widehat{\beta}2}$

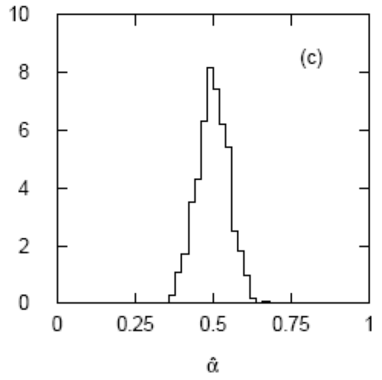
Variance of ML estimators: MC

To find the ML estimate itself one only needs the likelihood $L(\theta)$.

In principle to find the covariance of the estimators, one requires the full model $P(x|\theta)$. E.g., simulate many times independent data sets and look at distribution of the resulting estimates:







$$
\begin{aligned}
\overline{\hat{\alpha}} &= 0.499 \\
s_{\hat{\alpha}} &= 0.051 \\
\overline{\hat{\beta}} &= 0.498 \\
s_{\hat{\beta}} &= 0.111 \\
\widehat{\mathrm{cov}}[\hat{\alpha}, \hat{\beta}] &= 0.0024 \\
r &= 0.42
\end{aligned}
$$

Frequentist statistical tests

Consider a hypothesis H_0 and alternative H_1.

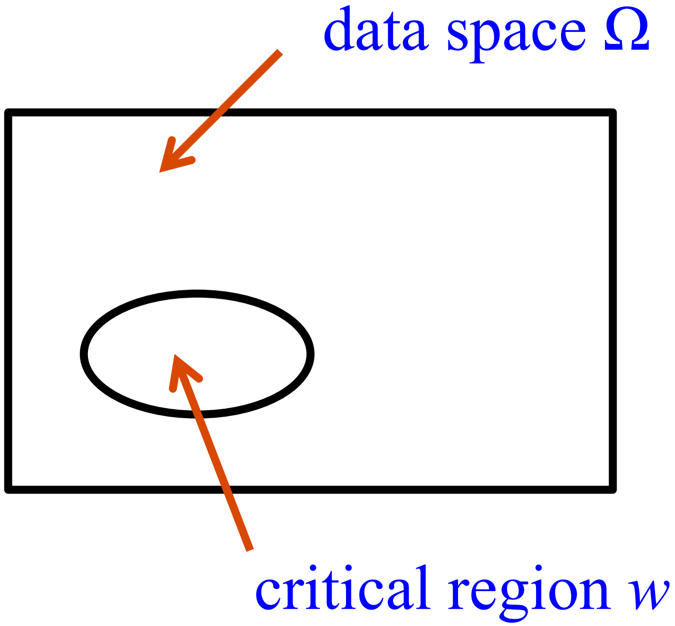
A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in w \mid H_0) \leq \alpha$$

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.

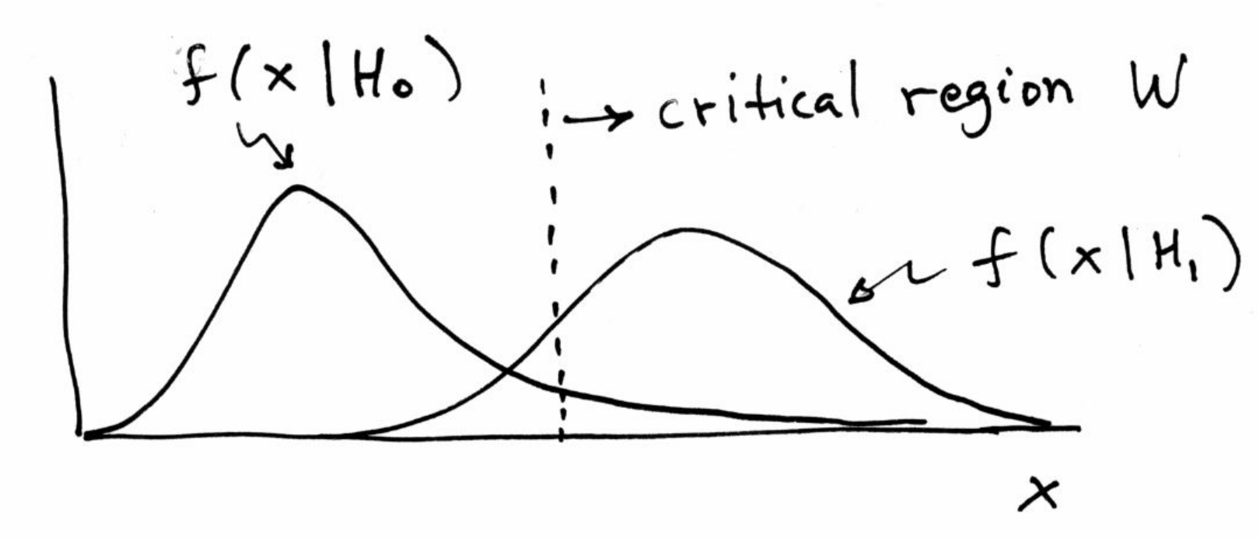


Definition of a test (2)

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:

$$P(x \in W \mid H_0) \le \alpha$$

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

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

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

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

p-values

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: $\vec{x}_{\mathrm{obs}}$

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

Express level of compatibility by giving the p-value for H:

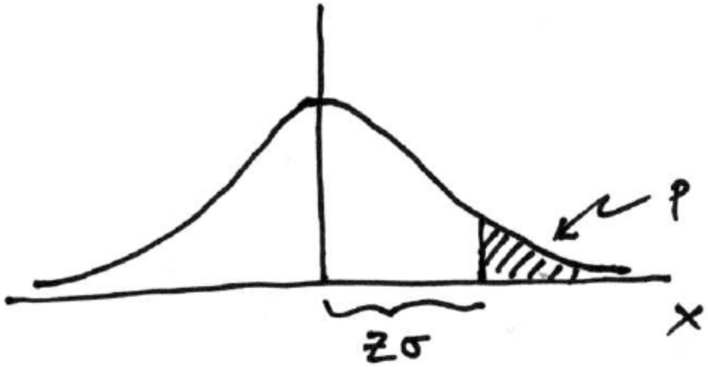
p = probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data we got.

⚠ This is not the probability that H is true!

Requires one to say what part of data space constitutes lesser compatibility with H than the observed data (implicitly this means that region gives better agreement with some alternative).

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 $p = 2.9 \times 10^{-7}$.

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.

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.

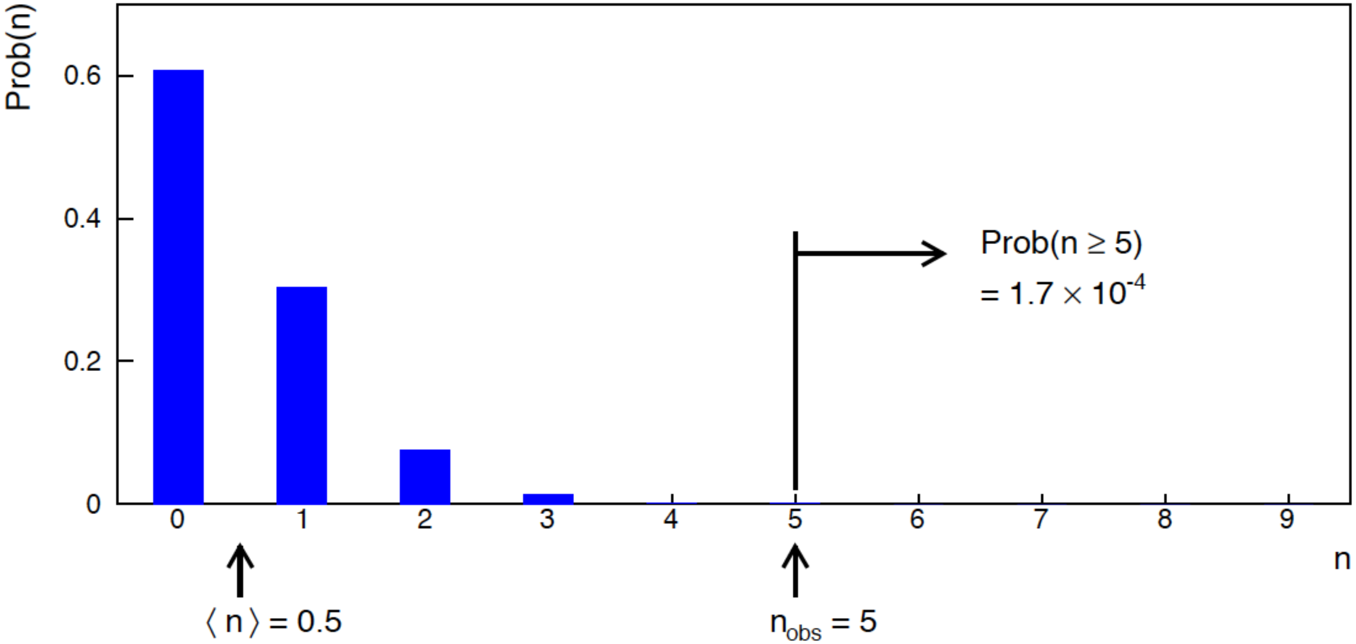
Poisson counting experiment: discovery p-value

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 $s = 0$:

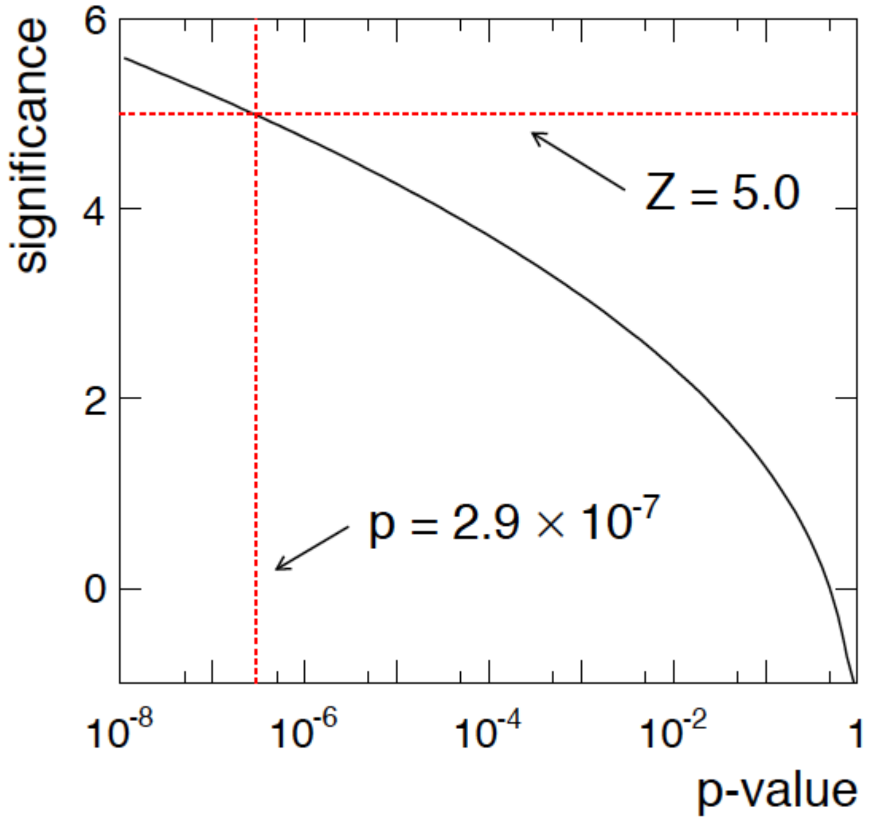
$$\begin{aligned} p\text{-value} &= P(n \geq 5; b = 0.5, s = 0) \\ &= 1.7 \times 10^{-4} \neq P(s = 0)! \end{aligned}$$



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.

Confidence intervals by inverting a test

Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

Specify values of the data that are 'disfavoured' by θ (critical region) such that P(data in critical region) $\leq \alpha$ for a prespecified α, e.g., 0.05 or 0.1.

If data observed in the critical region, reject the value θ.

Now invert the test to define a confidence interval as:

set of θ values that would not be rejected in a test of size α (confidence level is $1 - \alpha$).

The interval will cover the true value of θ with probability $\geq 1 - \alpha$.

Equivalently, the parameter values in the confidence interval have p-values of at least α.

To find edge of interval (the "limit"), set $p_\theta = \alpha$ and solve for θ.

Frequentist upper limit on Poisson parameter

Consider again the case of observing $n \sim \text{Poisson}(s + b)$.

Suppose $b = 4.5$, $n_{\text{obs}} = 5$. Find upper limit on s at 95% CL.

