# Basics in statistical analysis and hypothesis testing for physicists

Michael Winn

Department of Nuclear Physics IRFU/CEA, university Paris-Saclay

DPhN/Irfu, 05/06.12.2023

# Outline

- Introduction: context and scope
- Basic notions of statistical analyses
- parameter estimation and hypothesis testing with maximum likelihood
- goodness-of-fits

# Introduction: statistics in particle and nuclear physics (I)

- Statistical analysis: vast and well studied topic
  → problems treatable by existing methods

- however, time is limited in a PhD thesis

- knowledge prior to PhD thesis heterogeneous

- knowledge of experienced physicists heterogeneous

- standards & sophistication differ between subfields and groups
  → related to: typical precision, ´known´ standard case, typical knowledge on systematic effects, level of industrialisation, group sociology and history

- topic often not in the center of interest of students, supervisors

# Introduction: statistics in particle and nuclear physics (II)

- This situation can lead to:
  → not sufficiently advanced or simply wrong methods
  → ´over-design´: shooting with a bazooka on a fly
  → use of previously used code/methods blindly
  → trust blindly collaborators/common wisdom of group/collaboration

- can lead to wrong parameter/uncertainty estimates or failure to finish

- can lead to knowledge gaps after PhD

# Introduction: Scope

- myself: experimental physicist: ALICE, LHCb, phenomenology
- today´s audience very broad
- impossible to get very far in 2× 1.5 h
- Goals today:
  → basics, comments about Frequentist vs. Bayesian
  → useful references
  → notions of hypothesis testing
- Goals tomorrow: → residual topics of lecture
  → simple illustrations, just bring a pocket calculator tomorrow and a piece of paper and a pen
  → your questions

# Take home message

1. **Take the time** to understand the statistics questions that appear in or around your thesis yourself
2. **Look** in the literature, courses
3. **Decide yourself or follow collaboration reasoning** for an appropriate approach
4. **Discuss** with your supervisor/colleagues/group about open questions/points **early on!**
5. The invested time will pay off for your future:
   be it in physics or not

# Why statistics for physicists?

Broadly speaking: since we want to be quantitative
→ draw reliable conclusions from data

- experimental data: parameter and uncertainty estimation on
  - yields/rates/cross sections, exclusion limits
  - distribution moments/shapes
- analogue in theory computations or fits to experimental data
- compatibility between different data sets and data combination

# Used material

- Statistical methods in particle physics: K. Reygers, R. Stamen, M. Voelkl
  link,
  very rich and good lecture series, used for introduction, also good intro to Monte Carlo Method

- Beyond Standard Model physics: elements of statistical analysis (Master 2 NPAC), N. Morange
  link
  very condensed course by Beyond-the-Standard-Model practitioner, used partly, in particular for hypothesis testing

- Statistics for searches at the LHC: G. Cowan
  arXiv:1307.2487, short summary

- Introduction to Statistics and Data Analysis for Physicists: G. Bohm and G. Zech
  link
  broad text book tailored to high-energy physics; interesting for example collection

# Further reading

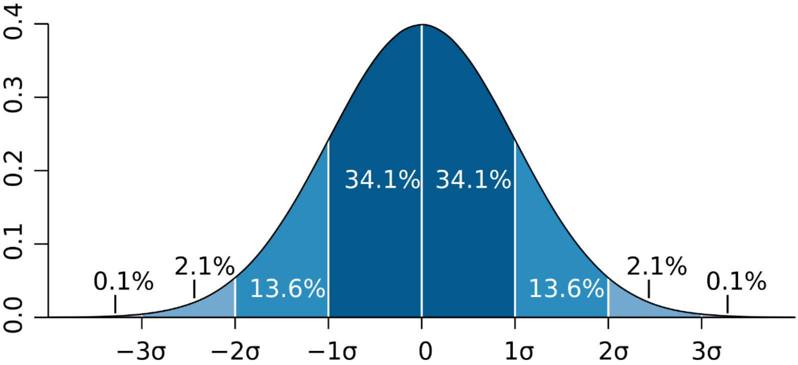
- Asymptotic formulae for likelihood-based tests of new physics. G. Cowan et al., arXiv:1007.1727v3, reference for high-energy physics searches, standard methodology
- ´How good are your fits? Unbinned multivariate goodness-of-fit tests in HEP´, Michael Williams, arXiv:1006.3019, interesting compilation of methods for goodness-of-fits
- ´Parameter uncertainties in weighted unbinned maximum likelihood fits´ C. Langenbruch, arXiv:1911.01303, presense of weights: subtleties
- ´sPlot: a statistical tool to unfold data distributions´, M. Pivk, Fr. Le Deberder, link, separate background and signal under independence condition, heavily used in B-physics community
- PDG review on statistics (G. Cowan) https://pdg.lbl.gov/2023/reviews/
- Pitfalls of Goodness-of-Fit from Likelihood, J. Heinrich arXiv:0310167
- Systematic Errros: Facts and Fiction, R. Barlow, not directly related, but very crucial, often most delicate part, arXiv:0207026
  → correct as much as possible, often detection of most important effect more crucial than quantifying precisely all effects

# Comment on communities

that I can relate to ...

