Statistical methods of data analysis

K. Desch

Uni Bonn & Köln   SS 2010

Course MSc physics716

Lectures: Mo 9:15 – 11:00   15 min break?

Excercises: 1h (2h every two weeks)   Dr. P. Wienemann and Dr. N. Vlasov

(including computer excercises, CIP pool)

Web: http://pi.physik.uni-bonn.de/~wienemann/teaching/ss2010/

Excercises: Thursday 16-18   room to be announced

Start: 29/04/10

Exam: written test

Literature and resources

Books:

G. Cowan: Statistical Data Analysis, Oxford University Press (1998), 35€

R. J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley (1993), 39€

S. Brandt: Datenanalyse, Spektrum Akademischer Verlag (1999), 26€

Computer programs, tools:

root http://root.cern.ch
data analysis

RooFit http://roofit.sourceforge.net/
fitting

RooStats https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome
statistical tests…

Contents

1. Describe data sets (e.g. sets of measurements) with few numbers (mean, variance, …) = descriptive statistics

2. Randomness in statististical physics
   Describe the properties of large ensembles and derive laws of nature for these ensembles (not for the individual particles)
   (note: classical physics is deterministic – but many unknown boundary conditions – "seemingly" random)

3. Randomness in quantum mechanics
   probability interpretation of wave function

4. Describe measurement errors (uncertaintiy)

   statististical errors: "known" (estimable, assessable) probability distribution

   systematic errors: "unknown" (or no) probability distribution

5. Hypothesis testing
   Compare data with a theoretical model

6. Determine (“fit”) the parameters of a theoretical model and their errors

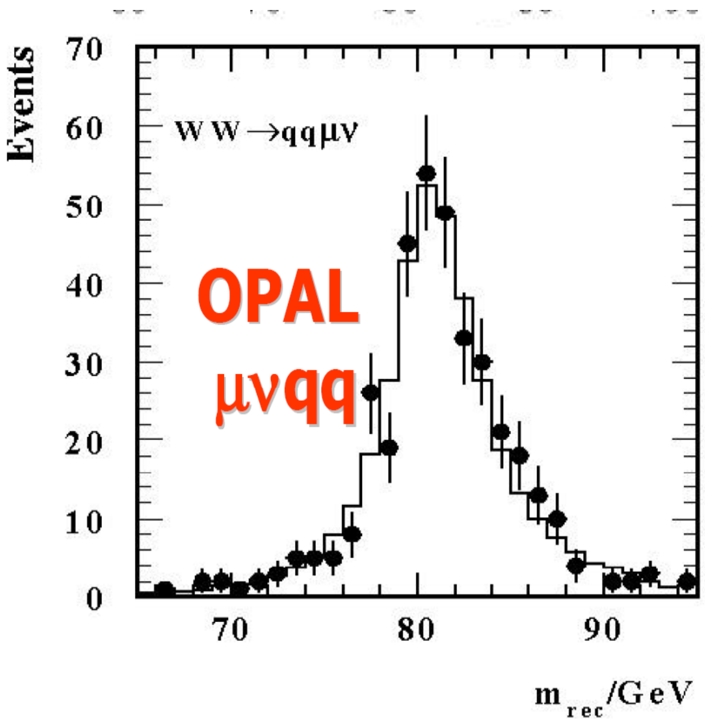
Note:

“Error” (= “uncertainty”) in physics:
True value lies within the error interval with a certain probability

(c.f. “tolerances” in engineering)

Measurement = Compare measured value to a scale
Measurement = Estimate a parameter of a theoretical model from data

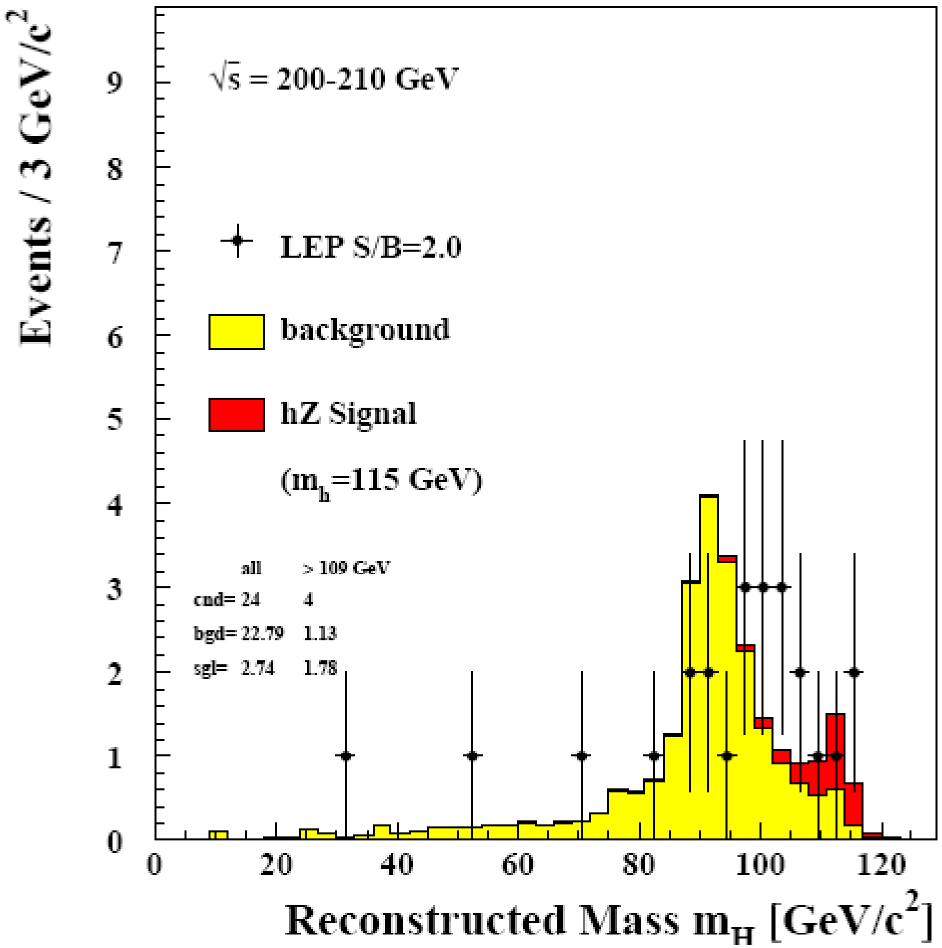
Example: measurement of the mass of the W-boson