When testing s values to find upper limit, relevant alternative is $s = 0$ (or lower s), so critical region at low n and p-value of hypothesized s is $P(n \le n_{\text{obs}}; s, b)$.

Upper limit s_{up} at CL = $1 - \alpha$ from setting $\alpha = p_s$ and solving for s:

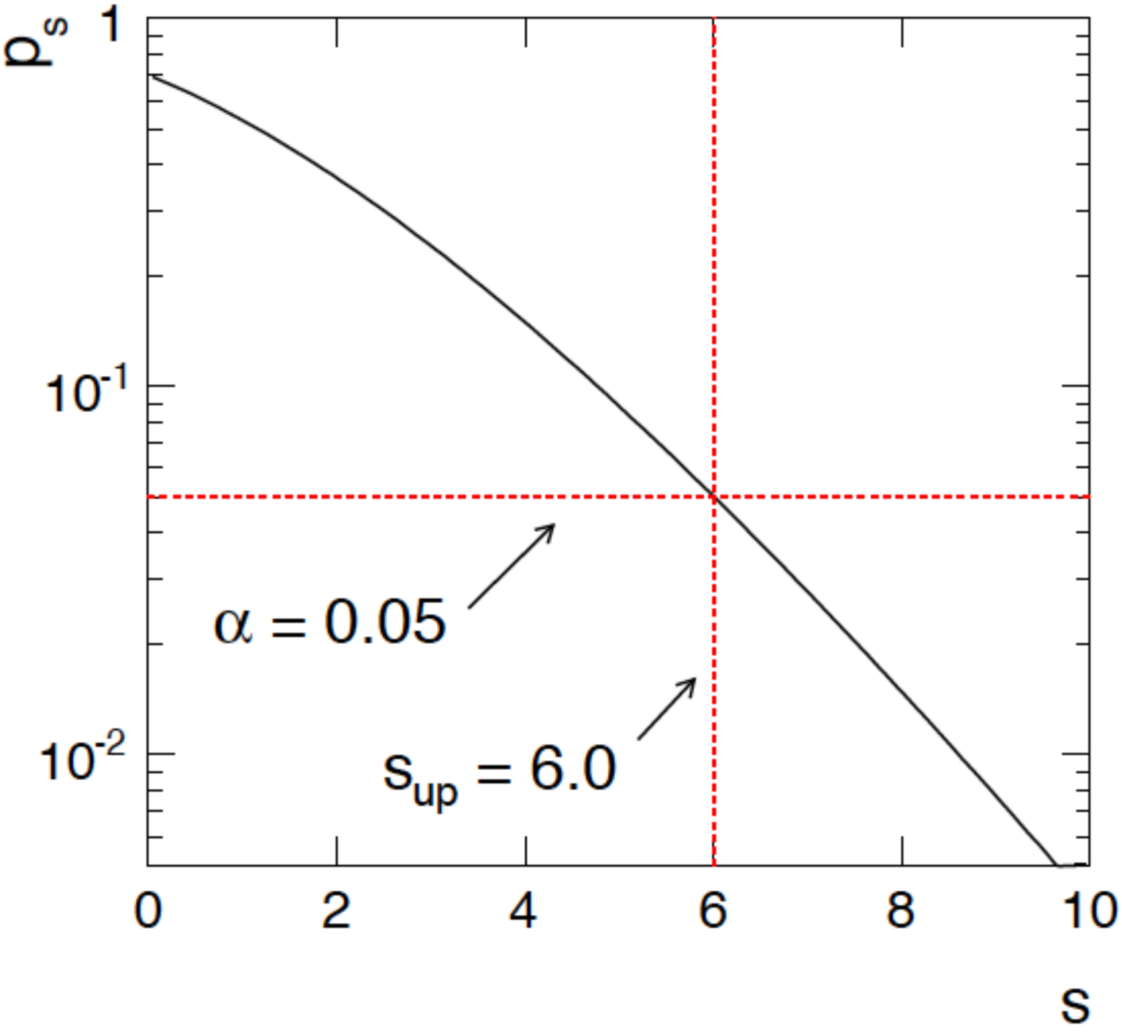
$$\alpha = P(n \le n_{\rm obs}; s_{\rm up}, b) = \sum_{n=0}^{n_{\rm obs}} \frac{(s_{\rm up} + b)^n}{n!} e^{-(s_{\rm up}+b)}$$

$$s_{\rm up} = \frac{1}{2} F^{-1}_{\chi^2}(1 - \alpha; 2(n_{\rm obs} + 1)) - b$$

$$= \frac{1}{2} F^{-1}_{\chi^2}(0.95; 2(5 + 1)) - 4.5 = 6.0$$

Frequentist upper limit on Poisson parameter

Upper limit s_{up} at CL = 1 – α found from $p_s = \alpha$.

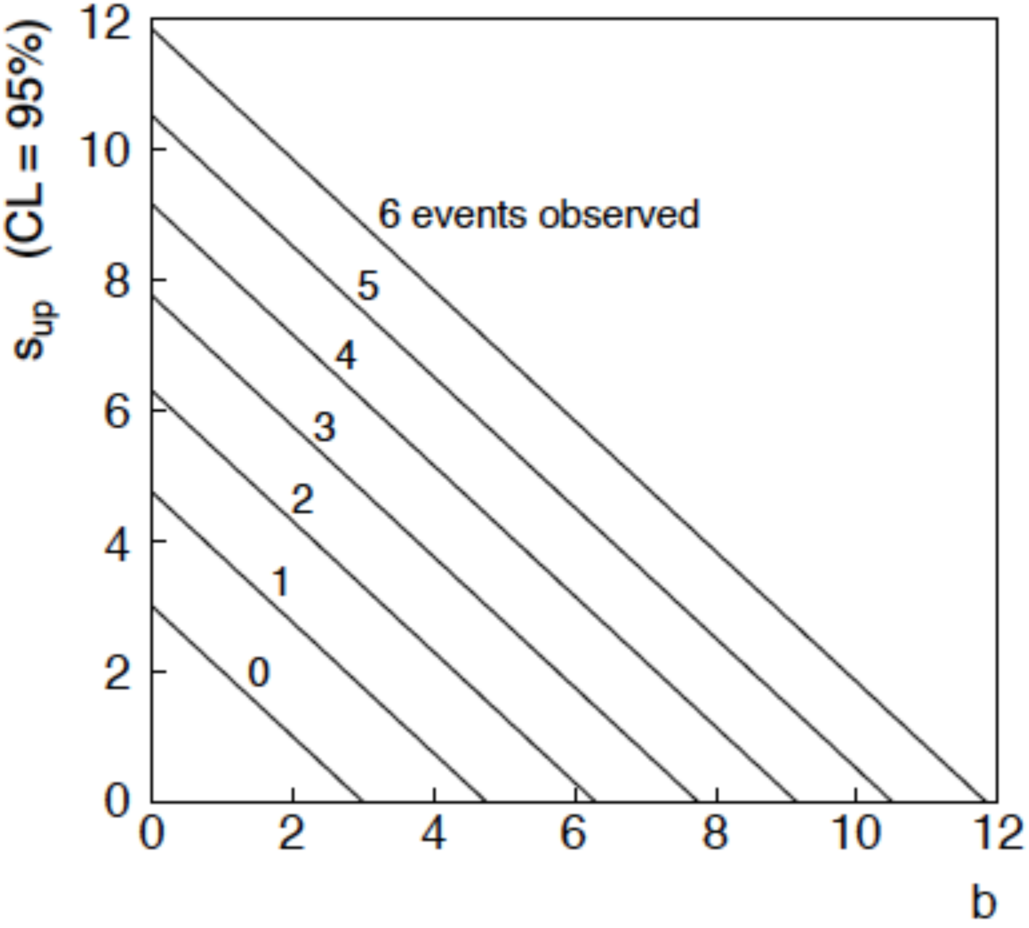


$n_{\mathrm{obs}} = 5$,

$b = 4.5$

$n \sim$ Poisson($s+b$): frequentist upper limit on s

For low fluctuation of n formula can give negative result for s_{up}; i.e. confidence interval is empty.



Limits near a physical boundary

Suppose e.g. $b = 2.5$ and we observe $n = 0$.

If we choose CL = 0.9, we find from the formula for s_{up}

$$s_{\mathrm{up}} = -0.197 \quad (\mathrm{CL} = 0.90)$$

Physicist:

We already knew $s \geq 0$ before we started; can't use negative upper limit to report result of expensive experiment!

Statistician:

The interval is designed to cover the true value only 90% of the time — this was clearly not one of those times.

Not uncommon dilemma when testing parameter values for which one has very little experimental sensitivity, e.g., very small s.

Expected limit for $s = 0$

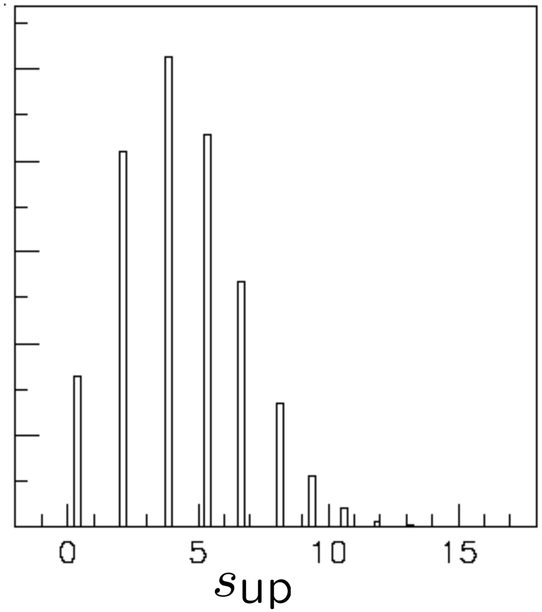
Physicist: I should have used CL = 0.95 — then $s_{up} = 0.496$

Even better: for CL = 0.917923 we get $s_{up} = 10^{-4}$!

Reality check: with $b = 2.5$, typical Poisson fluctuation in n is at least $\sqrt{2.5} = 1.6$. How can the limit be so low?

Look at the mean limit for the no-signal hypothesis ($s = 0$) (sensitivity).

Distribution of 95% CL limits with $b = 2.5$, $s = 0$.
Mean upper limit = 4.44



The Bayesian approach to limits

In Bayesian statistics need to start with 'prior pdf' $\pi(\theta)$, this reflects degree of belief about θ before doing the experiment.

Bayes' theorem tells how our beliefs should be updated in light of the data x:

$$p(\theta|x) = \frac{L(x|\theta)\pi(\theta)}{\int L(x|\theta')\pi(\theta')\,d\theta'} \propto L(x|\theta)\pi(\theta)$$

Integrate posterior pdf $p(\theta|\, x)$ to give interval with any desired probability content.

For e.g. $n \sim$ Poisson($s+b$), 95% CL upper limit on s from

$$0.95 = \int_{-\infty}^{s_{\rm up}} p(s|n)\,ds$$

Bayesian prior for Poisson parameter

Include knowledge that $s \geq 0$ by setting prior $\pi(s) = 0$ for $s < 0$.

Could try to reflect 'prior ignorance' with e.g.

$$\pi(s) = \begin{cases} 1 & s \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Not normalized but this is OK as long as $L(s)$ dies off for large s.

Not invariant under change of parameter — if we had used instead a flat prior for, say, the mass of the Higgs boson, this would imply a non-flat prior for the expected number of Higgs events.

Doesn't really reflect a reasonable degree of belief, but often used as a point of reference;

or viewed as a recipe for producing an interval whose frequentist properties can be studied (coverage will depend on true s).