- unbinned maximum likelihood approaches: vast experience in B-physics community (b-factories, LHCb), also a lot in ATLAS, CMS
- binned maximum likelihood approaches: vast experience in ATLAS, CMS $\rightarrow$ material for binned approaches from ATLAS/CMS authors, unbinned from LHCb
- Heavy-ion community (ALICE, STAR, PHENIX,..): often large statistics and large backgrounds, $\chi^2$-fits frequently used
- High-dimensional binned data to deduce models in theory/phenomenology:
  parton distribution functions, review; light-flavour baryon spectroscopy, recent review

# Basic notions of statistics for physics



# Sources of uncertainties

You will always have uncertainties in measurements

- Underlying theory (quantum mechanics) is probabilistic
  → true randomness

- Limited knowledge about the measurement process
  → present without quantum mechanics

Quantify uncertainty using probabilitiy

# Mathematical definition of probability

Let $A$ be an event. Then probability is a number obeying three conditions, the Kolmogorov axioms:

- $P(A) \geq 0$ (non-negative real number)
- $P(S) = 1$, where S is the set of all $A$, the sample space
- $P(A \cup B) = P(A) + P(B)$ for any $A$, $B$ which are exclusive, i.e. $A \cap B = 0$

From these axioms further properties can be derived, e.g.:

$$P(\overline{A}) = 1 - P(A) \tag{1}$$

$$P(\{\}) = 0 \tag{2}$$

$$\text{if}\, A \subset B, \text{then} P(A) \leqq P(B) \tag{3}$$

But what does P mean?

# Interpretations of probability

https://plato.stanford.edu/entries/probability-interpret/

- **Classical**
  → Assign equal probabilities based on symmetry of the problem, e.g., rolling dice: $P(6) = 1/6$

- **Frequentist**
  → Let $A, B, ....$ be outcomes of a repeatable experiment:
  $P(A) = lim_{n \to \infty} \frac{\text{times outcome is} A}{n}$

- **Bayesian** (subjective probability)
  → $A, B, ...$ are hypotheses (statements that are true or false)
  $P(A) =$ degree of believe that A is true

# Bayesian vs Frequentist statistics

Both philosophies coexist:

- Bayesian:
  → Closer to everyday reasoning, where probability is interpreted as a degree of belief that something will happen, or that a parameter will have a given value.

- Frequentist:
  → Closer to scientific reasoning, where probability means the relative frequency of something happening. This makes it more objective, since it can be determined independently of the observer, but restricts its application to repeatable phenomena.

- So what?
  → For practical matters, results tend to be very similar in the asymptotic regime of large numbers
  → There exist nonetheless some important differences (coverage, goodness-of-fit...)

# Bayesian vs Frequentist: take-home messages

"**Bayesian** address the questions everyone is interested in by using assumptions that no one believes. **Frequentists** use impeccable logic to deal with an issue that is of no interest to anyone."
(Louis Lyons)

- Communities tend to lean towards one approach
  → Cosmology is mostly using Bayesian statistics (there is only 1 universe...)
  → High-energy physics is more frequentist

- Will use frequentist approach in the following
  → by far the most common at the LHC, my background
  short discussion on differences in the following

# Conditional probability and independent events

For two events A and B, the conditional probability is defined as

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

Event A and B independent $<=>$ $P(A \cap B) = P(A) \cdot P(B)$

An Event A is independent of B if $P(A|B) = P(A)$

# Bayes´ theorem

Definition of conditional probability:

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

First published (posthumously) by the Reverend Thomas Bayes (1702-1761),
First modern formulation by Pierre-Simon Laplace in 1812

Accepted by everyone also if probabilities are not Bayesian probabilities

# A simple example: particle identification

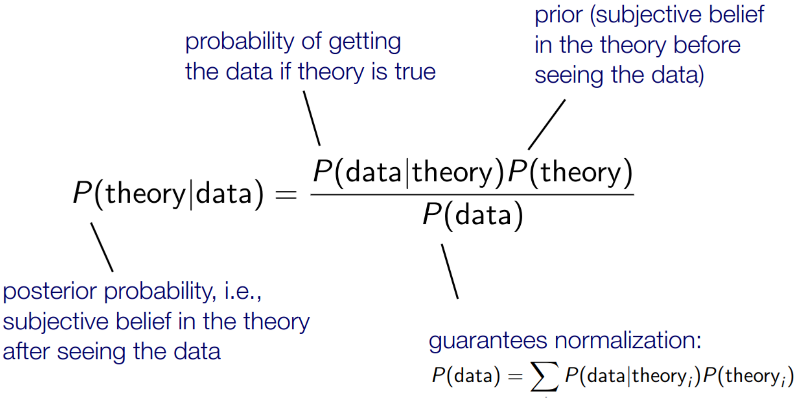
Consider a detector for electron identification

Assume you have the following information based on some calibration data or simulation:

- $P(A|electron)$, i.e., the efficiency to identify the electron
- $P(A|notanelectron)$, i.e. efficiency for background
- $P(notA|electron) = 1 - P(A|electron)$
  $P(notA|notanelectron) = 1 - P(A|notanelectron)$

**Question**: Given a sample of tracks $S$ passing selection A, what fraction of them are electrons?
→ i.e. what is $P(electron|A)$ ?
**Answer**: Cannot be determined with the provided information: Need in addition: $P(electron)$, the true fraction of electrons within $S$.
Provided this information, Bayes´ theorem inverts conditionality:
$P(electron|A) = P(A|electron) \cdot P(electron)/P(A)$

# Bayesian inference: Degree of believe in a theory given a certain set of data (I)



probability of getting the data if theory is true

prior (subjective belief in the theory before seeing the data)

$$P(\text{theory}|\text{data}) = \frac{P(\text{data}|\text{theory})P(\text{theory})}{P(\text{data})}$$

posterior probability, i.e., subjective belief in the theory after seeing the data

guarantees normalization:

$P(\text{data}) = \sum_i P(\text{data}|\text{theory}_i)P(\text{theory}_i)$

Addresses question: "What should I believe?"