Variety of mass extraction methods:

- Reweighting: generate MC 'template' distributions for various true W masses, find 'best' (A,L,O)
- Convolution: decompose observed distribution into physics function ⊗ detector response (D,O)
- Simple Breit-Wigner fit to reconstructed dist. (O)
  - Take into account resolution and ISR ⇒MC studies

Testing hypotheses

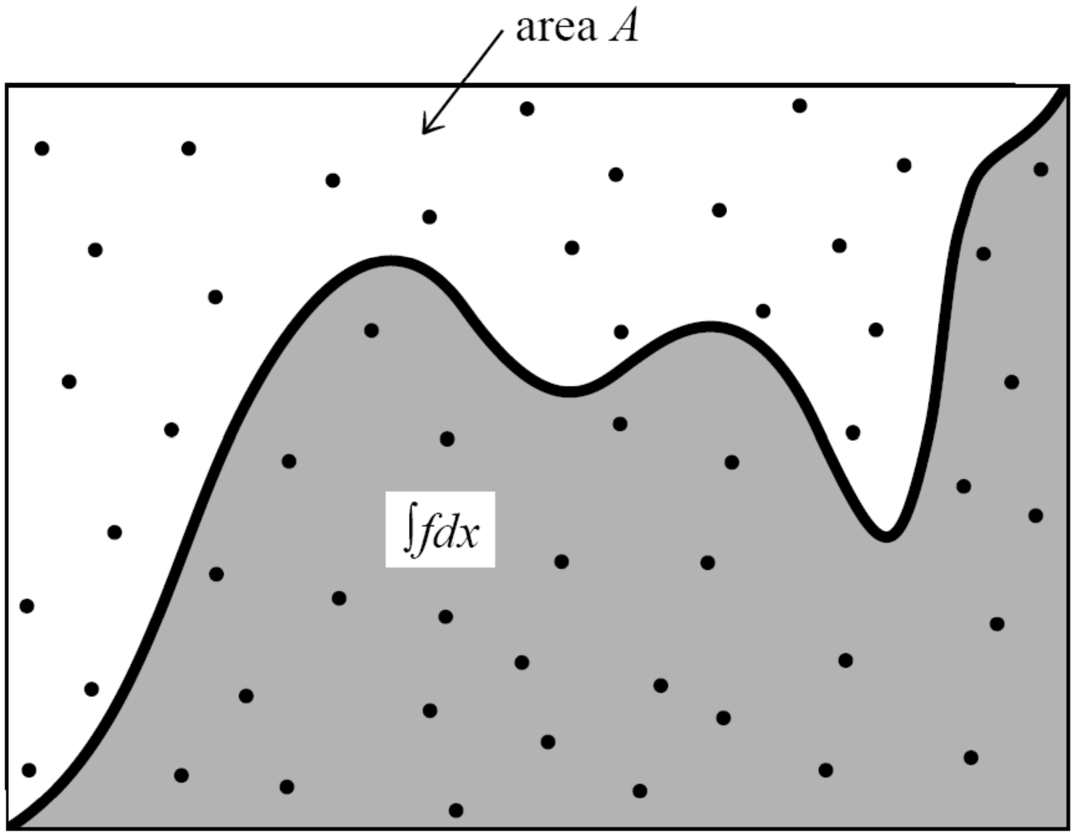


Higgs or no Higgs?