Bayesian interval with flat prior for *s*

Solve to find limit s_{up}:

$$s_{\mathrm{up}} = \frac{1}{2} F^{-1}_{\chi^2}\left[p, 2(n+1)\right] - b$$

where

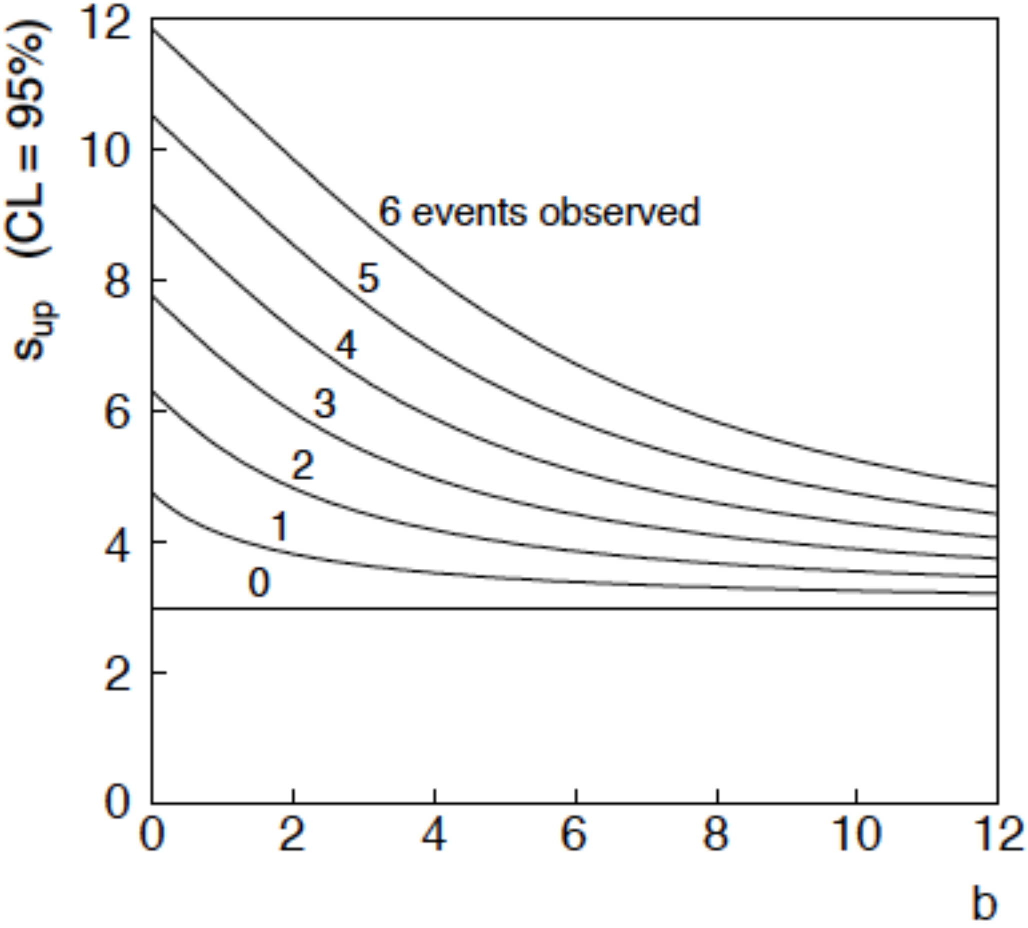
$$p = 1 - \alpha\left(1 - F_{\chi^2}\left[2b, 2(n+1)\right]\right)$$

For special case $b = 0$, Bayesian upper limit with flat prior numerically same as one-sided frequentist case (‘coincidence’).

Bayesian interval with flat prior for *s*

For $b > 0$ Bayesian limit is everywhere greater than the (one sided) frequentist upper limit.

Never goes negative. Doesn't depend on b if $n = 0$.



Priors from formal rules

Because of difficulties in encoding a vague degree of belief in a prior, one often attempts to derive the prior from formal rules, e.g., to satisfy certain invariance principles or to provide maximum information gain for a certain set of measurements.

Often called "objective priors"
Form basis of Objective Bayesian Statistics

The priors do not reflect a degree of belief (but might represent possible extreme cases).

In Objective Bayesian analysis, can use the intervals in a frequentist way, i.e., regard Bayes' theorem as a recipe to produce an interval with certain coverage properties.

Priors from formal rules (cont.)

For a review of priors obtained by formal rules see, e.g.,

Robert E. Kass and Larry Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Am. Stat. Assoc., Vol. 91, No. 435, pp. 1343-1370 (1996).

Formal priors have not been widely used in HEP, but there is recent interest in this direction, especially the reference priors of Bernardo and Berger; see e.g.

L. Demortier, S. Jain and H. Prosper, *Reference priors for high energy physics*, Phys. Rev. D 82 (2010) 034002, arXiv:1002.1111.

D. Casadei, *Reference analysis of the signal + background model in counting experiments*, JINST 7 (2012) 01012; arXiv:1108.4270.

Approximate confidence intervals/regions from the likelihood function

Suppose we test parameter value(s) $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad 0 \leq \lambda(\boldsymbol{\theta}) \leq 1$$

Lower $\lambda(\boldsymbol{\theta})$ means worse agreement between data and hypothesized $\boldsymbol{\theta}$. Equivalently, usually define

$$t_{\boldsymbol{\theta}} = -2 \ln \lambda(\boldsymbol{\theta})$$

so higher $t_{\boldsymbol{\theta}}$ means worse agreement between $\boldsymbol{\theta}$ and the data.

p-value of $\boldsymbol{\theta}$ therefore $$p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$$

need pdf

Confidence region from Wilks' theorem

Wilks' theorem says (in large-sample limit and providing certain conditions hold...)

$$f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta}) \sim \chi^2_n$$

chi-square dist. with # d.o.f. =
# of components in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

Assuming this holds, the *p*-value is

$$p_{\boldsymbol{\theta}} = 1 - F_{\chi^2_n}(t_{\boldsymbol{\theta}})$$

To find boundary of confidence region set $p_{\boldsymbol{\theta}} = \alpha$ and solve for $t_{\boldsymbol{\theta}}$:

$$t_{\boldsymbol{\theta}} = -2 \ln \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} = F^{-1}_{\chi^2_n}(1 - \alpha)$$

Confidence region from Wilks' theorem (cont.)

i.e., boundary of confidence region in $\boldsymbol{\theta}$ space is where

$$\ln L(\boldsymbol{\theta}) = \ln L(\hat{\boldsymbol{\theta}}) - \tfrac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

For example, for $1 - \alpha = 68.3\%$ and $n = 1$ parameter,

$$F^{-1}_{\chi^2_1}(0.683) = 1$$

and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

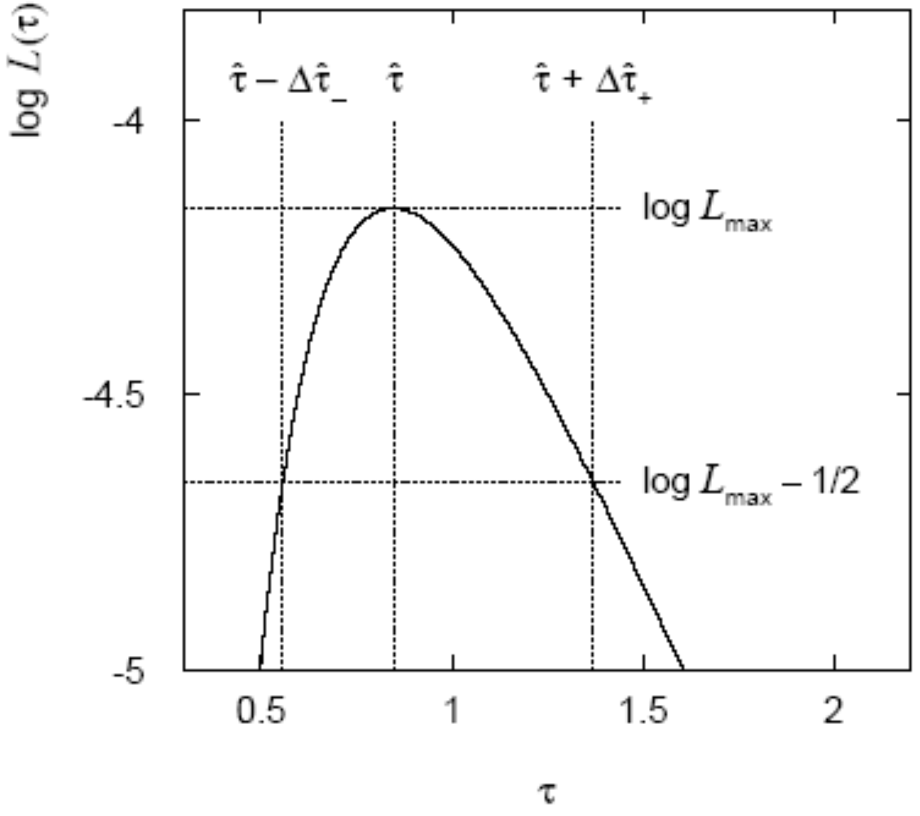
Same as recipe for finding the estimator's standard deviation, i.e.,

$[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Example of interval from ln L

For $n = 1$ parameter, CL = 0.683, $Q_\alpha = 1$.

Exponential example, now with only 5 events:



Parameter estimate and approximate 68.3% CL confidence interval:

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

Multiparameter case

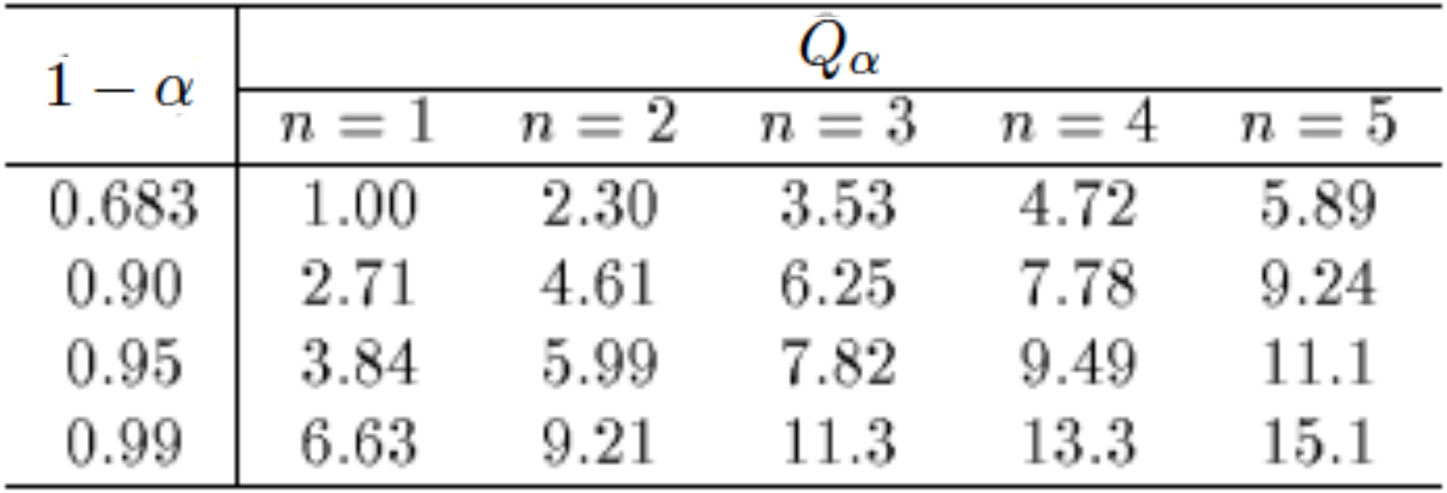
For increasing number of parameters, CL = 1 – *α* decreases for confidence region determined by a given

$$Q_\alpha = F^{-1}_{\chi^2_n}(1-\alpha)$$

Q_α	$1-\alpha$				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
1.0	0.683	0.393	0.199	0.090	0.037
2.0	0.843	0.632	0.428	0.264	0.151
4.0	0.954	0.865	0.739	0.594	0.451
9.0	0.997	0.989	0.971	0.939	0.891

Multiparameter case (cont.)

Equivalently, Q_α increases with n for a given CL = 1 – α.