# Bayesian inference: Degree of believe in a theory given a certain set of data (II)

For a continuous parameter $\lambda$:

$P_{posterior}(\lambda|m) = \frac{f(m|\lambda)P_{prior}(\lambda)}{f_1(m)}$ $\lambda$ : true value of a parameter of nature, $m$ : measurement,

$f_1(m) = \int f(m|\lambda^*)P(\lambda^*)d\lambda^*$

**Problems with Bayesian inference**

What functional form to chose for $P_{prior}(\lambda)$

Uniform prior: flat in continuous variable? In which variable, e.g. in $\lambda, \lambda^2, 1/\lambda, \ln(\lambda)$

→ criticism that can be addressed by ´Jeffrey´s prior´ constructions (however, does not solve the principle problem of missing knowledge)

**Bayesian reply**

Choice of prior usually unimportant after a few experiments

Not so easy for discovery physics

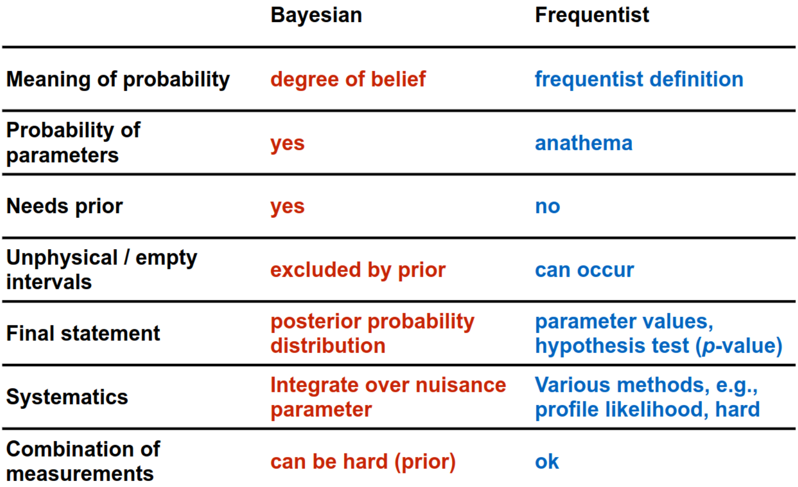
# Are you a Frequentist or a Bayesian?

Suppose mass of a particle is measured with a Gaussian resolution $\sigma$ and the result is reported as
$m \pm \sigma$

- **Bayesian**:
  $P(m|m_{true}) \propto e^{-(m-m_{true})^2/(2\sigma^2)} \rightarrow$ (flat prior for $m_{true}$)
  $P(m_{true}|m) \propto e^{-(m-m_{true})^2/(2\sigma^2)}$

- **Frequentist**:
  This is a statement about the interval $[m-\sigma, m+\sigma]$. For a large number of hypothetically repeated experiments, the interval would contain the true value in 68% of the cases. In the frequentist approach, a probabilistic statement about the true value is nonesense (the true value is what it is).

# Bayesian vs. Frequentist [based on L. Lyons]



| | Bayesian | Frequentist |
|---|---|---|
| Meaning of probability | degree of belief | frequentist definition |
| Probability of parameters | yes | anathema |
| Needs prior | yes | no |
| Unphysical / empty intervals | excluded by prior | can occur |
| Final statement | posterior probability distribution | parameter values, hypothesis test (*p*-value) |
| Systematics | Integrate over nuisance parameter | Various methods, e.g., profile likelihood, hard |
| Combination of measurements | can be hard (prior) | ok |

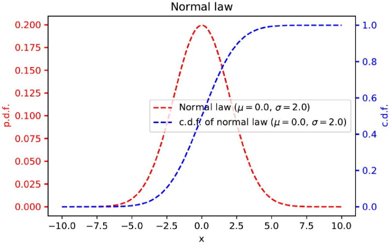
# Random variables and probability density functions

Random variables:

- Variable whose possible values are numerical outcomes of a random phenomenon
- Can be discrete or continuous

Probability density function (pdf) $f$ of a continuous variable:
$P(x \text{ found in } [x, x + dx]) = f(x)dx$
Normalization: $\int_a^b f(x)dx = 1$
It is related to the cumulative function F

- F so that $F(x_0) = P(x \leq x_0)$
  $\rightarrow F(a) = 0, F(b) = 1$
- $f(x)dx = F(x + dx) - F(x)$



# Histograms

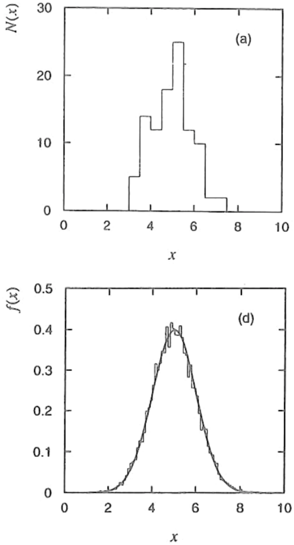
Histogram:

- representation of the frequencies of the numerical outcome of a random phenomenon

pdf = histogram for

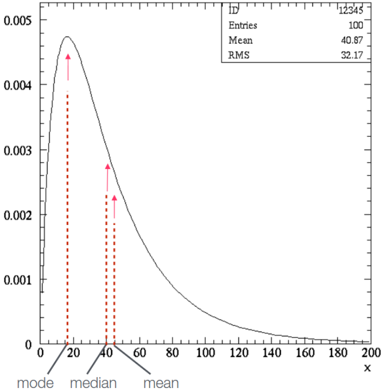
- infinite data sample
- zero bin width
- normalized to unit area

$f(x) = \frac{N(x)}{n\Delta x}$
$n$ = total number of entries, $\Delta x$ = bin width



# Mean, Median, and Mode

- **Mean**
  of a data sample: $\overline{x} = \frac{1}{N}\Sigma_{i=1}^{N} x_i$
  of a pdf: $\mu \equiv < x > \equiv \int xP(x)dx$
  $\equiv$ expectation value $E[x]$
- **Median**:
  point with 50% probability above and 50% probability below, can define other quantiles analoguously
- **Mode**: the most likely value



# Variance, standard deviation, moments

- **Variance**

  of a distribution: $V(x) = \int dx P(x)(x-\mu)^2 = E[(x-\mu)^2]$

  $V(x) =< x^2 > -\mu^2 =< x^2 > - < x >^2$

  Sample variance: $V(x) = \frac{1}{N}\sigma_i(x_i - \bar{x})^2 = \bar{x^2} - (\bar{x})^2$

  This formula underestimates the variance of underlying distribution as it uses the mean calculated from data!