exclusion limits

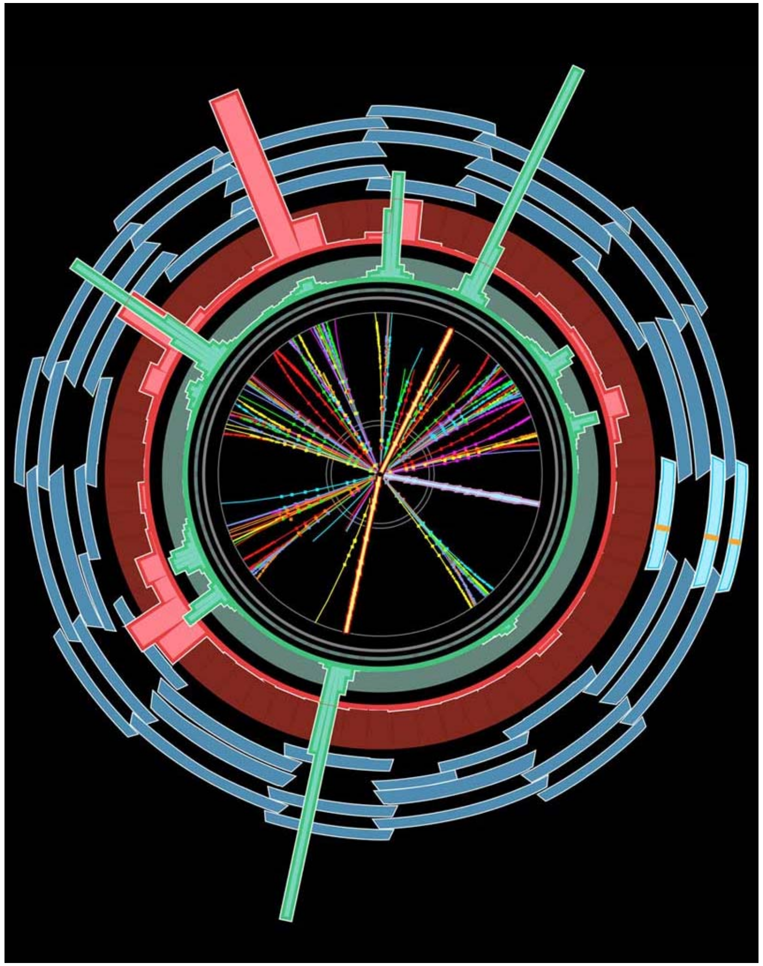
confidence levels

Statistical numerical methods: e.g. integration



Also works in N dimensionen (where classical methods fail…)

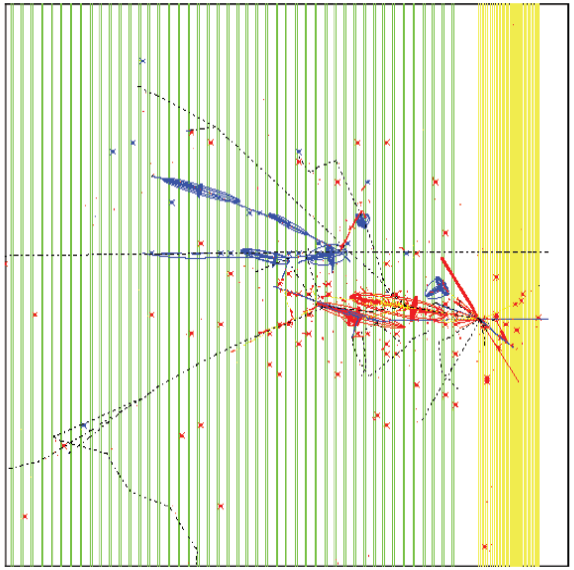
Simulation of particle/nuclear physics experiments



Variety of statistical processes

matrix element
particle transport
detector response

in detail…



Christian Huygens

De Ratiociniis in Ludo Aleae(1657)

14 theses about probability theory

Andrei Kolmogorov
(1903-1987)

Axiome zur mathematischen
Definition von Wahrscheinlichkeit
Heidelberg 1933

Mathematical definition of <u>probability</u>
(no meaning/interpretation what probability actually means,
good: conclusions following from Kolmogorov-Axioms are
independent of the interpretation of probability)

Set S of “samples” (“events”) (sample space)

Assign a real number P(A) to each subset A of S

P(A) := probability of A

with the following properties:

1. $P(A) \geq 0$ for each subset A of S
2. $P(S) = 1$
3. $P(A \cup B) = P(A) + P(B)$
   for disjoint subsets A and B ($A \cap B = \emptyset$)

Implications (w/o proof)

$P(\emptyset) = 0$
$0 \le P(A) \le 1$
$A \subset B \Rightarrow P(A) \le P(B)$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Complement of A:

$$P(\bar{A}) = 1 - P(A)$$

$$P(A \cup \bar{A}) = 1$$

A and B are called statistically independent (or uncorrelated) if
$P(A \cap B) = P(A)P(B)$

→ If an event belongs to A nothing is implied about its belonging to B

Important concept!

Example 1 (uncorrelated):

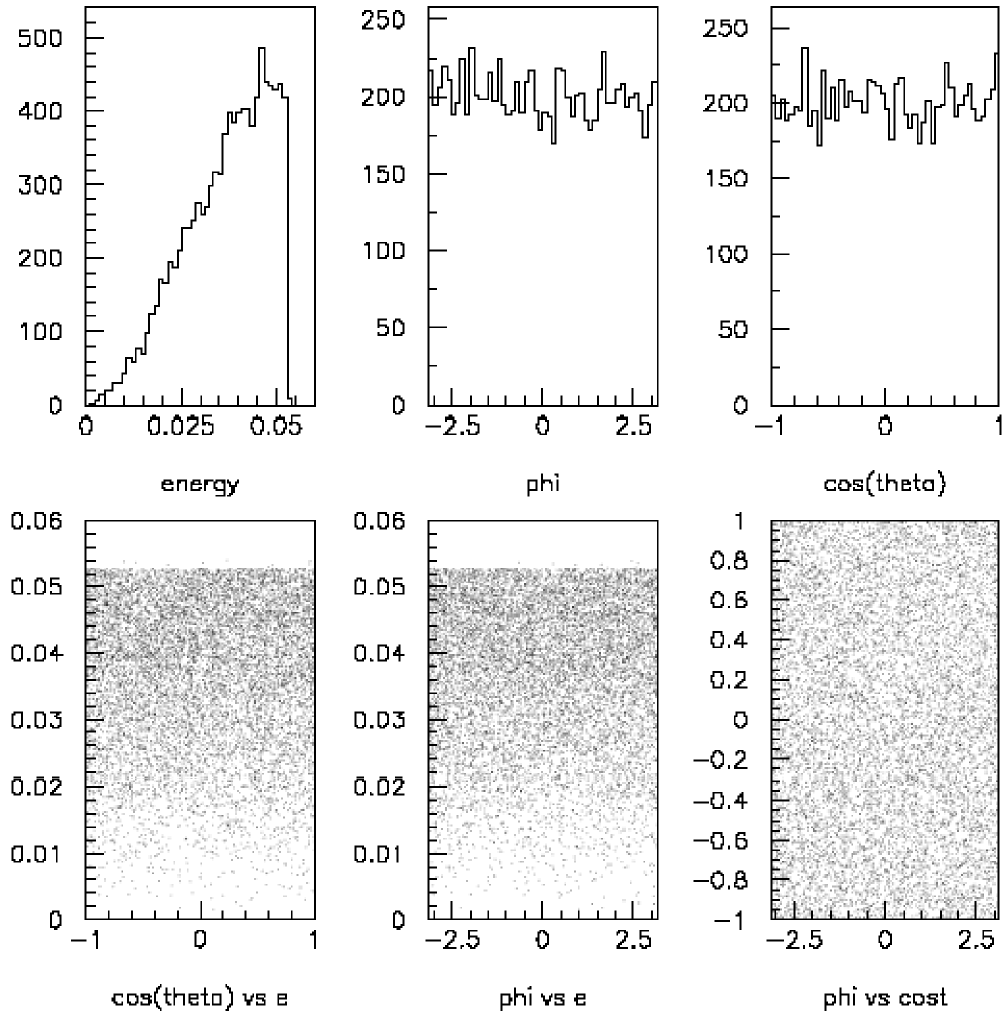
S = all students of Bonn university
A = all male students of Bonn university
B = all students whose birthday is between Jan 1st and April 30th
A and B are (presumably) uncorrelated, therefore $P(A \cap B) = P(A)P(B)$

