

Elements of statistical analysis

Nicolas **Morange**, *IJCLab*

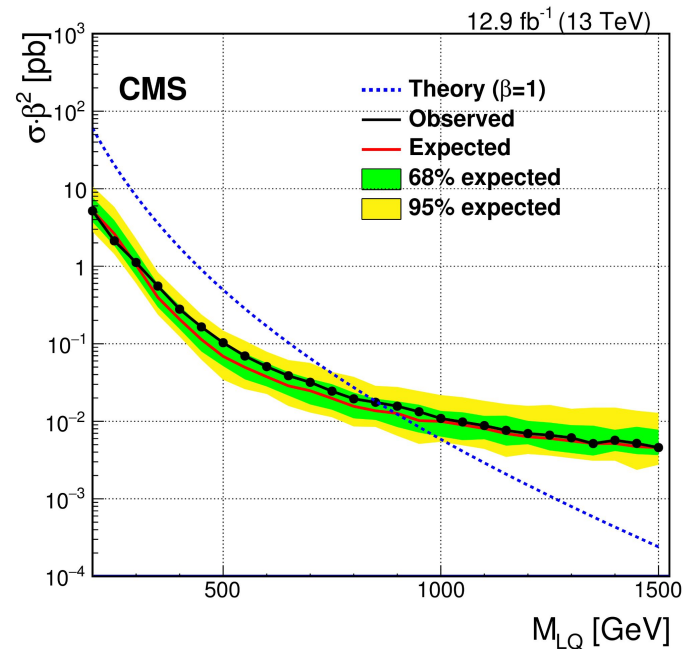
IDPASC, 19/07/2025



Introduction

or: why spending 3 hours to do maths ?

- Use of statistics in HEP is a very broad topic
 - There are complete courses on the topic
 - See bibliography / references
- Goals of the lectures:
 - Teach / remind some basic concepts
 - Focus on aspects used nowadays in the majority of HEP applications (LHC)
 - Understand the main plots often shown in searches or measurements