  Use this if you have to estimate the mean from data (**unbiased sample variance**): $\hat{V}(x) = \frac{1}{N-1}\Sigma_i(x_i - \bar{x})^2$

  Use this if you know the true mean $\mu$: $V(x) = \frac{1}{N}\sigma_i(x_i - \mu)^2$

- **Standard deviation**: $\sigma = \sqrt{V(x)}$

- **higher moments** $(E((X-\mu)^n))$:

  → skew $(n = 3)$: left-right asymmetry

  → kurtosis (n=4): measures the size of the tails, if positive larger than a Gaussian

# Multi-dimensional case

For example: 2 variables $x, y$ with joint pdf $f$

- A **marginal pdf** is defined as:
  $f_X(x) = \int dy^{''} f(x, y^{''})$
  $f_X$ is a projection of $f$, the other variables are integrated.
- A **conditional pdf** is defined as:
  $f_C(x, y_0) = f(x, y_0) = \frac{f(x,y_0)}{\int dx^{''} f(x^{''} y_0)}$
  $f_C$ is a slice of $f$

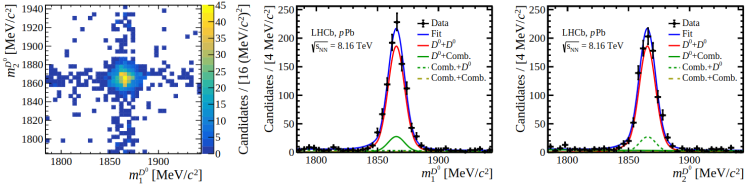


Figure 1: (Left) Two-dimensional invariant-mass distributions of $(m_1, m_2)$ for $D^0D^0$ pairs and the projections on (middle) $m_1$ and (right) $m_2$ with the fit results superimposed. Shown in the projection plots are (points with bars) $p$Pb data, (solid blue) the total fit and its four components.

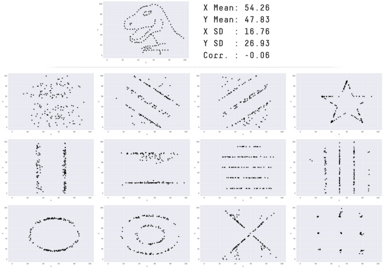
2D-data, projection of data and marginal fit-pdf from PRL 125 (2020) 212001

# Independence and correlation

- Two variables $X$ and $Y$ are independent if: $f(x,y) = f_X(x)f_Y(y)$
- Correlation coefficient between two variables $X$ and $Y$ $\rho(X,Y) = \frac{C(X,Y)}{\sigma_x \sigma_y}$ with $C(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - E(X)E(Y)$

Independent $=> \rho = 0$
The opposite is not true!