Example 2 (correlated):

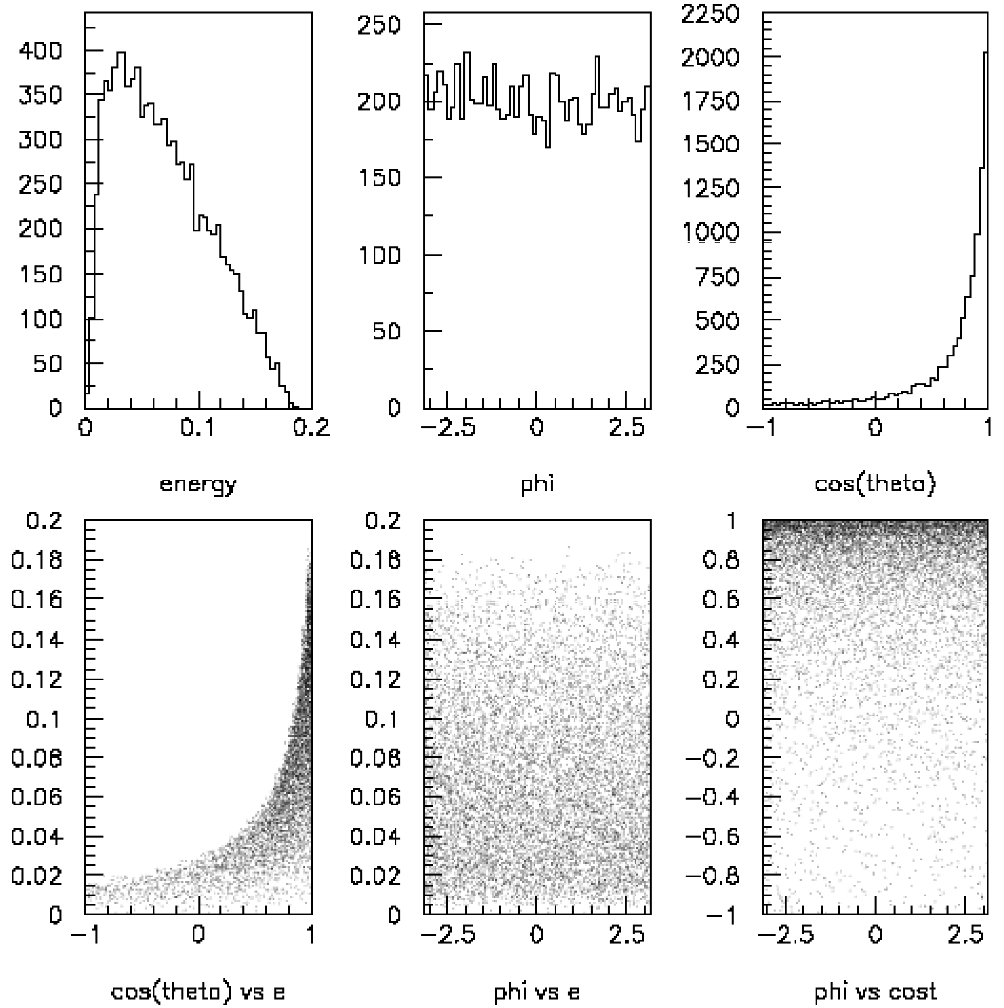
S = all people living in germany
A = all people living in germany under the age of 12 years
B = all people living in germany shorter than 150 cm

A and B are (strongly) positively correlated, i.e. $P(A \cap B) > P(A) * P(B)$

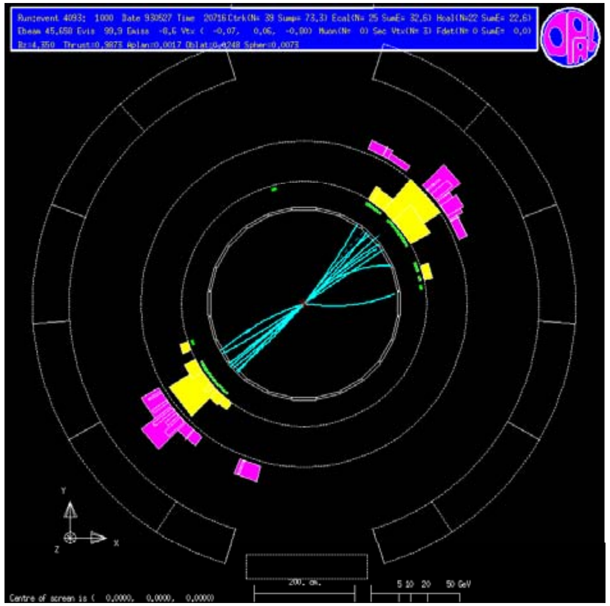
Example 3: Muon decay at rest $\mu \rightarrow e\,\nu\nu$