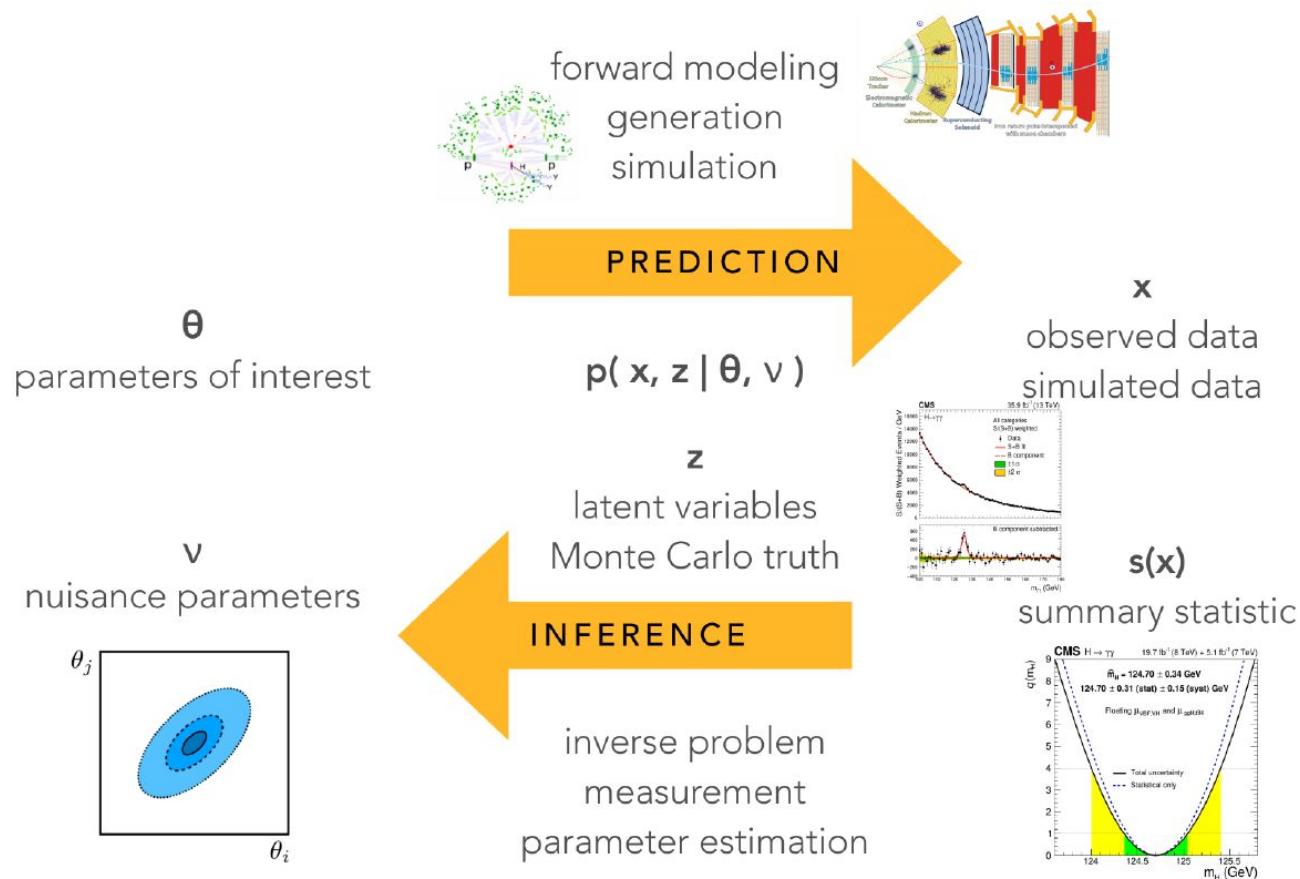


Outline

- Probabilities
- Parameter estimation
- Building a likelihood
- Hypothesis testing
 - Significances
 - Limits

Why probas / stats ?

Goals



Key tasks in statistics

- **Point estimation:** what single “measured” value of a parameter to report ?
 - $m_H = 125.09 \text{ GeV}$
- **Interval estimation:** what confidence interval to report ?
 - $m_H = 125.09 \pm 0.24 \text{ GeV}$
- **Hypothesis testing**
 - Tell apart different models: model selection
 - Is Higgs 0^+ or 0^- ?
 - Test a specific value of a parameter vs any other value
 - Goodness of fit: test a specific model vs anything else
 - Is the data consistent with the SM expectation ?
- **Decision making:** what action should be taken based on the observed data ?
 - Usually based on more or less explicit conventions
 - Ex: The small difference between the measurement and theory is probably a fluctuation, more data is needed.

Bayesian vs Frequentist statistics

Two 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 asymptotics regime
 - There exist nonetheless some important differences (coverage, goodness of fit...)

Bayesian vs Frequentist: take-home messages

“**Bayesians** 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)

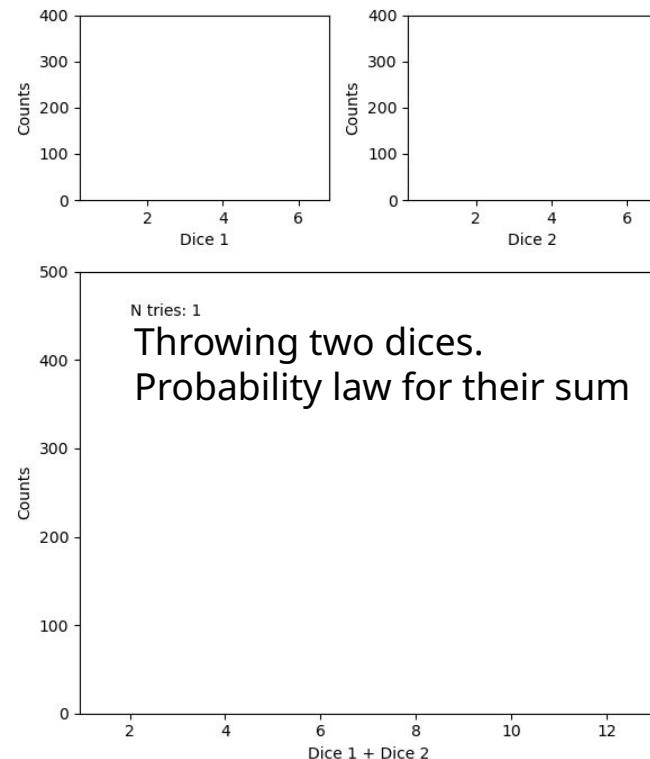
- Communities tend to lean towards one approach
 - Cosmology is mostly using Bayesian statistics (there is only 1 universe...)
 - See Hands-on on cosmology by Thibaut Louis
 - HEP is more frequentist
- Will use frequentist approach in the following
 - By far the most common at the LHC