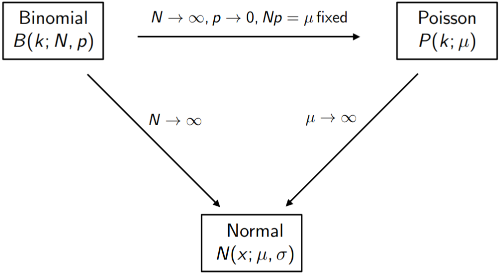
https://www.research.autodesk.com/publications/same-stats-different-graphs/



```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

# Important distributions

- Poisson: $p(k,\mu) = \frac{\mu^k}{k!}e^{-\mu}$, $E[k] = \mu$, $V[k] = \mu$
- Binomial:
  $f(k,N,p) = \frac{N!}{k!(N-k)!}p^k(1-p)^{N-k}$, $E[k] = Np$, $V[k] = Np(1-p)$
- Normal (or Gaussian) distribution: $g(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}exp(-\frac{(x-\mu)^2}{2\sigma^2})$,
  $E[x] = \mu$, $V[x] = \sigma^2$



Binomial $B(k;N,p)$

$N \to \infty$, $p \to 0$, $Np = \mu$ fixed

Poisson $P(k;\mu)$

$N \to \infty$

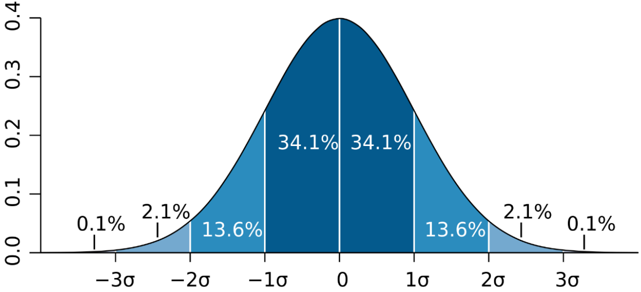
$\mu \to \infty$

Normal $N(x;\mu,\sigma)$

Poisson $P(k;\mu)$ : $\quad \frac{k-\mu}{\sqrt{\mu}} \to N(0,1)$ as $\mu \to \infty$

Binomial $B(k;n,p)$ : $\quad \frac{k-np}{\sqrt{np(1-p)}} \to N(0,1)$ as $n \to \infty$

# Deviation in units of $\sigma$ for a Gaussian



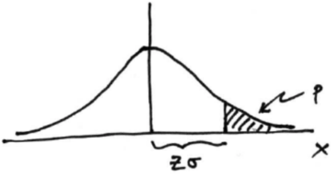
$P(Z\sigma) = \frac{1}{2\pi}\int_{-Z}^{Z} e^{-x^2/2} dx$

- 68.27%/95.45%/99.73% of area within $\pm 1/2/3\sigma$
- 90% of area within $\pm 1.645\sigma$
- 95% of area within $\pm 1.960\sigma$
- 99% of area within $\pm 2.576\sigma$

2D gaussian $1\sigma$-ellipse: contains 39%

# *p*-value and significance

**p-value**:
probability that a random process produces a measurement thus far, or further, from the true mean
p-value and significance $Z$ (one-tailed):



$$p = 1 - \Phi(Z), \;\; Z = \Phi^{-1}(1-p)$$

One-tailed Gaussian *p*-values

| Deviation | p-value |
|---|---|
| 1 σ | 0.16 |
| 2 σ | 0.023 |
| 3 σ | 0.0013 |
| 4 σ | $3.2 \times 10^{-5}$ |
| 5 σ | $2.9 \times 10^{-7}$ |

standard to report a “discovery” → 5 σ

Φ cumulative Gaussian function.

# Why Gaussians are so useful?

# Why Gaussians are so useful?

**Central limit theorem:**

- When independent random variables are added, their properly normalized sum tends towards a normal distribution even if the original variables themselves are not normally distributed, but have a finite variance

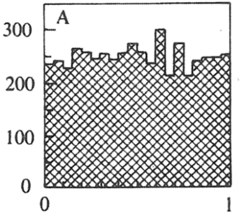
More specifically:

- Consider $n$ random variables with finite variance $\sigma_i^2$ and arbitrary pdfs:
  $y = \Sigma_{i=1}^n x_i \rightarrow_{n \rightarrow \infty} y$ follows Gaussian with $E[y] = \Sigma_{i=1}^n \mu_i$,
  $V[y] = \Sigma_{i=1}^n \sigma_i^2$

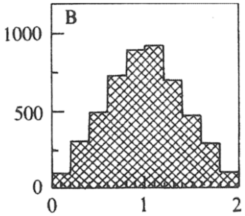
Measurement uncertainties are often the sum of many independent contributions. The underlying pdf for a measurement can therefore be assumed to be a Gaussian.
Many convenient features in addition, e.g. sum or difference of two Gaussian random variables is again a Gaussian.

# The central limit theorem at work

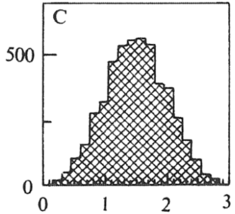
A: x taken from a uniform PD in [0,1], with μ=0.5 and σ2=1/12, N=5000



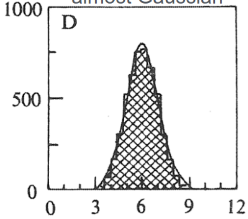
B: X = $x_1+x_2$ from A, *N*=5000, flat shoulders



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



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

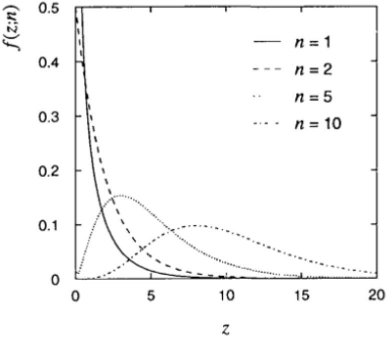


# $\chi^2$-distribution

Let $x_1$, .., $x_n$ be n independent standard normal $(\mu = 0, \sigma = 1)$ random variables. Then the sum of their squares
$z = \Sigma_{i=1}^{n} x_i^2$
follows a $\chi^2$ distribution with $n$ degrees of freedom.

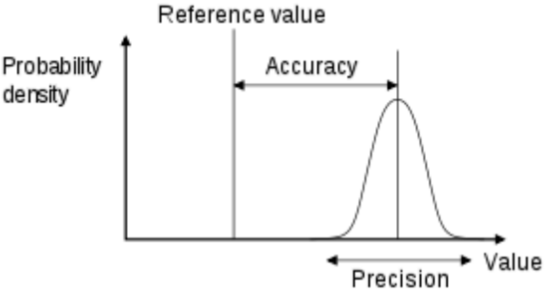
$\chi^2$ distribution

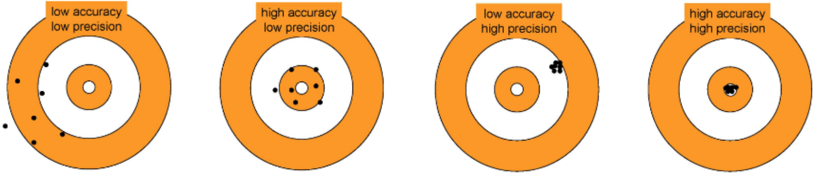
- $f(z; n) = \frac{z^{n/2-1} e^{-z/2}}{2^{n/2}\Gamma(\frac{n}{2})}$ $(z \leq 0)$
- $E[z] = n, V[z] = 2n$
- mode: $max(n-2, 0)$
- $\approx$ Gauss for $n > 100$



Application: goodness of fit $\chi^2 = \Sigma_{i=1}^{n} (\frac{y_i - h(x_i)}{\sigma_i})^2$

# Precision and accuracy





# Parameter estimation

- Suppose we have a model, represented by a pdf $f(x|\Theta)$
  $\rightarrow$ x is a **random variable**
  $\rightarrow$ $\Theta$ represents parameters that affect the shape of the pdf

- Collect a sample of observed data $x = (x_1, x_2, ..., x_N)$

- Intend to say something about $\Theta$ using the observed data

- An **estimator** is a function of the data (*a statistic*) that is used to **estimate the value** of a parameter:

- $t_N(x)$
- $t_N(x) \rightarrow \Theta$ ?