Example 3: Muon decay in flight $\mu \rightarrow e\,\nu\nu$



Example 4: $e^+e^- \to b\bar{b}$



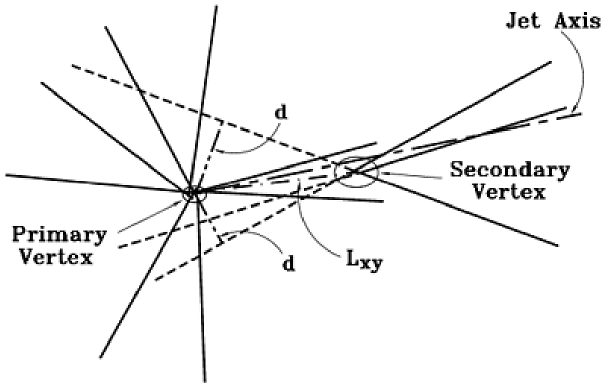
Measurement of the ratio $R_b = \dfrac{\sigma(b\bar{b})}{\sigma(q\bar{q})}$

Identification of b-quarks through secondary vertex

Needed: efficiency to "tag" a b-jet P(b)

Determine from "double tags"

$$\frac{N(b\bar{b})}{N(b)} \approx P(b)$$



assuming $P(b\bar{b}) = P(b) \cdot P(\bar{b}) = P(b)^2$

works only if probability to tag either jet are uncorrelated!

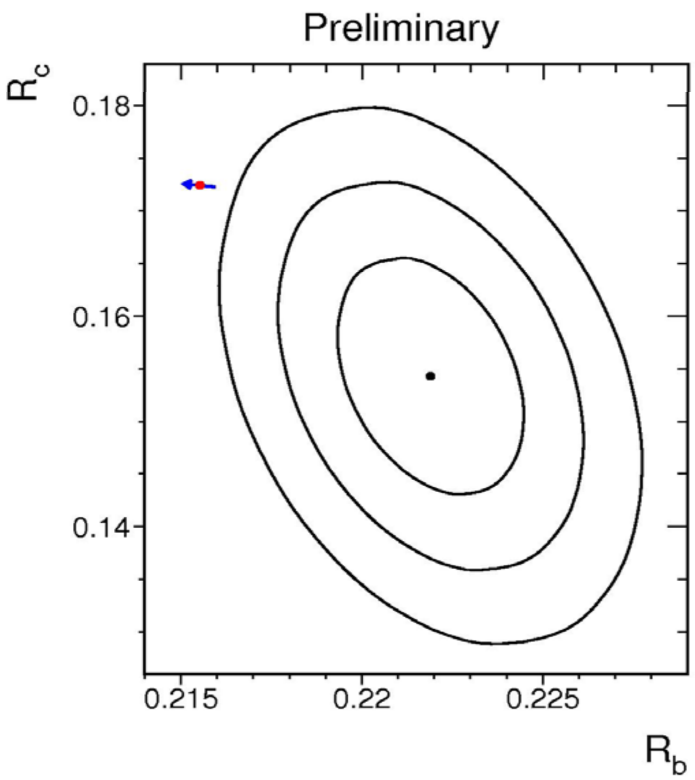


Figure 5: Contours in the R_b-R_c plane derived from LEP data, corresponding to 68%, 95% and 99.7% confidence levels assuming Gaussian systematic errors. The Standard Model prediction for $m_t = 180 \pm 12$ GeV is also shown. The arrow points in the direction of increasing values of m_t.

31

1994

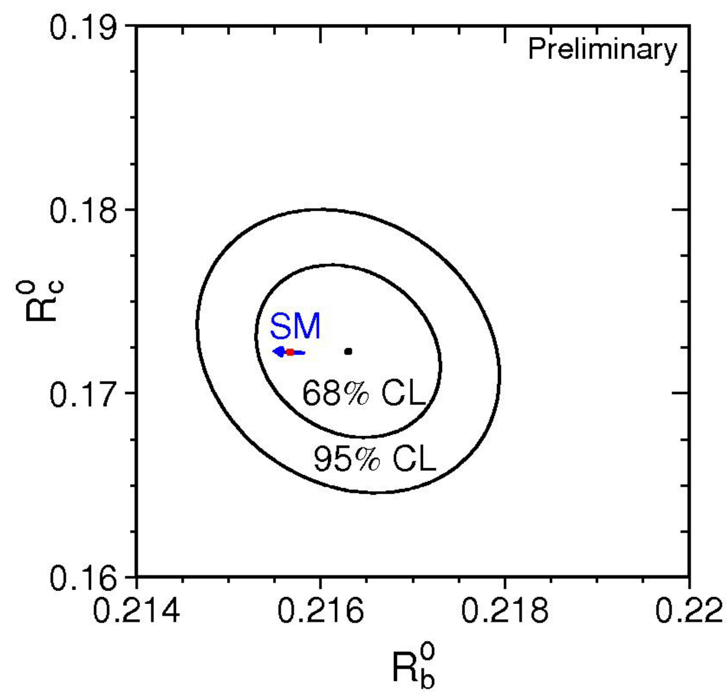
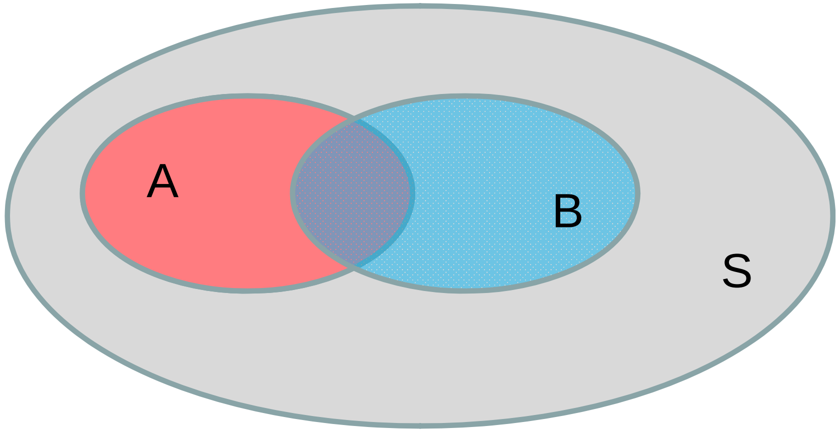