$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

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

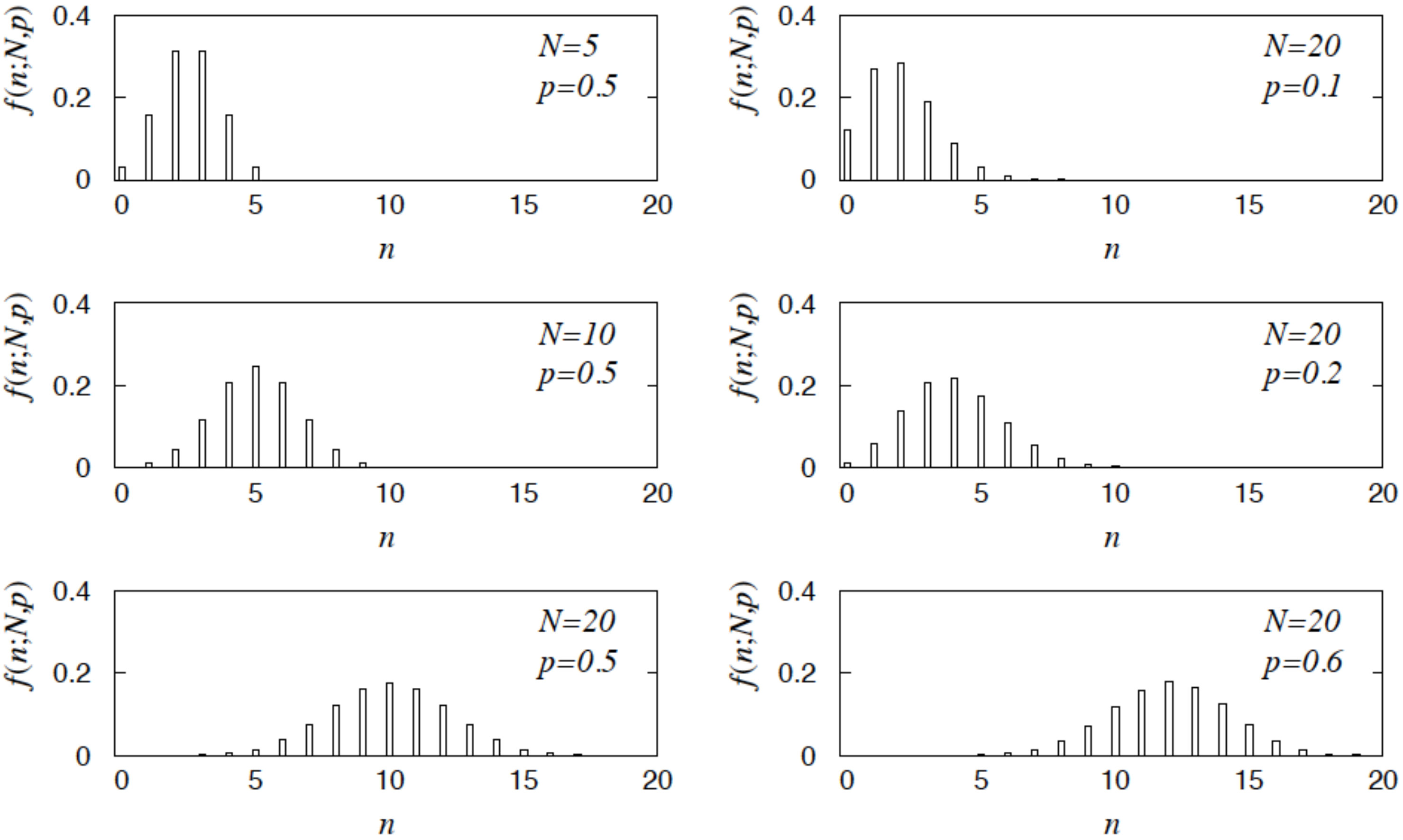
For the expectation value and variance we find:

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

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

Binomial distribution (3)

Binomial distribution for several values of the parameters:



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

Multinomial distribution

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

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

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

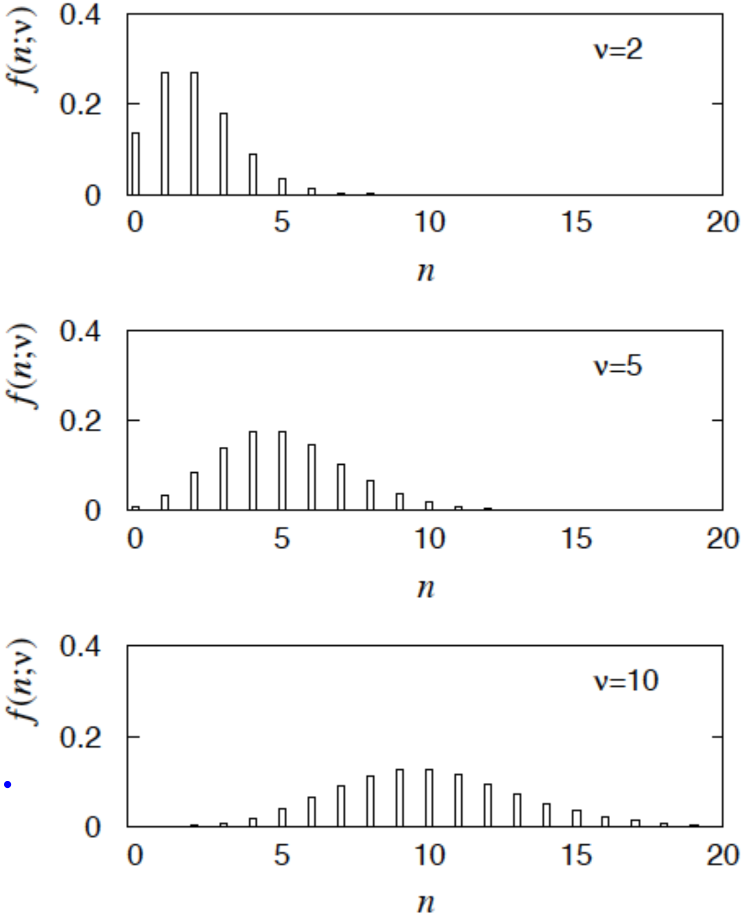
Consider binomial *n* in the limit

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

→ *n* follows the Poisson distribution:

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

$$E[n] = \nu \, , \qquad V[n] = \nu \; .$$



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

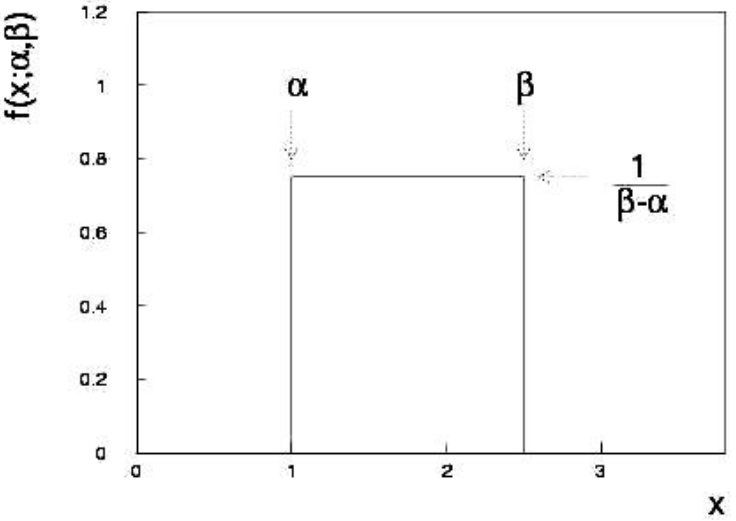
Uniform distribution

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 $[E_{\min}, E_{\max}]$, with

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

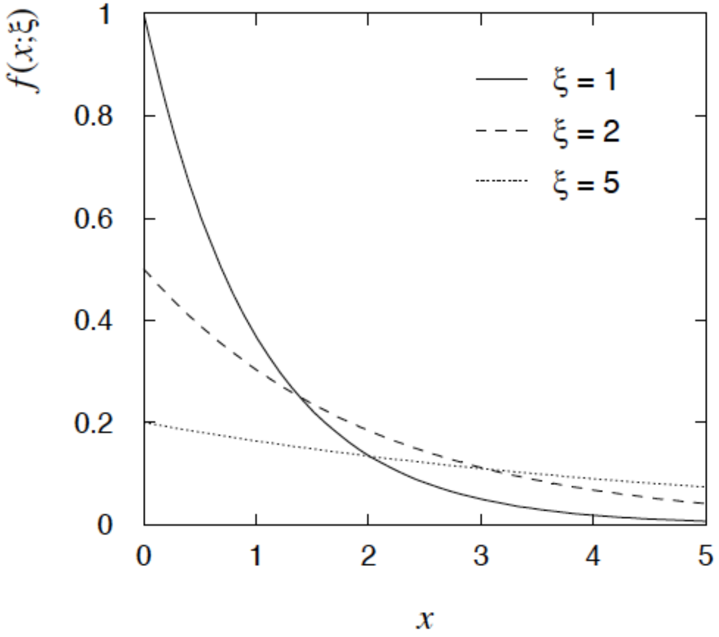
Exponential distribution

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

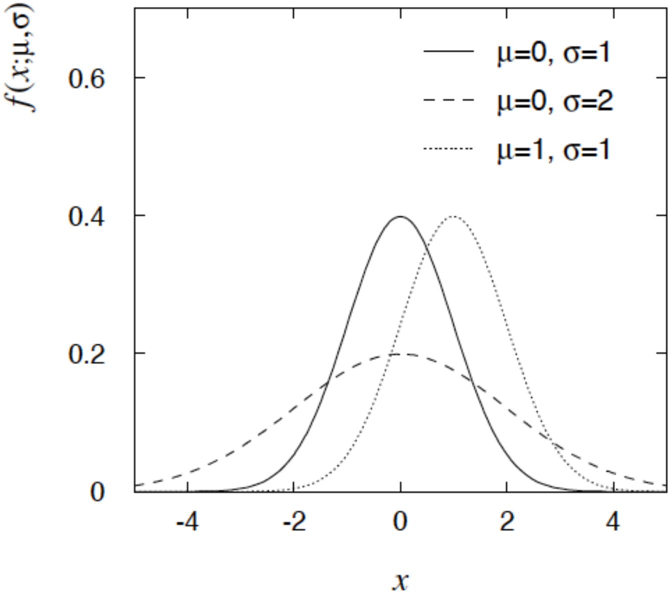
The Gaussian (normal) pdf for a continuous r.v. *x* is defined by:

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

$$E[x] = \mu$$

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

(N.B. often μ, σ^2 denote mean, variance of any r.v., not only Gaussian.)



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

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

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,

$$E[x_i] = \mu_i, \, , \qquad \mathrm{cov}[x_i, x_j] = V_{ij} \; .$$

For $n = 2$ this is

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

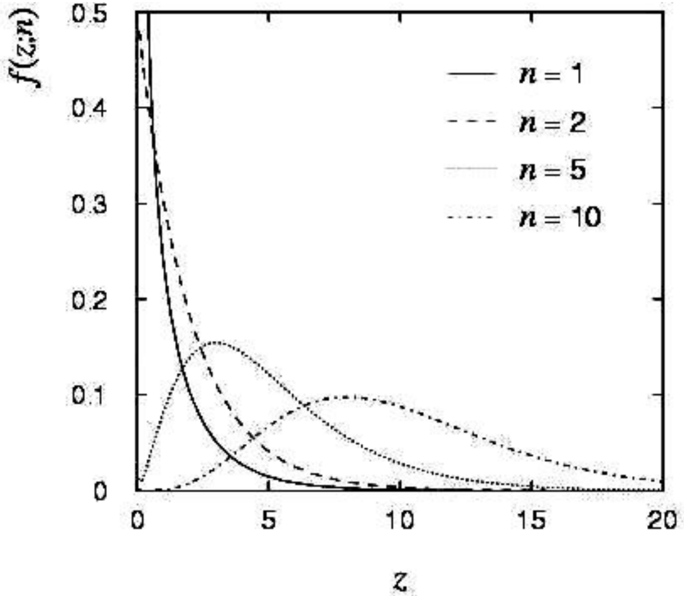
$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ 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, ... = number of 'degrees of freedom' (dof)

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

For independent Gaussian x_i, i = 1, ..., 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.

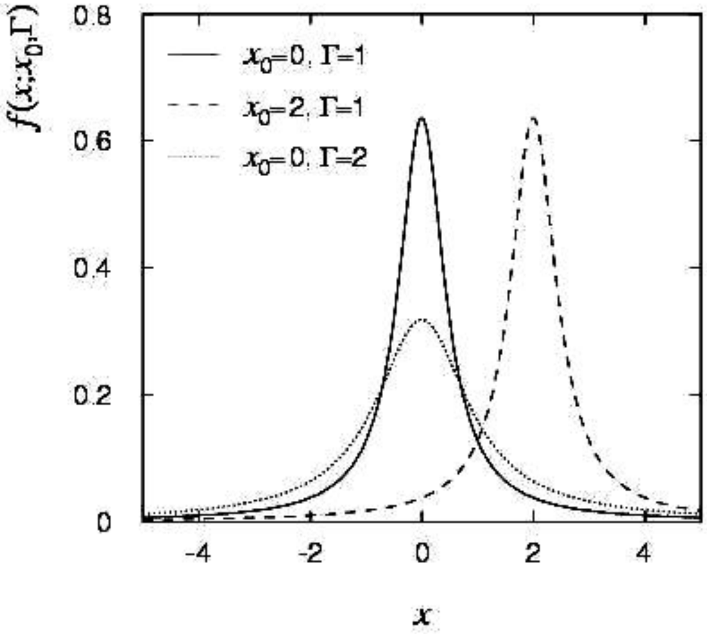
Example: goodness-of-fit test variable especially in conjunction with method of least squares.

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, $V[x] \rightarrow \infty$.