# Estimator properties

*Not all estimators behave the same*

$X$ is a random variable of pdf $f(x|\Theta_0)$ with $\Theta_0$ unknown. An estimator $t_N$ of $\Theta_0$ can be:

- **unbiased**(**accuracy**): if the bias $b_N = E(t_N) - \Theta_0 = 0$
- **convergent** (or consistent): mathematical convergence towards the true value for large enough $N$
- **efficient**(**precision**): if the variance of the estimator $V(t_N)$ converges towards a minium variance bound
- **optimal**: if $t_N$ minimises the Mean Square Error (MSE): $MSE(t_N) = V(t_N) + b_N^2$
- **robust**: if it does not depend on a hypothesis on the pdf

# Usual methods to build estimators

- Moments method
  → e.g. sample mean

- Maximum likelihood
  → focus on this for hypothesis testing examples

- Least squares method

# Likelihood function

A random variable $x$ follows a pdf $f(x|\Theta)$ where $\Theta$ represents parameter(s).

N independent observations of $x$ are obtained: $x_1, ...x_N$

The joint pdf of the N observations is then:
$p(\mathbf{X}|\theta) = \Pi_{i=1}^{N} f(x_i|\Theta)$

The likelihood function is this pdf, evaluated with **fixed data X** and regarded **as a function of the parameter** $\Theta$ only:
$L(\Theta) = P(\mathbf{X}|\Theta)$
Notes:

- $L(\theta)$ is **NOT** a pdf for $\Theta$. The area under L is **meaningless**
- It is not even normalised to unity. The **absolute value of the likelihood is also meaningless**

# Maximum likelihood estimators

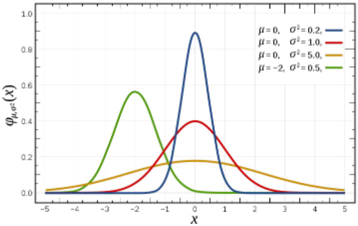
- If the hypothesized $\Theta$ is close to the true value, then there is a high probability to get data like the observed one.
- **The maximum likelihood (ML) estimator(s) are defined as the parameter value(s) for which the likelihood is maximum.**
- In practice, usually - $lnL(\Theta)$ or $-2lnL(\Theta)$ minimized

ML estimators are **not** guaranteed to be always unbiased, neither optimal

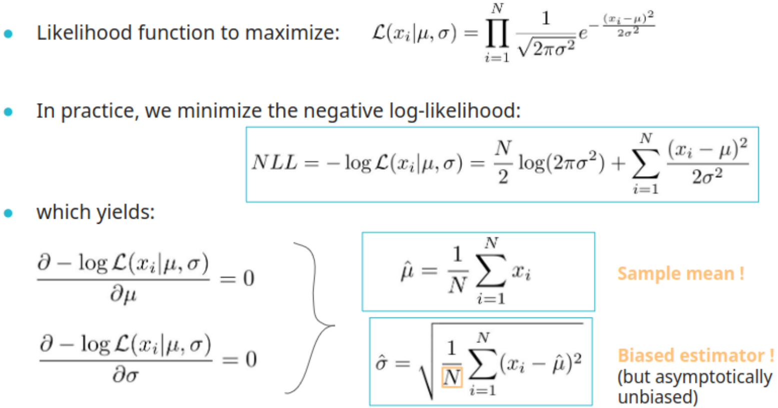
→ however, often method of choice if full pdf available in some analytical or numerical form, see discussion on parameter inference in link as useful guide

# Example: estimation of a Gaussian

- Random process following a Gaussian law of unknown mean and variance:
  - Example: Invariant mass distribution of $Z \rightarrow e^+e^-$
  - Parameters: θ ↦ μ mean, σ standard error
  - Observables: $x_i$
  - PDF: f ↦ $G(x \mid \mu, \sigma) = 1/\sqrt{(2\pi\sigma^2)} \exp(-(x-\mu)^2 / 2\sigma^2)$



# Example: estimation of a Gaussian



- Likelihood function to maximize: $\mathcal{L}(x_i|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$
- In practice, we minimize the negative log-likelihood:

$$\boxed{NLL = -\log\mathcal{L}(x_i|\mu,\sigma) = \frac{N}{2}\log(2\pi\sigma^2) + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- which yields:

$$\left.\begin{array}{l} \dfrac{\partial - \log\mathcal{L}(x_i|\mu,\sigma)}{\partial\mu} = 0 \\[2ex] \dfrac{\partial - \log\mathcal{L}(x_i|\mu,\sigma)}{\partial\sigma} = 0 \end{array}\right\}$$

$$\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

Sample mean !

$$\boxed{\hat{\sigma} = \sqrt{\frac{1}{\boxed{N}}\sum_{i=1}^{N}(x_i-\hat{\mu})^2}}$$

Biased estimator !
(but asymptotically unbiased)

# Coverage probability and confidence interval

## Beyond parameter estimation: parameter uncertainty

- Frequentists report **confidence intervals**, which will contain the true value of the parameter Θ a certain fraction of the time (called the **confidence level**)

- **Frequentist Principle** (Neyman): Construct statements such that a fraction $f \geq 1 - \alpha$ of them are true over an ensemble of statements
  f is called coverage probability
  $1 - \alpha$ is called confidence level

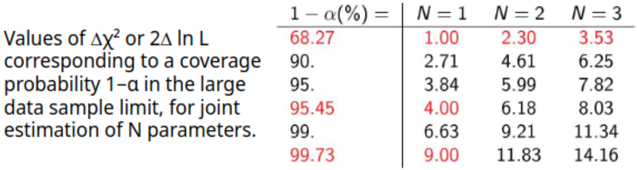
  An ensemble of statements that obyes the Frequentists Principle is said to ´cover´

- Application of confidence intervals: if we report a confidence interval I and we repeat the experimenta N times, then a fraction $f$ of the intervales I will contain the true value of the parameter

# Confidence intervals for Maximum-likelihood estimators

- Finding procedures that give correct coverage (i.e. neither undercoverage nor overcoverage) is not trivial