Probability and random variables

Random Variables

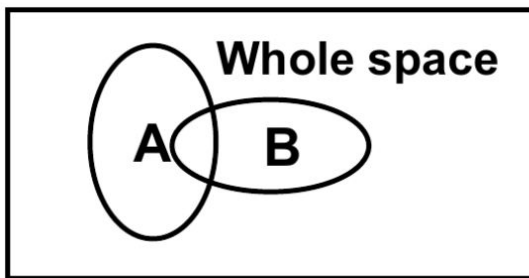
Random variable: a variable that represents the outcome of a random phenomenon.

- Examples: tossing a coin, lifetime of a particle, throwing dices...
- Random variables are usually denoted with a capital letter (e.g. X)
- A function P is a probability function of X if (Kolmogorov axioms):
 - $P(x_i) \geq 0$ for all i
 - $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$
 - $\sum P(x_i) = 1$
- Frequentist probability: $P(A) = \lim_{N \rightarrow \infty} n/N$
- **Probability distribution** of the random variable



Bayes' theorem

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



$$P(A) = \frac{\text{Area of A}}{\text{Area of Whole space}}$$

$$P(B) = \frac{\text{Area of B}}{\text{Area of Whole space}}$$

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

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

$$P(A \cap B) = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}}$$

$$P(A) \times P(B|A) = \frac{\text{Area of A}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of A}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

$$P(B) \times P(A|B) = \frac{\text{Area of B}}{\text{Area of Whole space}} \times \frac{\text{Area of A} \cap B}{\text{Area of B}} = \frac{\text{Area of A} \cap B}{\text{Area of Whole space}} = P(A \cap B)$$

Useful example

A muon detection experiment measures:

- $P(\text{muon ID} | \text{muon})$, i.e., efficiency for tagging muons
- $P(\text{muon ID} | \text{not a muon})$, i.e., efficiency for background
- $P(\text{no muon ID} | \text{muon}) = 1 - P(\text{muon ID} | \text{muon})$
- $P(\text{no muon ID} | \text{not a muon}) = 1 - P(\text{muon ID} | \text{not a muon})$

Question: Given a selection of particles identified as muons, what fraction of them is muons? I.e., what is $P(\text{muon} | \text{muon ID})$?

Answer: Cannot be determined from the given information ! Need in addition: $P(\text{muon})$, the true fraction of all particles that are muons.

Then Bayes' theorem inverts the conditionality:

$$P(\text{muon} | \text{muon ID}) = P(\text{muon ID} | \text{muon})P(\text{muon}) / P(\text{muon ID})$$

Useful example, contd.

- $P(\text{muon ID} | \text{muon})$ is the **efficiency** for tagging muons
- $P(\text{muon} | \text{muon ID})$ is the **purity** of a sample of particles identified as muons

⇒ helpful to keep in mind when you are in doubt

⇒ avoid making the logical error of thinking $P(A | B) = P(B | A)$.

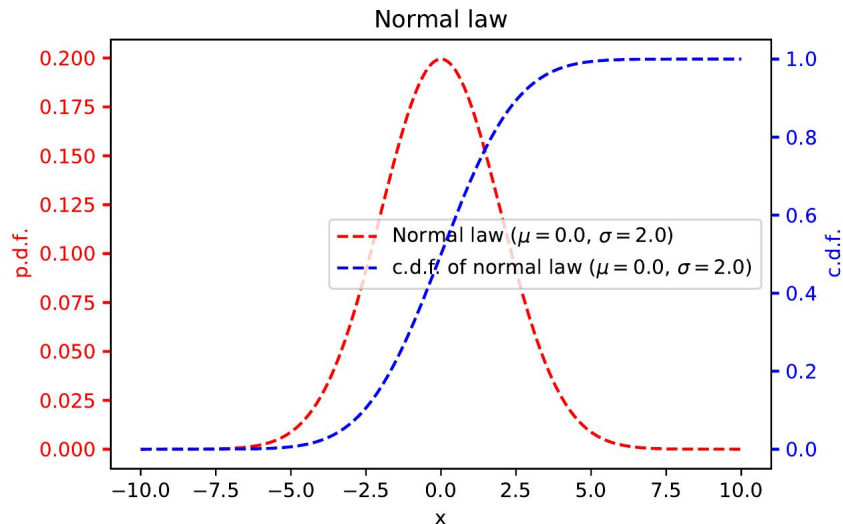
Probability density function

For **continuous** random variables, the **probability density function (pdf)** f is defined by:

$$f(x) dx = P(X \in [x, x + dx])$$

It is related to the cumulative function:

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



Quantiles

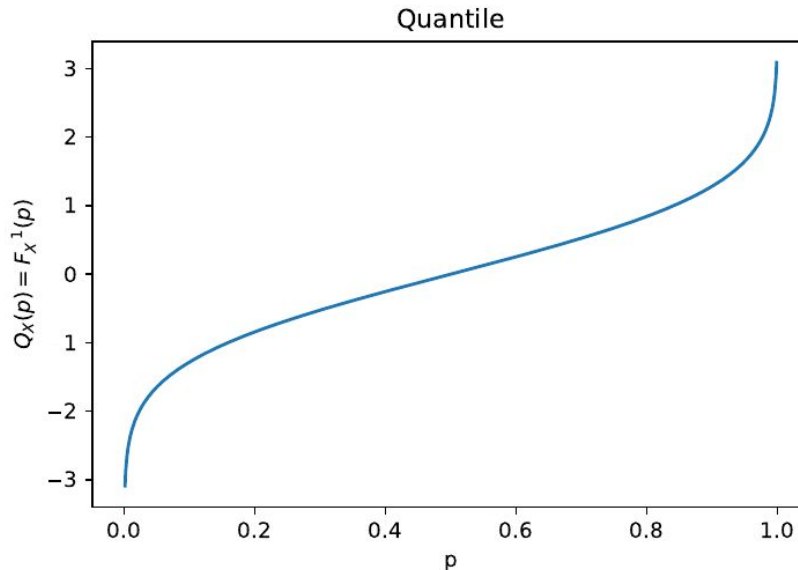
The **quantile** x_α is the value of the random variable x at which the cumulative distribution is equal to α . It is the inverse of the cumulative distribution function:

$$x_\alpha = F^{-1}(\alpha)$$

Special case:

- **Median:** $x_{\text{med}} = x_{1/2}$

Quantiles of the Normal law:



- **Mean μ**

$$\mu = E(X) = \int x f(x) dx$$

- **Variance V , standard deviation σ**

$$V = \sigma^2 = E((X - \mu)^2) = E(X^2) - \mu^2$$

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

- **skew** (n=3): measures left-right asymmetry of the pdf
- **kurtosis** (n=4): measures the size of the tails of the distribution (if positive, then larger tails than a Gaussian).

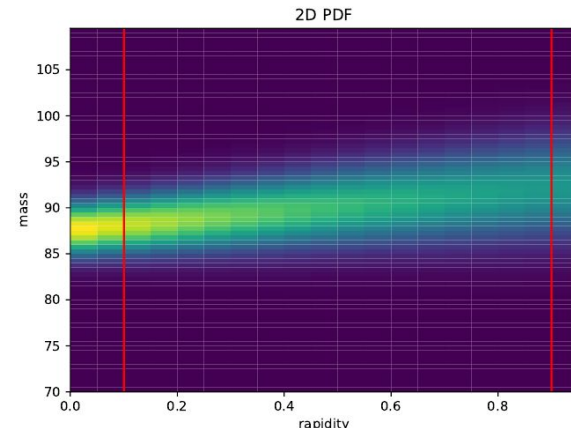
Multi-dimensional case

Example: 3 variables x, y, z , with a joint pdf f

A **marginal pdf** is defined as:

$$f_X(x) = \int dy' dz' f(x, y', z')$$

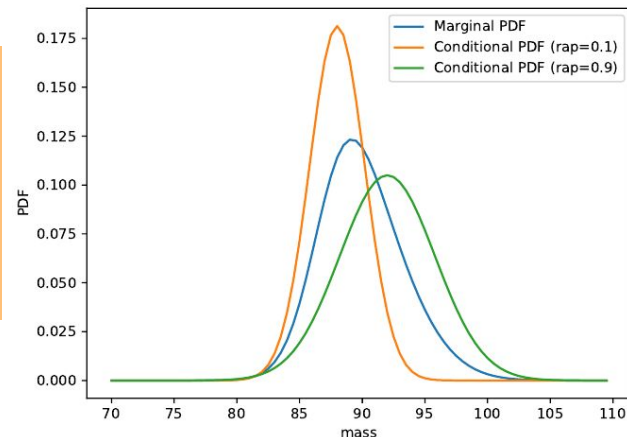
f_X is a **projection** of f . The other variables are **integrated**.