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

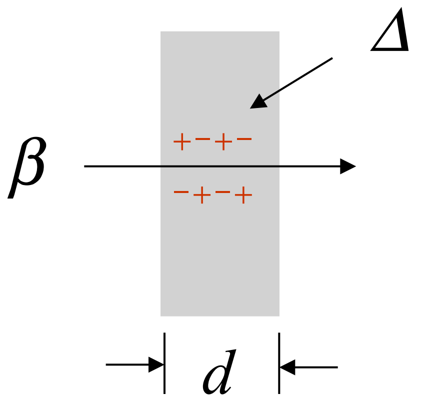
For a charged particle with $\beta = v/c$ traversing a layer of matter of thickness d, the energy loss Δ follows the Landau pdf:

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

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

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

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

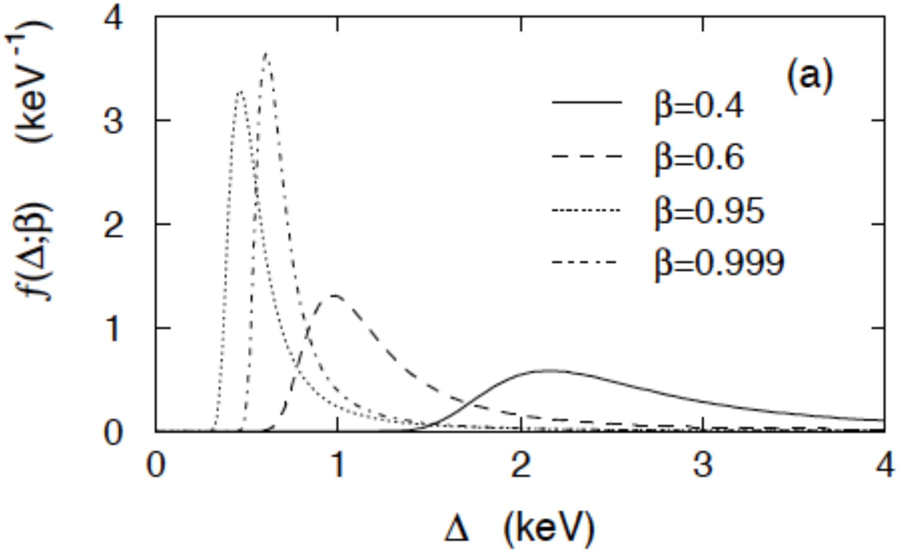


L. Landau, J. Phys. USSR **8** (1944) 201; see also
W. Allison and J. Cobb, Ann. Rev. Nucl. Part. Sci. **30** (1980) 253.

Landau distribution (2)

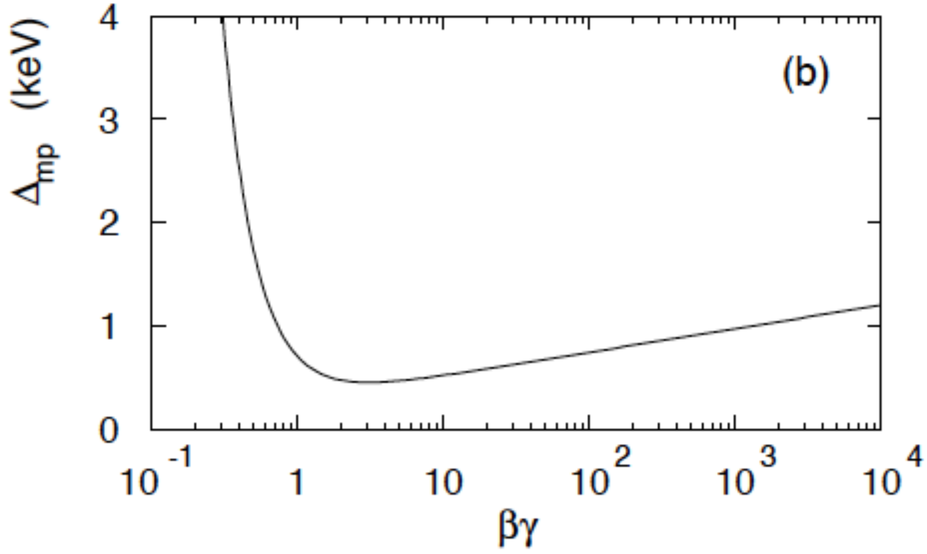
Long 'Landau tail'

→ all moments ∞



Mode (most probable value) sensitive to β,

→ particle i.d.

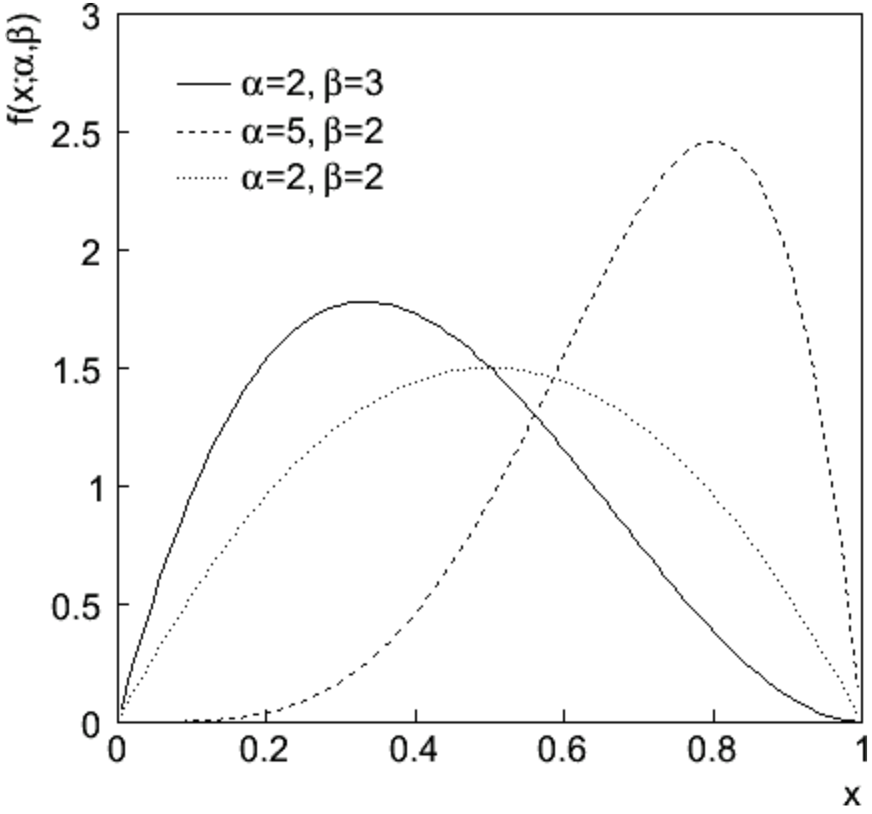


Beta distribution

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

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

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



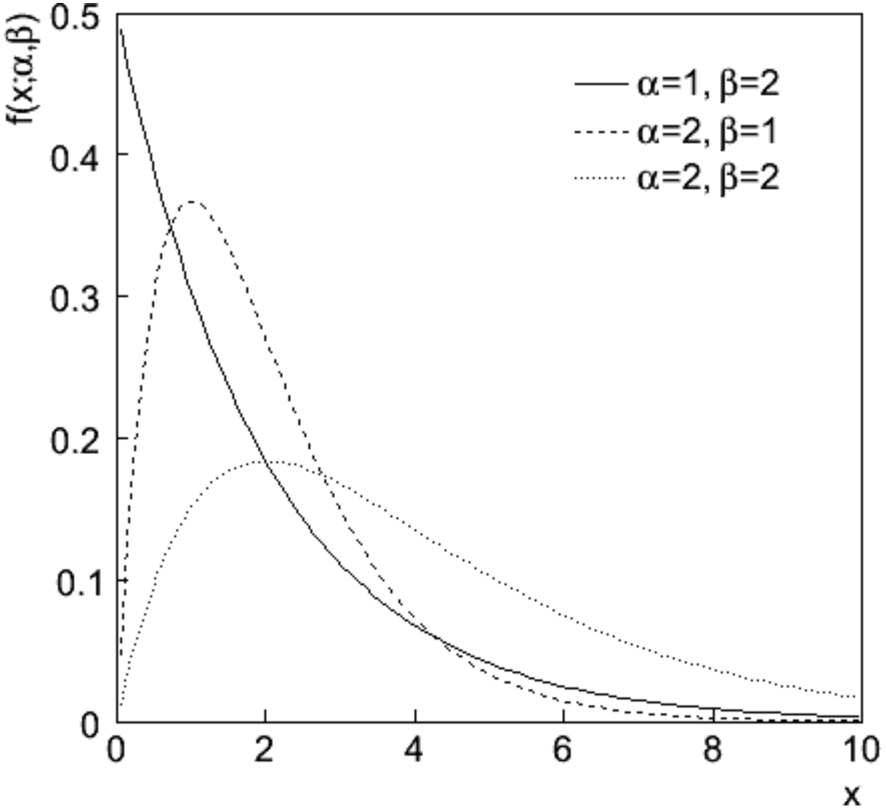
Often used to represent pdf of continuous r.v. nonzero only between finite limits.

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,∞].

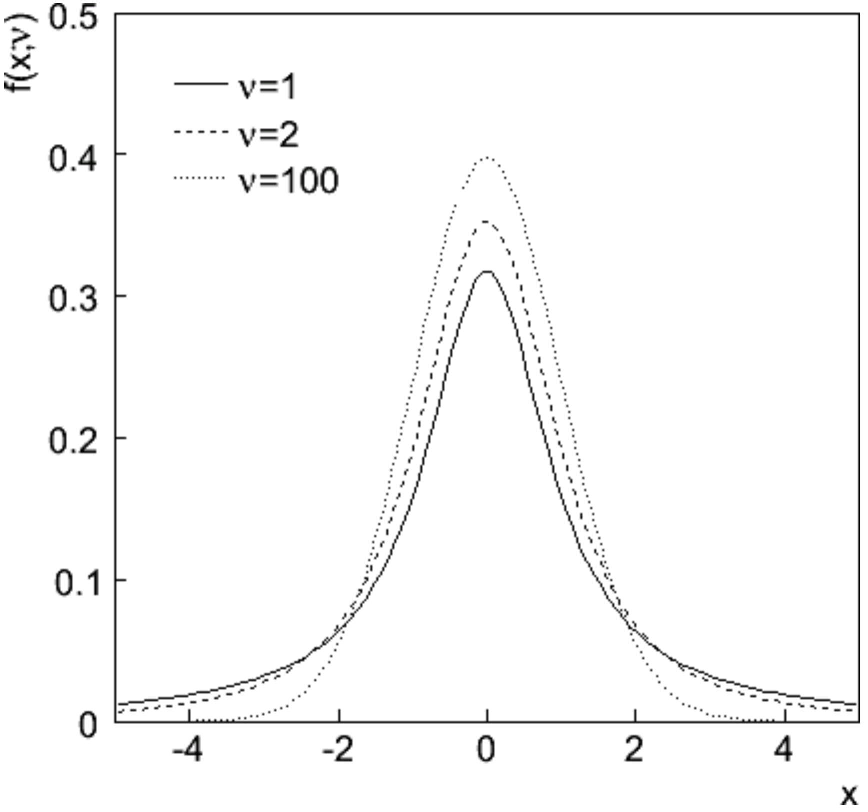
Also e.g. sum of *n* exponential r.v.s or time until *n*th event in Poisson process ~ Gamma

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)

ν = 1 gives Cauchy,

$\nu \rightarrow \infty$ gives Gaussian.

Student's t distribution (2)

If $x \sim$ Gaussian with $\mu = 0$, $\sigma^2 = 1$, and
$z \sim \chi^2$ with n degrees of freedom, then
$t = x \,/\, (z/n)^{1/2}$ follows Student's t with $\nu = n$.

This arises in problems where one forms the ratio of a sample mean to the sample standard deviation of Gaussian r.v.s.

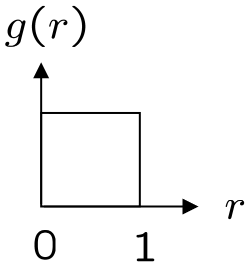
The Student's t provides a bell-shaped pdf with adjustable tails, ranging from those of a Gaussian, which fall off very quickly, ($\nu \rightarrow \infty$, but in fact already very Gauss-like for $\nu =$ two dozen), to the very long-tailed Cauchy ($\nu = 1$).

Developed in 1908 by William Gosset, who worked under the pseudonym "Student" for the Guinness Brewery.

The Monte Carlo method

What it is: a numerical technique for calculating probabilities and related quantities using sequences of random numbers.