Figure 5.2: Contours in the (R_b^0,R_c^0) plane derived from the LEP+SLD data, corresponding to 68% and 95% confidence levels assuming Gaussian systematic errors. The Standard Model prediction for $m_t = 178.0 \pm 4.3$ GeV is also shown. The arrow points in the direction of increasing values of m_t.

today

Probability for A given B:

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



Probability for B given A:

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

→ Bayes theorem: $P(A|B)\,P(B) = P(B|A)\,P(A)$

Split S in disjoint subsets A_i i.e. $S = \cup_i A_i$ then

$$P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

combined with Bayes theorem: $P(A|B) = \frac{P(B\,|\,A)\,P(A)}{\sum_i P(B|A_i)P(A_i)}$

for any subset A, e.g. for one of the A_i

Example:

disease known to be carried by 0.2% of the population. Prior probability:

$$P(\text{disease})=0.002$$

$$P(\text{no disease}) = 0.998$$

a rather reliable blood test for the disease yields

$$P(+\,|\,\text{disease}) = 0.98$$

$$P(-\,|\,\text{disease}) = 0.02$$

and a small probability for a false positive result

$$P(+\,|\,\text{no disease}) = 0.03$$

$$P(-\,|\,\text{no disease}) = 0.97$$

What is the probability to have the disease if you are tested positive?

$$P(\text{disease} \mid +) = \frac{P(+ \mid \text{disease})P(\text{disease})}{P(+ \mid \text{disease})P(\text{disease}) + P(+ \mid \text{no disease})P(\text{no disease})}$$

$$= \frac{0.98 * 0.002}{0.98 * 0.002 + 0.03 * 0.998} \approx 0.06$$

→ small probability of having the disease even if tested positive by a highly reliable test

1.5.1. Probability as a relative frequency: “Frequentist interpretation”

- elements of S are the possible outcomes of a measurement
- assume measurement can be (at least hypothetically) repeated
- Subset A: occurrence of of any of the outcomes in the subset A

Define probability

$$P(A) = \lim_{n\to\infty} \frac{\text{number of occurences of outcome A in n measurements}}{n}$$

Simon Laplace
1749-1827

This the natural interpretation of probability in
- quantum mechanics
- statistical mechanics

- consistent with Kolmogorov axioms
- can never be determined perfectly (→ estimation)

1.5.2. Subjective („Bayesian“) probability

„Bayes interpretation“

- elements of sample space S are „hypotheses“ or „propositions“ (i.e. statements which are true or false)

Rev. Thomas Bayes
1702-1761

Interpretation of probability

P(A) = degree of belief that the hypothesis A is true

Bayes statistics:

A = hypothesis that a theory is true
B = hypothesis that experiment yields a certain result

Bayes theorem:

$$P(\text{theory} \mid \text{data}) \propto P(\text{data} \mid \text{theory}) \cdot P(\text{theory})$$

Problematic interpretation of P(theory) („a priori“ probability, „prior probability“)
No fundamental rule how to define a prior – but once it is done, no need to estimate the „posterior“ probability, it can be calculated

In the following we will mainly work in the „frequentist“ picture

There are certain limits within which frequentist and Bayesian statistics will yield the same result

We will discuss Bayesian statistics in the context of the principle of maximum likelihood and when setting confidence limits

Simplest case: measurement can only take discrete values x_i
(e.g. couting experiment, x_i = number of counts)

probability to measure x_i : $P(x_i) =: f_i$

$$\sum_i f_i = 1$$

Often the result of a measurement is a continous quantitiy x
Probability to obtain exactly x is zero
Better: probability to obtain a measurement in the interval [x+dx]
P([x+dx]) =: f(x) dx

f(x) = probability densitiy function (p.d.f.) with

$$\int_S f(x)dx = 1 \quad f(x) \geq 0$$

Integration of f(x) yields a probability:

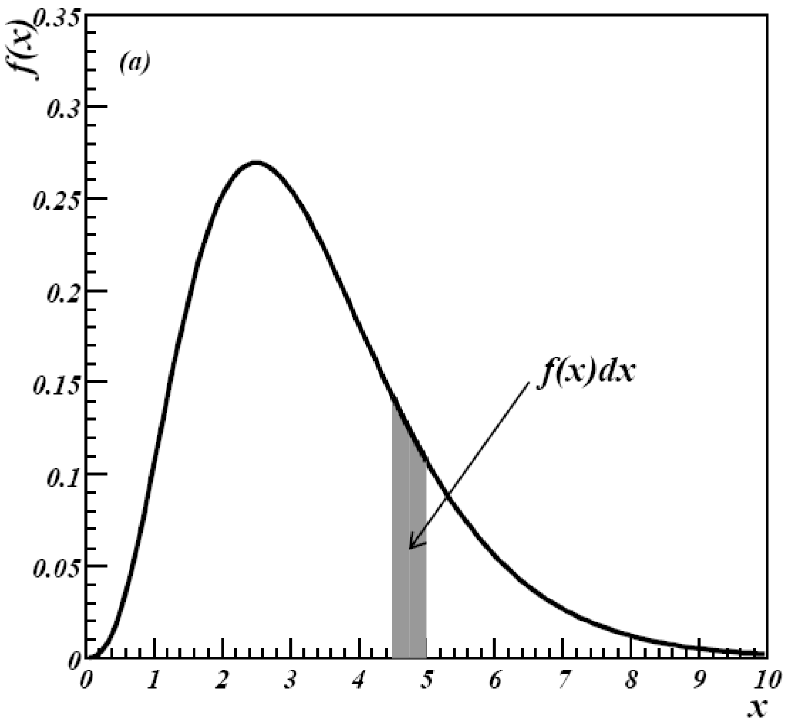
Cumulative Distribution: $F(x) := \int_{-\infty}^{x} f(x')dx'$

yields the probability to obtain a measurement smaller than x

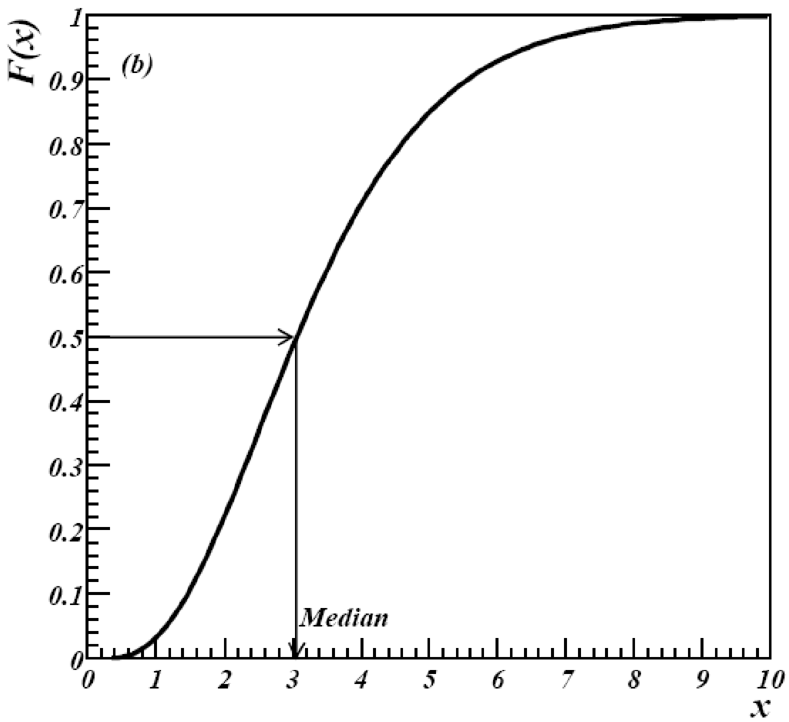
Probability that measurement lies in the interval [a,b] is F(b)-F(a)

For discrete random variables $F(x) := \sum_{x_i \leq x} x_i$

Example for a p.d.f.



and its cumulative distribution:



Quantile: $x_\alpha : F(x_\alpha) = \int_{-\infty}^{x_\alpha} f(x)\,dx = \alpha$

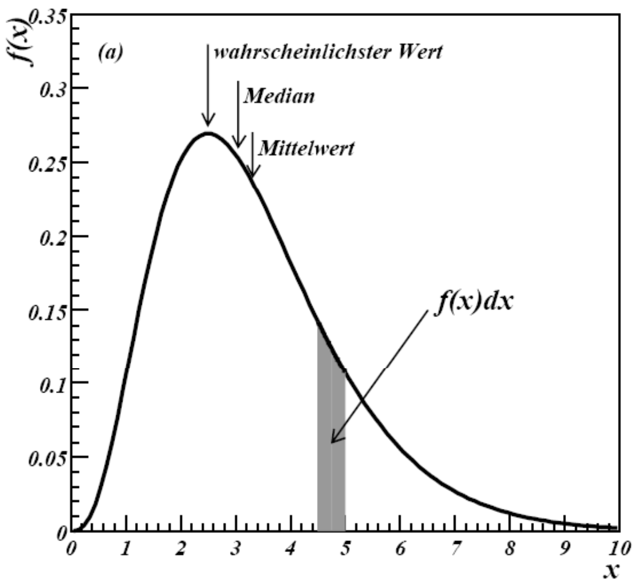
Median: $x_{0.5}$ (50%-value)

Most probable value: maximum of f(x)

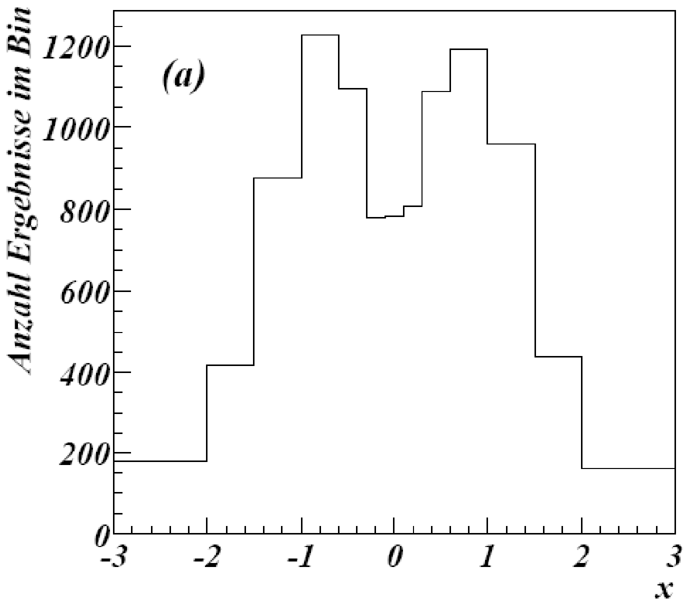
Mean value: $\mu_x = \int_{-\infty}^{\infty} x\, f(x)\,dx$

discrete distribution:

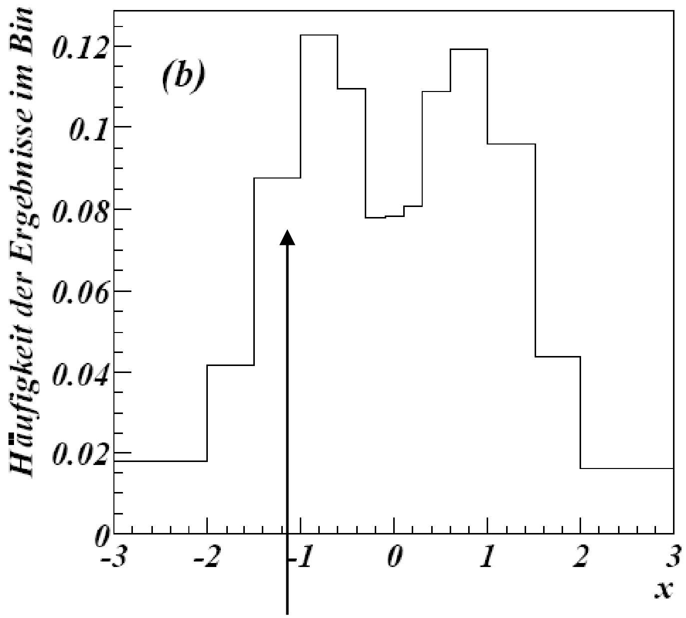
$$\mu_x = \sum_i x_i P(x_i)$$



Histogram: frequency distribution of events



Normalized to area = 1:



bin contents/bin width ~ f(x)
(in the limit bin width → 0)

Expectation value of a function a(x): $$E[a] = \int_{-\infty}^{\infty} a(x)\, f(x)\, dx$$

For a(x) = x one yields the mean value $$\mu \equiv E[x] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

Expectation values of powers of x are called moments of a p.d.f.

algebraic moments: $$E[x^n] \equiv \mu_n' = \int_{-\infty}^{\infty} x^n\, f(x)\, dx$$

central moments: $$E[(x-\mu)^n] \equiv \mu_n = \int_{-\infty}^{\infty} (x-\mu)^n\, f(x)\, dx$$

A measure for the width of a distribution (p.d.f.) is the

variance: $$V[x] \equiv \sigma_x^2 \equiv E[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

The variance is the second central moment von f(x), i.e. its
mean square deviation from the mean value.

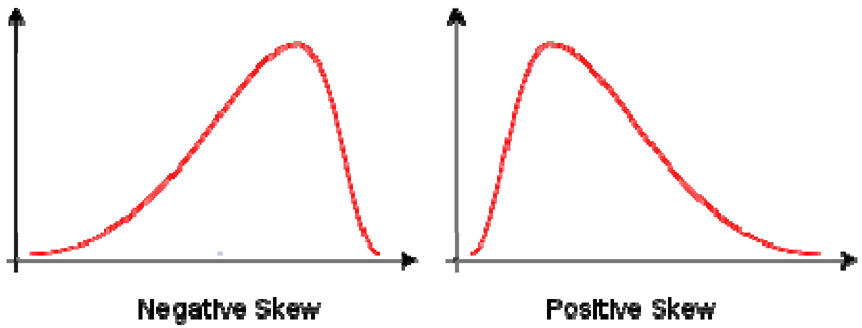
standard deviation: $\sigma_x = \sqrt{V[x]}$ (same units as x, μ)

From linearity of expectation values follows

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

(will be useful later for estimating the variance of discrete p.d.f.´s)

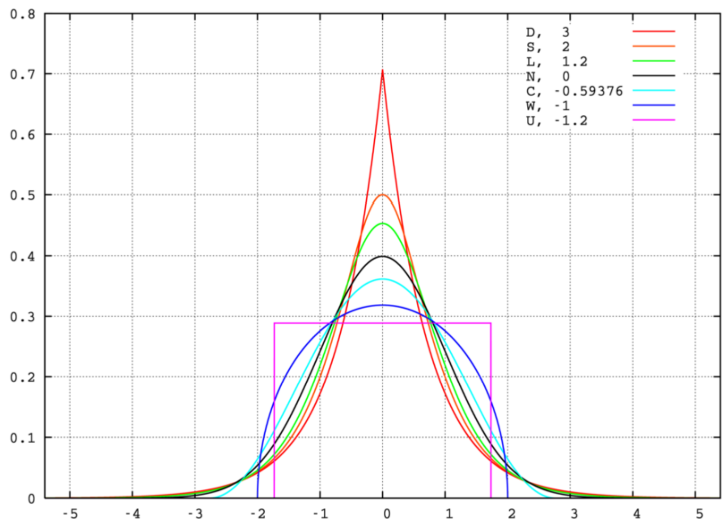
Higher moments:



Skewness: $\gamma = E[(x-\mu)^3]/\sigma^3$

Measure for the asymmetry of a distribution

Kurtosis: $\kappa = E[(x-\mu)^4]/\sigma^4 - 3$



Measures importance of “tails” of a distribution

Larger tails than Gaussian: $\kappa > 0$
Smaller tails than Gaussian: $\kappa < 0$