A **conditional pdf** is defined as:

$$f_C(x, y_0, z_0) = \frac{f(x, y_0, z_0)}{\int dx' f(x', y_0, z_0)}$$

f_C is a **slice** of f



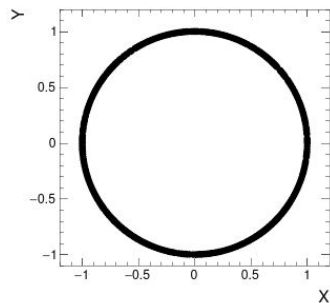
Independence and correlation

Two variables X and Y are **independent** iff $f(x,y) = f_X(x)f_Y(y)$

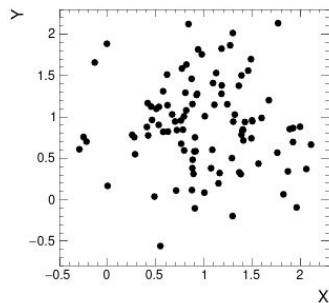
Correlation coefficient between two variables X and Y :

$$\rho(X, Y) = \frac{C(X, Y)}{\sigma_x \sigma_y}$$

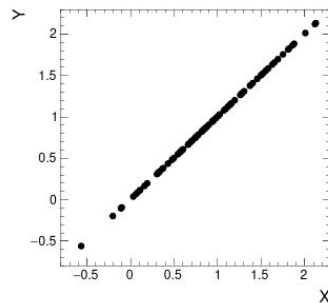
with $C(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y)$



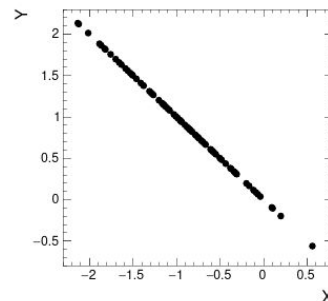
$\rho = 0$,
not independent



$\rho = 0$,
independent



$\rho = 1$,
complete correlation



$\rho = -1$,
complete
anti-correlation

Independent
 $\Rightarrow \rho = 0$

The opposite is
not true

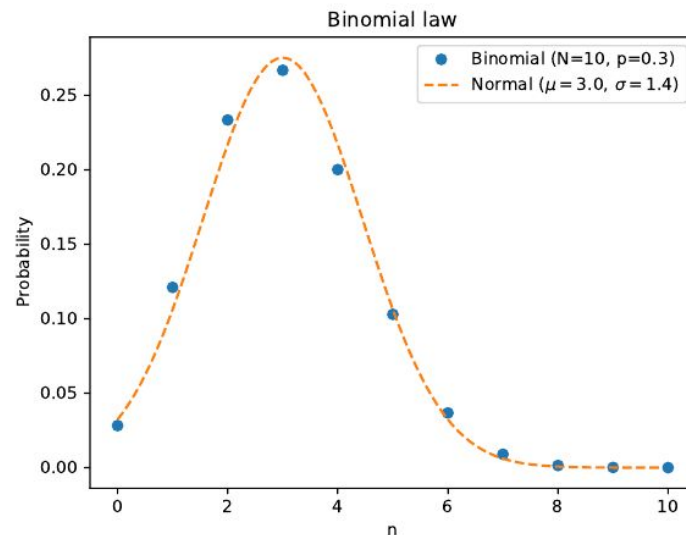
Binomial law

A random process of probability of success p is repeated N times.