The usual steps:



(1) Generate sequence $r_1, r_2, \ldots, r_m$ uniform in [0, 1].

(2) Use this to produce another sequence $x_1, x_2, \ldots, x_n$ distributed according to some pdf $f(x)$ in which we're interested (x can be a vector).

(3) Use the x values to estimate some property of $f(x)$, e.g., fraction of x values with $a < x < b$ gives $\int_a^b f(x)\,dx$.

→ MC calculation = integration (at least formally)

MC generated values = 'simulated data'

→ use for testing statistical procedures

Random number generators

Goal: generate uniformly distributed values in [0, 1].

Toss coin for e.g. 32 bit number... (too tiring).

→ 'random number generator'

= computer algorithm to generate $r_1, r_2, \ldots, r_n$.

Example: multiplicative linear congruential generator (MLCG)

$n_{i+1} = (a\, n_i) \bmod m$,   where

n_i = integer

a = multiplier

m = modulus

n_0 = seed (initial value)

N.B. mod = modulus (remainder), e.g. 27 mod 5 = 2.

This rule produces a sequence of numbers $n_0, n_1, \ldots$

Random number generators (2)

The sequence is (unfortunately) periodic!

Example (see Brandt Ch 4): $a = 3$, $m = 7$, $n_0 = 1$

$$
\begin{aligned}
n_1 &= (3 \cdot 1)\mathrm{mod}7 = 3 \\
n_2 &= (3 \cdot 3)\mathrm{mod}7 = 2 \\
n_3 &= (3 \cdot 2)\mathrm{mod}7 = 6 \\
n_4 &= (3 \cdot 6)\mathrm{mod}7 = 4 \\
n_5 &= (3 \cdot 4)\mathrm{mod}7 = 5 \\
n_6 &= (3 \cdot 5)\mathrm{mod}7 = 1
\end{aligned}
$$

← sequence repeats

Choose a, m to obtain long period (maximum = $m - 1$); m usually close to the largest integer that can represented in the computer.

Only use a subset of a single period of the sequence.

Random number generators (3)

$r_i = n_i/m$ are in [0, 1] but are they ‘random’?

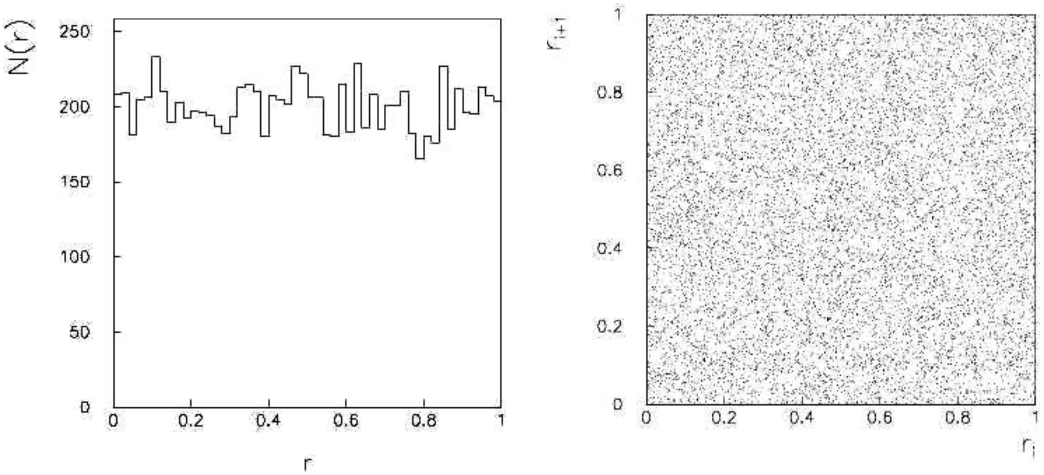
Choose a, m so that the r_i pass various tests of randomness:

uniform distribution in [0, 1],

all values independent (no correlations between pairs),

e.g. L’Ecuyer, Commun. ACM **31** (1988) 742 suggests

$a = 40692$
$m = 2147483399$

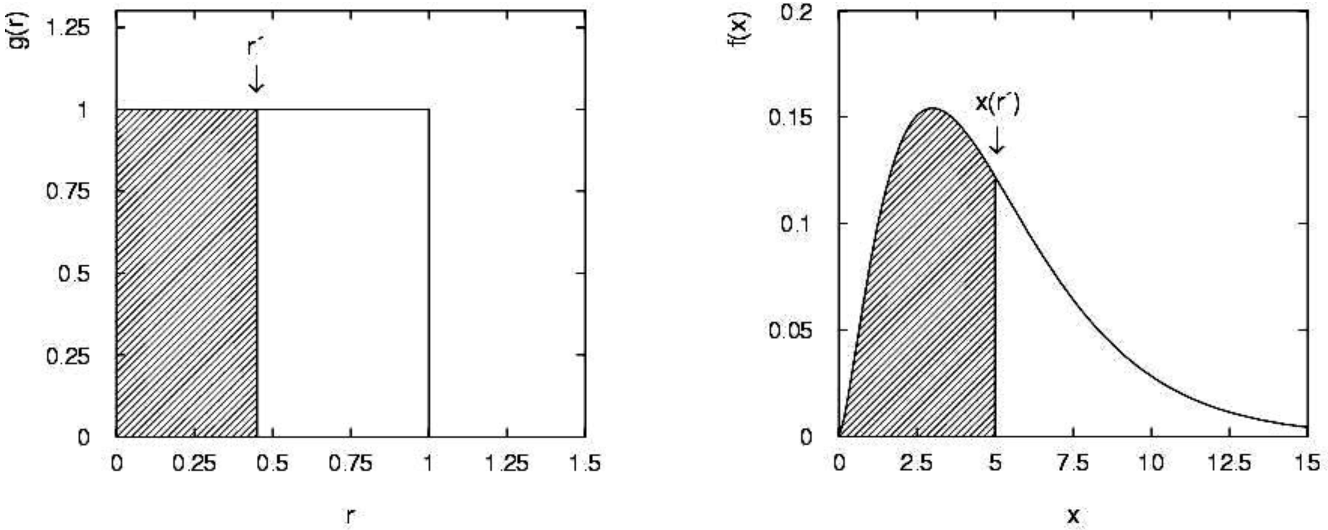


Far better generators available, e.g. `TRandom3`, based on Mersenne twister algorithm, period = $2^{19937} - 1$ (a “Mersenne prime”).

See F. James, Comp. Phys. Comm. 60 (1990) 111; Brandt Ch. 4

The transformation method

Given $r_1, r_2, \ldots, r_n$ uniform in [0, 1], find $x_1, x_2, \ldots, x_n$ that follow $f(x)$ by finding a suitable transformation $x(r)$.



Require: $P(r \leq r') = P(x \leq x(r'))$

i.e. $$\int_{-\infty}^{r'} g(r)\,dr = r' = \int_{-\infty}^{x(r')} f(x')\,dx' = F(x(r'))$$

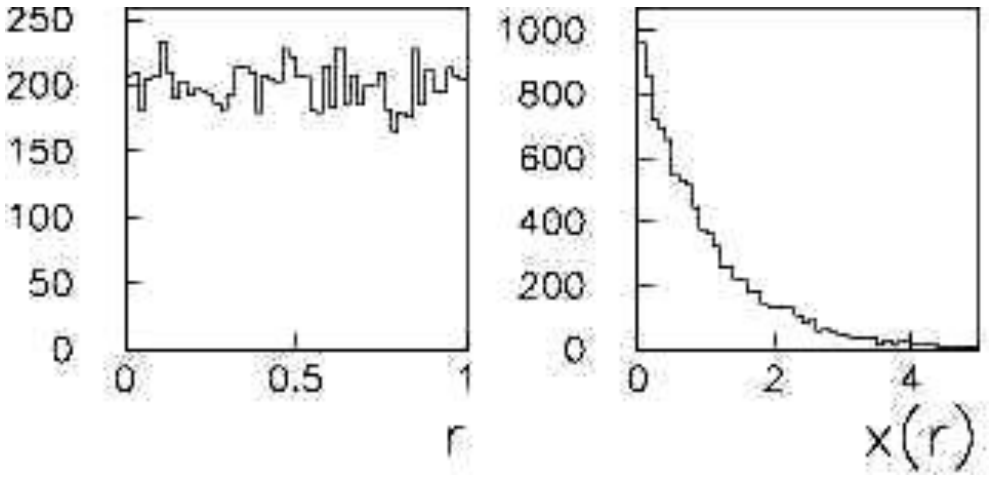
That is, set $F(x) = r$ and solve for $x(r)$.

Example of the transformation method

Exponential pdf: $f(x;\xi) = \frac{1}{\xi} e^{-x/\xi} \quad (x \geq 0)$

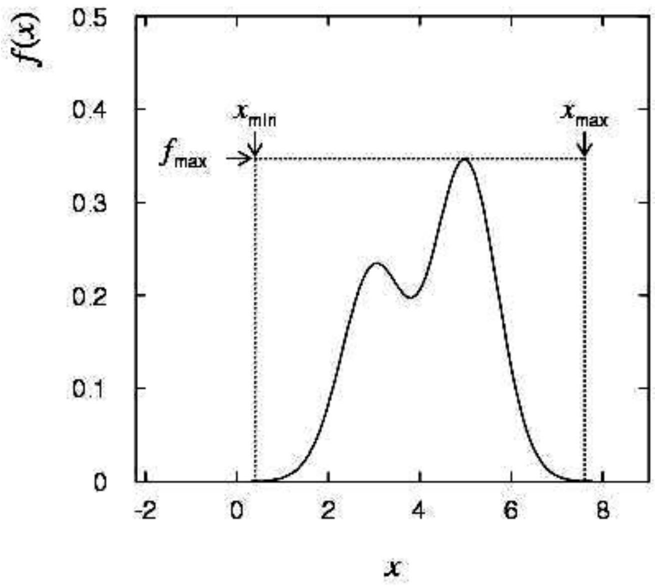
Set $\int_0^x \frac{1}{\xi} e^{-x'/\xi}\,dx' = r$ and solve for $x(r)$.

→ $x(r) = -\xi \ln(1-r)$ ($x(r) = -\xi \ln r$ works too.)



The acceptance-rejection method

Enclose the pdf in a box:



(1) Generate a random number x, uniform in $[x_{\min}, x_{\max}]$, i.e. $x = x_{\min} + r_1(x_{\max} - x_{\min})$, r_1 is uniform in [0,1].

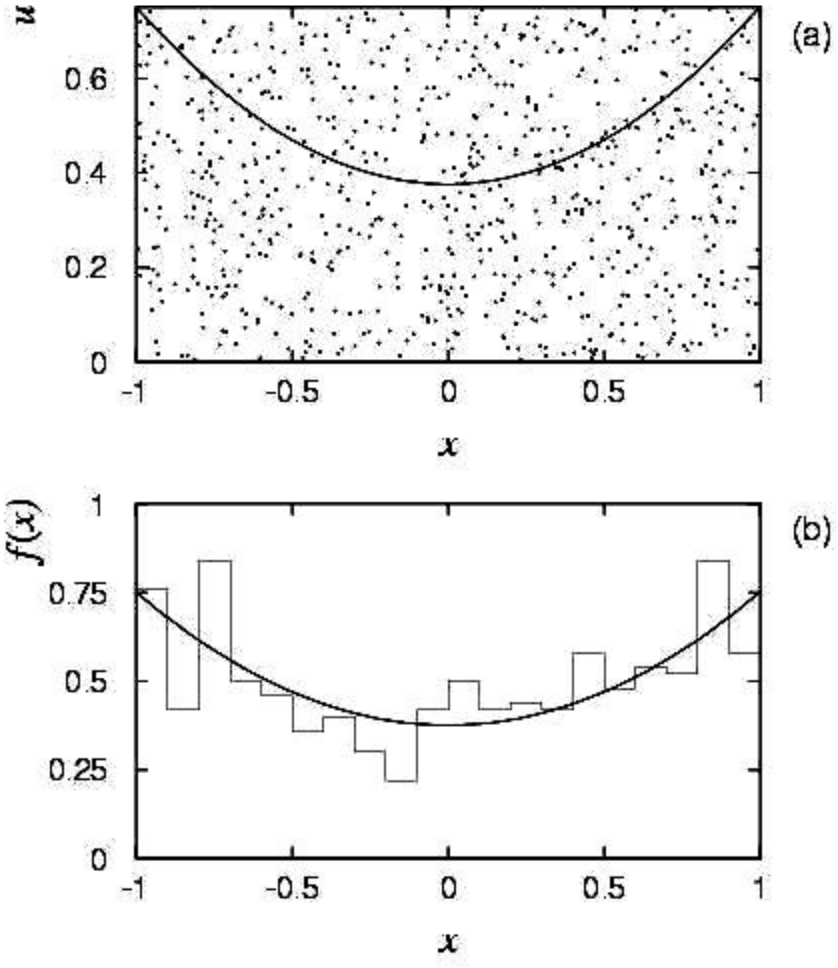
(2) Generate a 2nd independent random number u uniformly distributed between 0 and $f_{\max}$, i.e. $u = r_2 f_{\max}$.