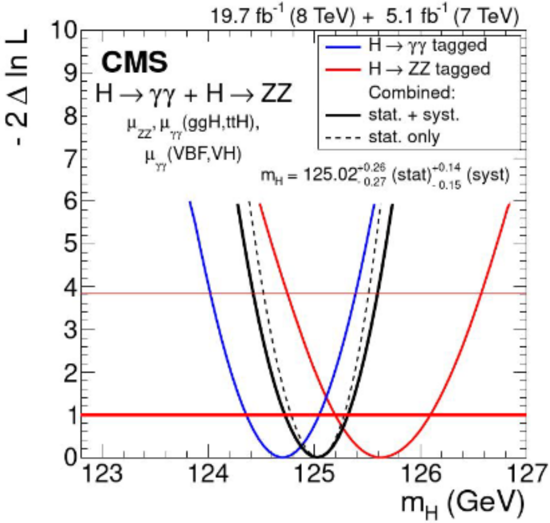
- Asymptotic properties of log-likelihoods to the rescue:
  → **Wald´s approximation**: the likelihood shape is asymptotically gaussian around its maximum
  → **Wilk´s theorem**: $-2 lnL(\Theta)$ asymptotically follows a a $\chi^2$ distribution with d degrees of freedom, where d is the dimensionality of $\Theta$
  → **Consequence**: Confidence intervals can be obtained from the inverse quantiles of a $\chi^2$ distribution with d degrees of freedom, the so-called **likelihood intervals**



Values of $\Delta\chi^2$ or $2\Delta \ln L$ corresponding to a coverage probability 1−α in the large data sample limit, for joint estimation of N parameters.

| $1-\alpha(\%) =$ | $N=1$ | $N=2$ | $N=3$ |
|---|---|---|---|
| 68.27 | 1.00 | 2.30 | 3.53 |
| 90. | 2.71 | 4.61 | 6.25 |
| 95. | 3.84 | 5.99 | 7.82 |
| 95.45 | 4.00 | 6.18 | 8.03 |
| 99. | 6.63 | 9.21 | 11.34 |
| 99.73 | 9.00 | 11.83 | 14.16 |

# Example: Higgs mass measurement



# Hypotheses and tests

**Hypothesis test**

- Statement about the validity of a model
- Tells you which of two competing models is more consistent with the data

**Simple hypothesis**: a hypothesis with no free parameters

- Example: the detected particle is an electron; data follow Poissonian with mean 5

**Composite hypothesis**: contains unspecified parameter(s)

- Example: data follow Poissonian with mean $> 5$

**Null hypothesis $H_0$ and alternative hypothesis $H_1$**

- $H_0$ often the background-only hypothesis
- $H_1$ often signal or signal $+$ background hypothesis

Question to be replied by hypothesis test: Can null hypothesis be rejected by the data?

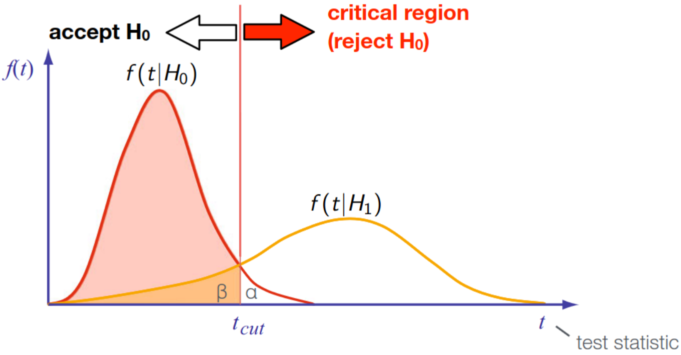
- special case **´Goodness-of-fit´** test:
  Can null hypothesis be rejected by the data without specification of alternative $H_1$?

# Hypothesis testing ingredients

We need:

- A **test statistic** $t(\vec{x})$: a (usually scalar) variable which is a function of the the data alone that is used to test the hypothesis
  $\vec{x} = (x_1, ..., x_n)$: measured features/data
- A **critical region** w such that the hypothesis $H_0$ is false (with a given probability) if $t$ in $w$

# Critical region



The probability for $H_0$ to be rejected while $H_0$ is true:

$$\int_{t_{\mathrm{cut}}}^{\infty} f(t|H_0)\,\mathrm{d}t = \alpha$$

α: "size" or "significance level" of the test

Probability to reject $H_1$ even though it is true:

$$\int_{-\infty}^{t_{\mathrm{cut}}} f(t|H_1)\,\mathrm{d}t = \beta$$

1– β: "power of the test", prob. to reject $H_0$ if $H_1$ is true

# Type I and Type II errors

Type I error:
Null hypothesis is rejected while it is actually true

Type II error:
Test fails to reject null hypothesis while it is actually false

Type I and type II errors and their probabilities:

| | $H_0$ is true | $H_0$ is false (i.e., $H_1$ is true) |
|---|---|---|
| $H_0$ is rejected | Type I error ($\alpha$) | Correct decision ($1-\beta$) |
| $H_0$ is not rejected | Correct decision ($1-\alpha$) | Type II error ($\beta$) |

# Neyman-Pearson lemma

In the comparison of two simple hypothesis $H_0$ and $H_1$, the optimal discriminator is the **likelihood ratio** (LR):

$$t(x) = \frac{L(x|H_1)}{L(x|H_0)} \tag{4}$$

**Notes:**

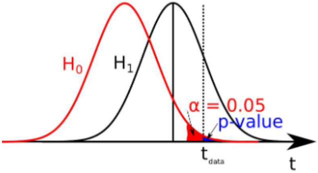
- **Optimal**: minimizes Type II error for a given Type I level of significance
- Valid for **monotonic function of t**
  → Ex: $q(x) = -2\ln(t(x))$
  → Ex: in a counting experiment, number of events
- Strictly valid for simple hypotheses only.
  → However, in practice, **works well for hypotheses in typical LHC applications**

# Procedure for hypothesis testing

- Specify the null and the alternate hypotheses
  $\rightarrow$ Ex: $H_0$ background only, $H_1$ background and additional resonnace
- Build a test statistics $t(x)$ using e.g. Neyman-Pearson lemma
- Specify the significance of the test (what we accept as a false discovery rate)
  $\rightarrow$ Ex: $2.9 \cdot 10^{-7}$ ($5\sigma$) for discovery
  $\rightarrow$ Ex: 0.05 for exclusion
- See where the measurement is $t_{obs}$
- Depending on whether $t_{obs}$ is in or out of the critical region: decide on $H_0$

# p-value and significance

- p-value: $p_0 = p(t \geq t_{obs} | H_0)$