(3) If $u < f(x)$, then accept x. If not, reject x and repeat.

Example with acceptance-rejection method

$$f(x) = \frac{3}{8}(1 + x^2)$$

$$(-1 \leq x \leq 1)$$



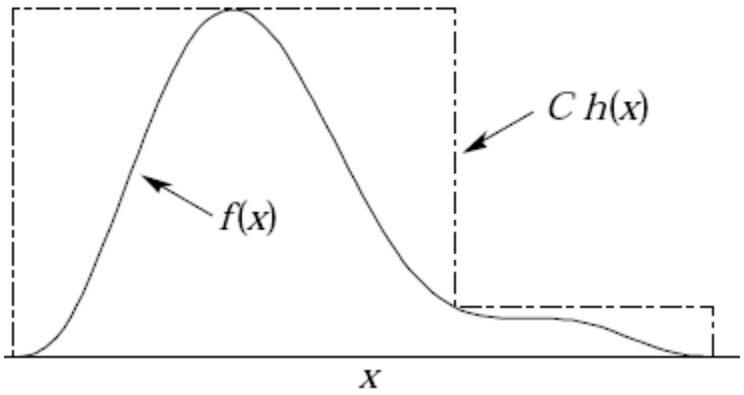
If dot below curve, use *x* value in histogram.

Improving efficiency of the acceptance-rejection method

The fraction of accepted points is equal to the fraction of the box's area under the curve.

For very peaked distributions, this may be very low and thus the algorithm may be slow.

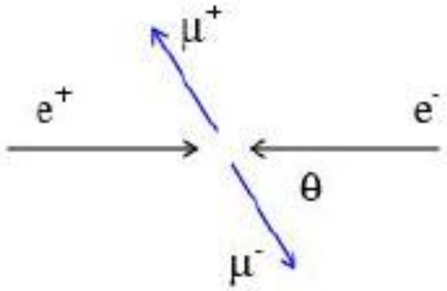
Improve by enclosing the pdf $f(x)$ in a curve $C\,h(x)$ that conforms to $f(x)$ more closely, where $h(x)$ is a pdf from which we can generate random values and C is a constant.



Generate points uniformly over $C\,h(x)$.

If point is below $f(x)$, accept x.

Monte Carlo event generators



Simple example: $e^+e^- \rightarrow \mu^+\mu^-$

Generate $\cos\theta$ and ϕ:

$$f(\cos\theta; A_{\mathrm{FB}}) \propto \left(1 + \frac{8}{3} A_{\mathrm{FB}} \cos\theta + \cos^2\theta\right) ,$$

$$g(\phi) = \frac{1}{2\pi} \quad (0 \leq \phi \leq 2\pi)$$

Less simple: ‘event generators’ for a variety of reactions:

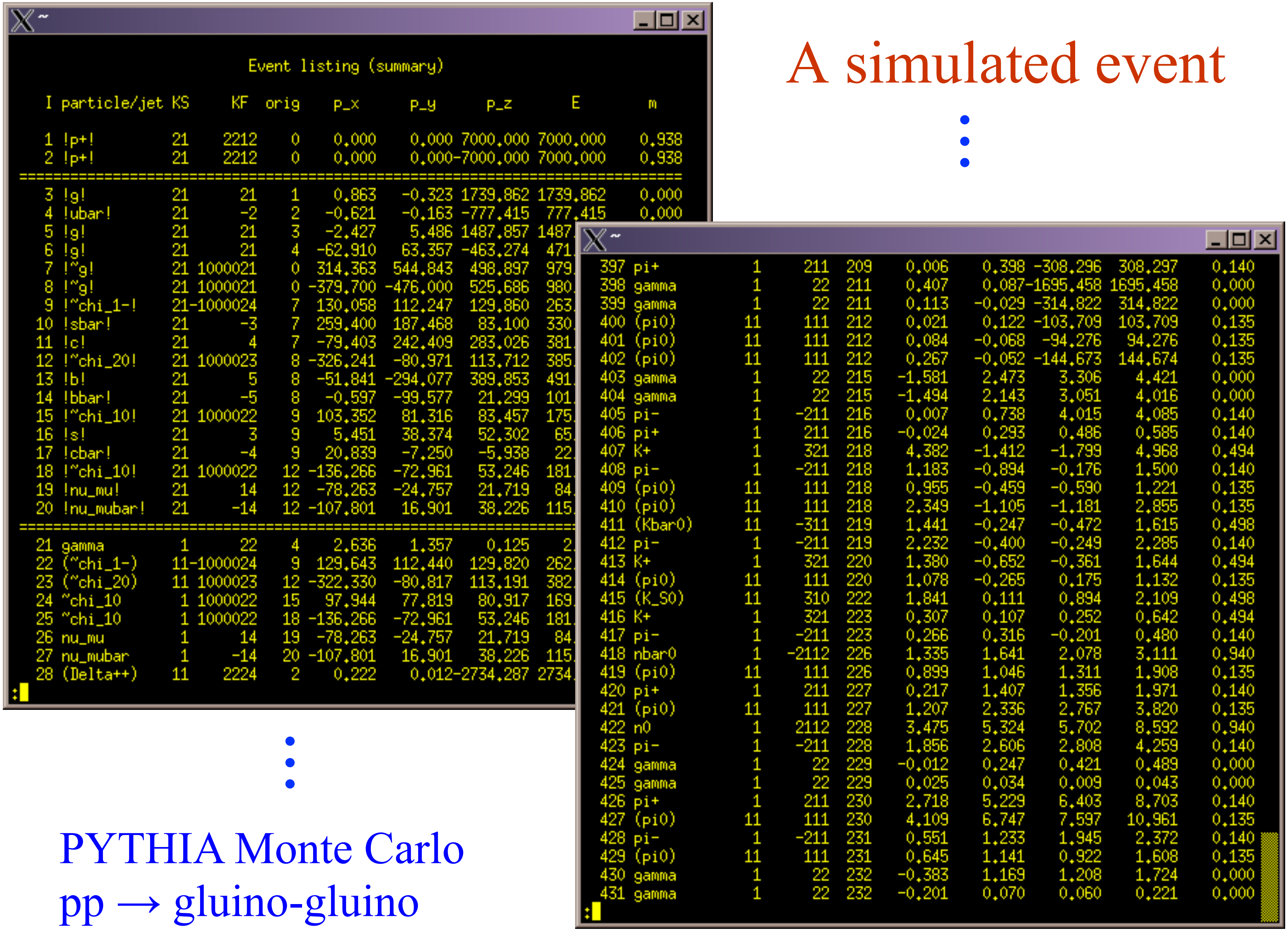
$e^+e^- \rightarrow m^+m^-$, hadrons, ...
pp → hadrons, D-Y, SUSY,...

e.g. PYTHIA, HERWIG, ISAJET...

Output = ‘events’, i.e., for each event we get a list of generated particles and their momentum vectors, types, etc.

A simulated event

⋮



```
                      Event listing (summary)

    I particle/jet KS     KF  orig    p_x      p_y      p_z       E          m

    1 !p+!         21    2212    0    0.000    0.000 7000.000 7000.000     0.938
    2 !p+!         21    2212    0    0.000    0.000-7000.000 7000.000     0.938
 ==============================================================================
    3 !g!          21      21    1    0.863   -0.323 1739.862 1739.862     0.000
    4 !ubar!       21      -2    2   -0.621   -0.163 -777.415  777.415     0.000
    5 !g!          21      21    3   -2.427    5.486 1487.857 1487.
    6 !g!          21      21    4  -62.910   63.357 -463.274  471.
    7 !~g!         21 1000021    0  314.363  544.843  498.897  979.
    8 !~g!         21 1000021    0 -379.700 -476.000  525.686  980.
    9 !~chi_1-!    21-1000024    7  130.058  112.247  129.860  263.
   10 !sbar!       21      -3    7  259.400  187.468   83.100  330.
   11 !c!          21       4    7  -79.403  242.409  283.026  381.
   12 !~chi_20!    21 1000023    8 -326.241  -80.971  113.712  385.
   13 !b!          21       5    8  -51.841 -294.077  389.853  491.
   14 !bbar!       21      -5    8   -0.597  -99.577   21.299  101.
   15 !~chi_10!    21 1000022    9  103.352   81.316   83.457  175.
   16 !s!          21       3    9    5.451   38.374   52.302   65.
   17 !cbar!       21      -4    9   20.839   -7.250   -5.938   22.
   18 !~chi_10!    21 1000022   12 -136.266  -72.961   53.246  181.
   19 !nu_mu!      21      14   12  -78.263  -24.757   21.719   84.
   20 !nu_mubar!   21     -14   12 -107.801   16.901   38.226  115.
 =============================================================
   21 gamma         1      22    4    2.636    1.357    0.125    2.
   22 (~chi_1-)    11-1000024    9  129.643  112.440  129.820  262.
   23 (~chi_20)    11 1000023   12 -322.330  -80.817  113.191  382.
   24 ~chi_10       1 1000022   15   97.944   77.819   80.917  169.
   25 ~chi_10       1 1000022   18 -136.266  -72.961   53.246  181.
   26 nu_mu         1      14   19  -78.263  -24.757   21.719   84.
   27 nu_mubar      1     -14   20 -107.801   16.901   38.226  115.
   28 (Delta++)    11    2224    2    0.222    0.012-2734.287 2734.
:
```

```
  397 pi+          1     211  209    0.006    0.398 -308.296  308.297     0.140
  398 gamma        1      22  211    0.407    0.087-1695.458 1695.458     0.000
  399 gamma        1      22  211    0.113   -0.029 -314.822  314.822     0.000
  400 (pi0)       11     111  212    0.021    0.122 -103.709  103.709     0.135
  401 (pi0)       11     111  212    0.084   -0.068  -94.276   94.276     0.135
  402 (pi0)       11     111  212    0.267   -0.052 -144.673  144.674     0.135
  403 gamma        1      22  215   -1.581    2.473    3.306    4.421     0.000
  404 gamma        1      22  215   -1.494    2.143    3.051    4.016     0.000
  405 pi-          1    -211  216    0.007    0.738    4.015    4.085     0.140
  406 pi+          1     211  216   -0.024    0.293    0.486    0.585     0.140
  407 K+           1     321  218    4.382   -1.412   -1.799    4.968     0.494
  408 pi-          1    -211  218    1.183   -0.894   -0.176    1.500     0.140
  409 (pi0)       11     111  218    0.955   -0.459   -0.590    1.221     0.135
  410 (pi0)       11     111  218    2.349   -1.105   -1.181    2.855     0.135
  411 (Kbar0)     11    -311  219    1.441   -0.247   -0.472    1.615     0.498
  412 pi-          1    -211  219    2.232   -0.400   -0.249    2.285     0.140
  413 K+           1     321  220    1.380   -0.652   -0.361    1.644     0.494
  414 (pi0)       11     111  220    1.078   -0.265    0.175    1.132     0.135
  415 (K_S0)      11     310  222    1.841    0.111    0.894    2.109     0.498
  416 K+           1     321  223    0.307    0.107    0.252    0.642     0.494
  417 pi-          1    -211  223    0.266    0.316   -0.201    0.480     0.140
  418 nbar0        1   -2112  226    1.335    1.641    2.078    3.111     0.940
  419 (pi0)       11     111  226    0.899    1.046    1.311    1.908     0.135
  420 pi+          1     211  227    0.217    1.407    1.356    1.971     0.140
  421 (pi0)       11     111  227    1.207    2.336    2.767    3.820     0.135
  422 n0           1    2112  228    3.475    5.324    5.702    8.592     0.940
  423 pi-          1    -211  228    1.856    2.606    2.808    4.259     0.140
  424 gamma        1      22  229   -0.012    0.247    0.421    0.489     0.000
  425 gamma        1      22  229    0.025    0.034    0.009    0.043     0.000
  426 pi+          1     211  230    2.718    5.229    6.403    8.703     0.140
  427 (pi0)       11     111  230    4.109    6.747    7.597   10.961     0.135
  428 pi-          1    -211  231    0.551    1.233    1.945    2.372     0.140
  429 (pi0)       11     111  231    0.645    1.141    0.922    1.608     0.135
  430 gamma        1      22  232   -0.383    1.169    1.208    1.724     0.000
  431 gamma        1      22  232   -0.201    0.070    0.060    0.221     0.000
:
```

⋮

PYTHIA Monte Carlo
pp → gluino-gluino

Monte Carlo detector simulation

Takes as input the particle list and momenta from generator.

Simulates detector response:

multiple Coulomb scattering (generate scattering angle),
particle decays (generate lifetime),
ionization energy loss (generate Δ),
electromagnetic, hadronic showers,
production of signals, electronics response, ...

Output = simulated raw data → input to reconstruction software:

track finding, fitting, etc.

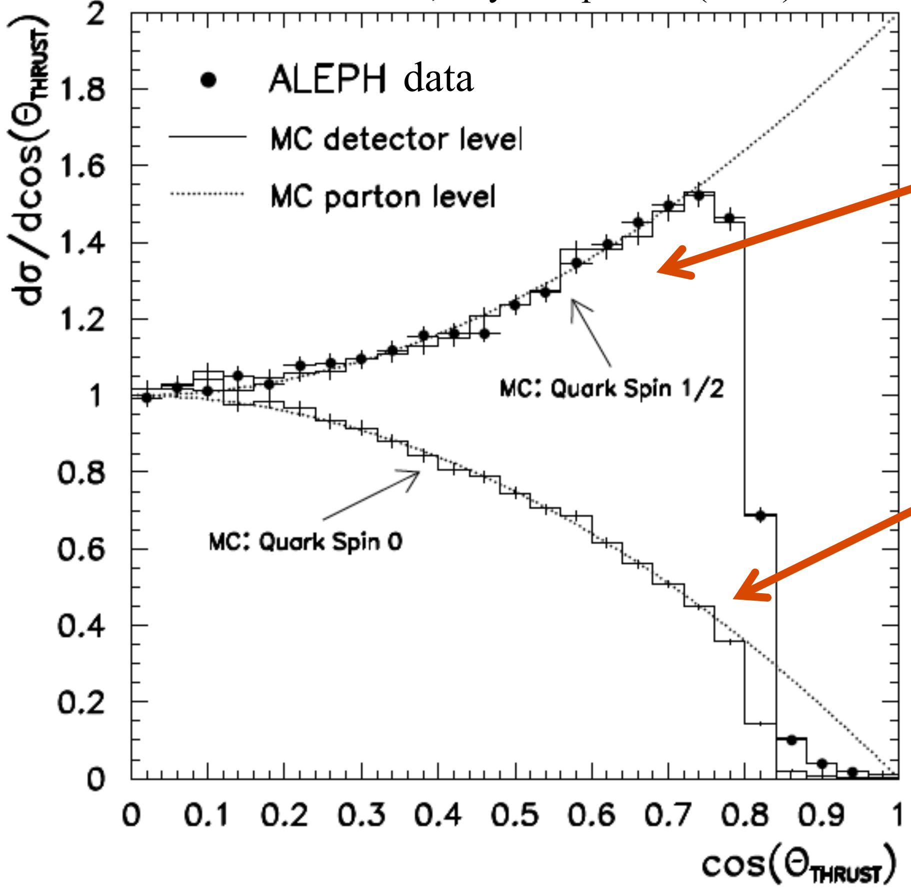
Predict what you should see at ‘detector level’ given a certain hypothesis for ‘generator level’. Compare with the real data.

Estimate ‘efficiencies’ = #events found / # events generated.

Programming package: **`GEANT`**

Data analysis in particle physics: testing hypotheses

Test the extent to which a given model agrees with the data:



ALEPH, Phys. Rept. 294 (1998) 1-165

spin-1/2 quark
model "good"

spin-0 quark
model "bad"

In general need tests
with well-defined properties
and quantitative results.

Choosing a critical region

To construct a test of a hypothesis H_0, we can ask what are the relevant alternatives for which one would like to have a high power.

Maximize power wrt H_1 = maximize probability to reject H_0 if H_1 is true.

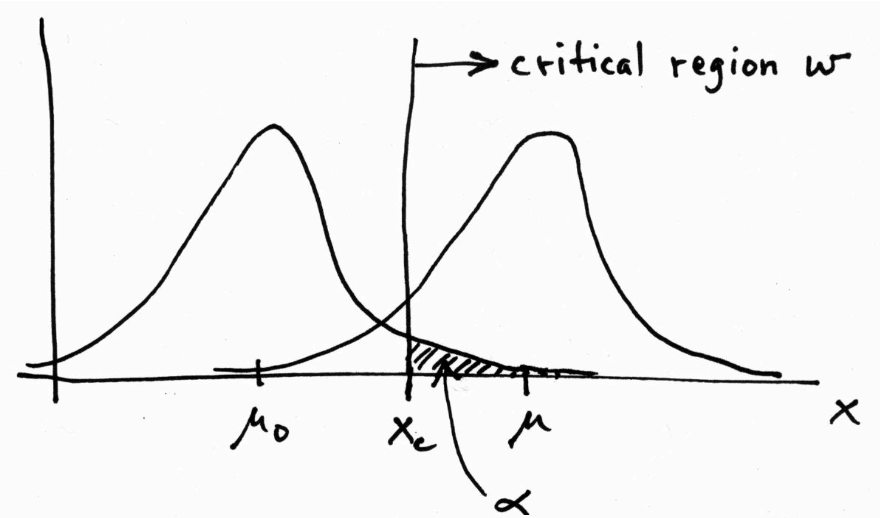
Often such a test has a high power not only with respect to a specific point alternative but for a class of alternatives. E.g., using a measurement $x \sim$ Gauss (μ, σ) we may test

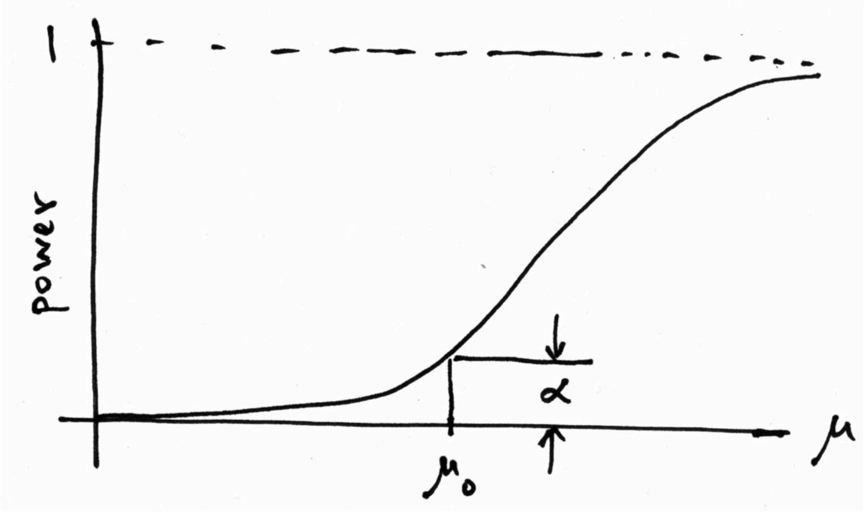
$H_0 : \mu = \mu_0$ versus the composite alternative $H_1 : \mu > \mu_0$

We get the highest power with respect to any $\mu > \mu_0$ by taking the critical region $x \geq x_\mathrm{c}$ where the cut-off x_c is determined by the significance level such that

$$\alpha = P(x \geq x_\mathrm{c} | \mu_0).$$

Test of $\mu = \mu_0$ vs. $\mu > \mu_0$ with $x \sim$ Gauss(μ,σ)





Standard Gaussian cumulative distribution

$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$

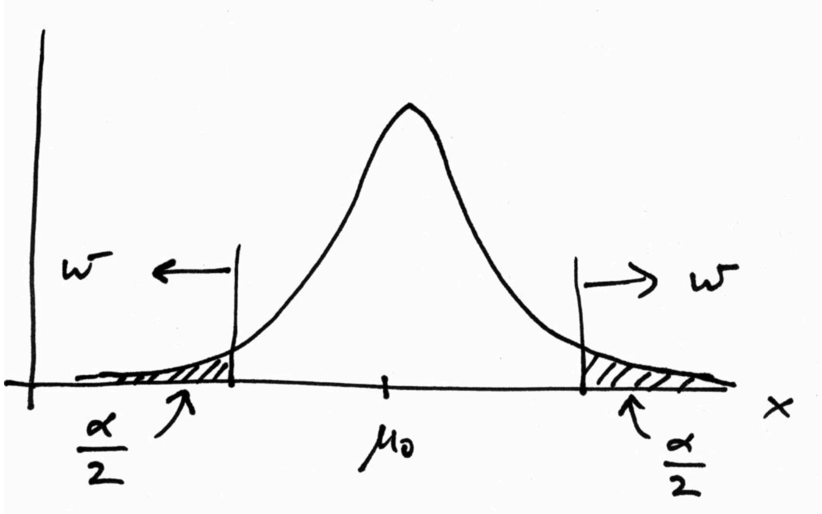
$$x_c = \mu_0 + \sigma\Phi^{-1}(1-\alpha)$$

Standard Gaussian quantile

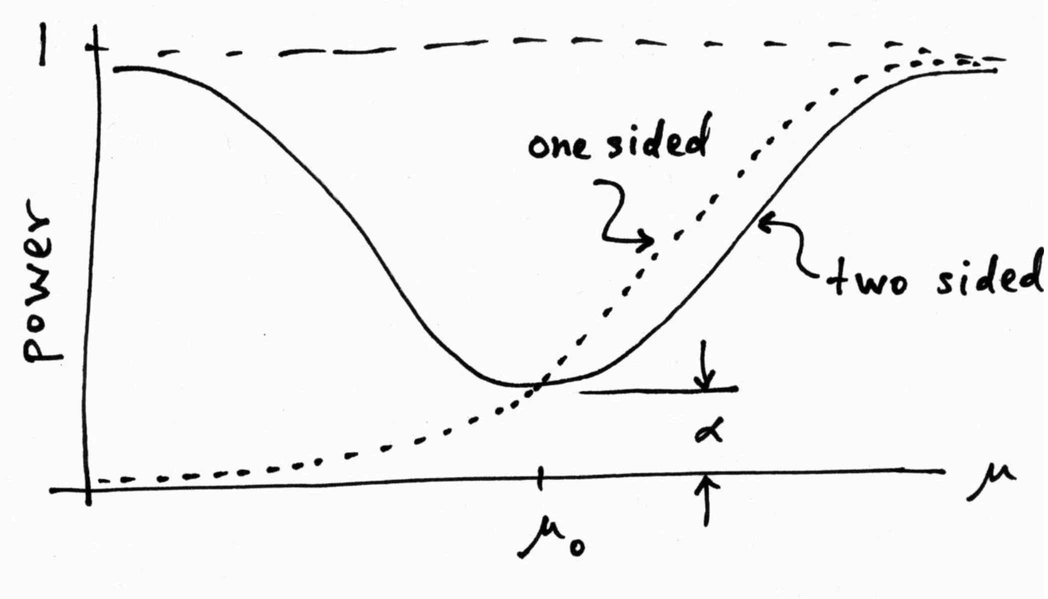
$$\text{power} = 1 - \beta = P(x > x_c|\mu) =$$

$$1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma} + \Phi^{-1}(1-\alpha)\right)$$

Choice of critical region based on power (3)



But we might consider $\mu < \mu_0$ as well as $\mu > \mu_0$ to be viable alternatives, and choose the critical region to contain both high and low x (a two-sided test).



New critical region now gives reasonable power for $\mu < \mu_0$, but less power for $\mu > \mu_0$ than the original one-sided test.

No such thing as a model-independent test

In general we cannot find a single critical region that gives the maximum power for all possible alternatives (no "Uniformly Most Powerful" test).

In HEP we often try to construct a test of

H_0 : Standard Model (or "background only", etc.)

such that we have a well specified "false discovery rate",

α = Probability to reject H_0 if it is true,

and high power with respect to some interesting alternative,

H_1 : SUSY, Z', etc.

But there is no such thing as a "model independent" test. Any statistical test will inevitably have high power with respect to some alternatives and less power with respect to others.

Rejecting a hypothesis

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 true. In frequentist statistics only associate probability with outcomes of repeatable observations (the data).

In Bayesian statistics, probability of the hypothesis (degree of belief) would be found using Bayes' theorem:

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

which depends on the prior probability $\pi(H)$.

What makes a frequentist test useful is that we can compute the probability to accept/reject a hypothesis assuming that it is true, or assuming some alternative is true.