  $\rightarrow$ **Significance level** of a test $\alpha$: chosen **prior** to look at data
  $\rightarrow$ **p-value**: quantity computed **when looking at the data**
- Interpretation:
  $\rightarrow$ probability for the test statistic $t$ to be larger than the observed one $t_{obs}$, under the null hypothesis
- **NOT!!!** the probability that $H_0$ is true
- "Significance" in number of signals:
  $\rightarrow$ translation of the p-value using the integral in one tail of a Gaussian
  $p_0 = \int_Z^\infty G(x|0,1)dx = 1 - \Phi(Z)$
  $\Phi$: cumulative Gaussian function.
- **Convention: 3 sigma evidence, 5 sigma is discovery**



| z-value ($\sigma$) | p-value |
|---|---|
| 1.0 | 0.159 |
| 2.0 | 0.0228 |
| 3.0 | 0.00135 |
| 5.0 | $2.87 \times 10^{-7}$ |

# Spurious exclusion

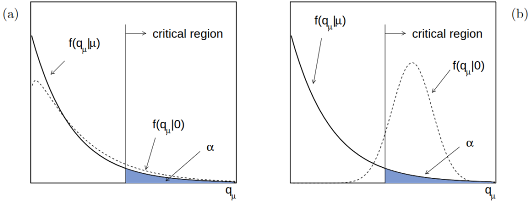


Figure 1: Illustration of statistical tests of parameter values $\mu$ for the cases of (a) little sensitivity and (b) substantial sensitivity (see text).

taken from Cowan et al. arXiv:1105.3166

- for a tested very small signal strength, $H_0$ and $H_1$ may yield very close test statistics t
- can lead to exclusion of hypotheses for which there is little sensitivity
- $\alpha$ for exclusion often set to 0.05, hence in one out of 20 cases, we reject $H_0$
- weak sensitivity can be quantified by the power of the test $\beta$
- one possible strategy:
  standard procedure, but only for the power $\beta(\mu)$ larger than a given threshold, otherwise 100% confidence level
- ´power-constrained´ limit, threshold is pure convention
- see for details in arXiv:1105.3166, see application arXiv:2303.1429

# A word of caveat and the last item of today

- methodology can take into account systematic uncertainties as nuisance parameters, see arXiv:1007.1727

- maximum-loglikelihood estimators and hypothesis testing makes sense and have good properties:
  **Assuming that the PDF (the model) is correct**
  → this is where the physics input is needed
  → Pull distributions, **Goodness-of-fit-tests**

# Goodness-of-fit

Test consistency of the model with the data.

# Goodness-of-fit for least squares fit (1)

The minimum $\chi^2(\vec{(}\Theta))$ of a least-squares fit is a measure of the level of agreement between the model and the data:

$$\chi^2_{min} = \Sigma_{i=1}^{n} \left( \frac{y_i - f(x_i, \vec{\Theta})}{\sigma_i} \right)^2 \tag{5}$$

- Large $\chi^2_{min}$: the model can be rejected.
- If the model is correct, then $\chi^2_{min}$ for repeated experiments follows a $\chi^2$ distribution $f(t, n_{df})$, $t = \chi^2_{min}$ with $n_{df} = n - m =$number of data points $-$ number of fit parameter
- N.B.: even if you don´t do a least square fit and the property of approaching the $\chi^2$ gets lost, it might be still a useful test depending on the circumstances, see discussion by M. Williams arXiv:1006.3019

# Goodness-of-fit for least squares

Expectation value of the $\chi^2$ distribution is $n_{df}$
$\rightarrow \chi^2 \approx n_{df}$ indicates a good fit

Consistency of a model with the data is quantified with the p-value:
$p - value = \int_{\chi 2min}^{\infty} f(t; n_{df})dt$
The p-value is the probability to get a $\chi^2_{min}$ as high as the observed one, or higher, if the model is correct.
The p-value is **not** the probablity that the model is correct

- straight-forward test and method
  be happy, if it is applicable in your case!
- however, be aware of its applicability bounds (Gaussianity in bins) and limitations (binning sensitivity)

→ selection of different useful methods in arxiv:1006.3019

# Instead of a summary - some advice

- **no one for all**: there is not a single method for everything
  make it as simple as possible and as complicated as needed
- **don´t trust** common wisdom blindly
- **think** about your statistics question and their importance early on
- **read** the literature, **use** established methods and codes if adequate, don´t reinvent the wheel
- **visualise** the (test data for blinded analysis) data as much as possible
- **do** self-consistency checks, bias and convergence tests based on your model (pseudo-experiments)
- **discuss** within your group