The number of successes n follows a **binomial distribution**:

$$P(n) = B(n|N, p) = \binom{N}{n} p^n (1 - p)^{N-n}$$

- Mean $E(n) = pN$
- Variance $V(n) = Np(1-p)$
- Example: out of 1000 collisions, how many will produce a W boson ?
- In the limit of small p and large N , with pN constant, the binomial law converges towards the **Poisson law**



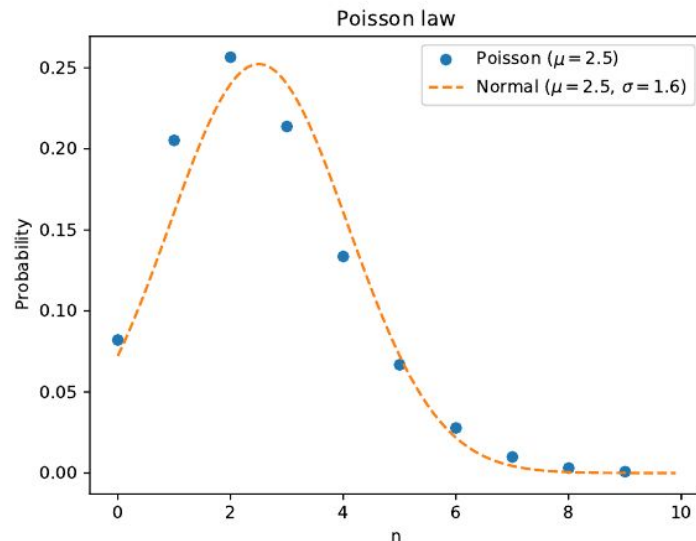
Poisson law

Typical case of random memoryless processes.

The probability to observe n events in a given interval follows a **Poisson distribution**:

$$P_{\mu}(n) = \frac{\mu^n e^{-\mu}}{n!}$$

- Mean $E(n) = \mu$
- Variance $V(n) = \mu$
- $P_{\mu_1} + P_{\mu_2} = P_{\mu_1 + \mu_2}$
- Example: how many Higgs bosons are produced for a luminosity $L = 140 \text{ fb}^{-1}$?
- In the limit of large μ , the Poisson distribution converges towards a **Gaussian distribution**

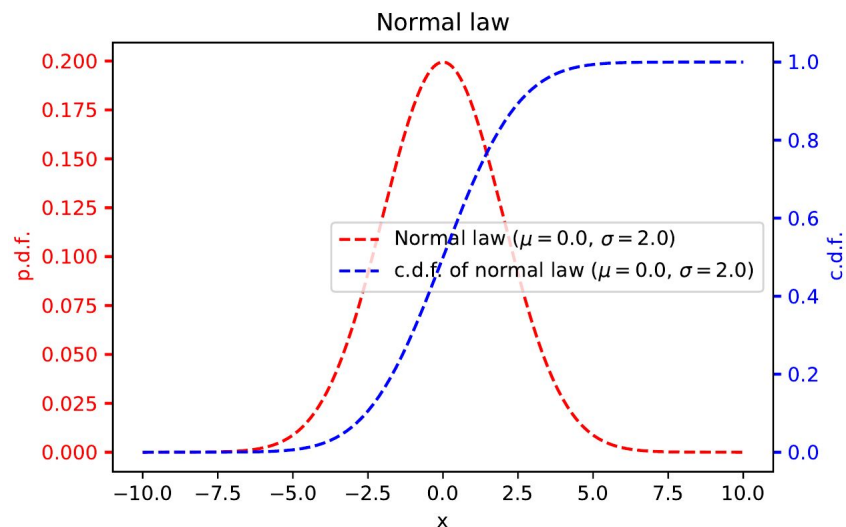


Gaussian distribution

A continuous random variable x follows a **Gaussian distribution** of parameters μ and σ :

$$f_{\mu,\sigma}(x) = G(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean $E(x) = \mu$
- Variance $V(x) = \sigma^2$
- Special case $\mu=0, \sigma=1$ is called the **Normal law**
- Gaussian distributions play a very special role in statistics because of the **Central Limit Theorem**



Normal law: useful numbers

FWHM ²	
$P(-\sigma \leq x - \mu \leq \sigma)$	$\approx 2 \times 1.176\sigma$ ≈ 0.68
$P(-1.64\sigma \leq x - \mu \leq 1.64\sigma)$	0.90
$P(-1.96\sigma \leq x - \mu \leq 1.96\sigma)$	0.95
$P(-3.24\sigma \leq x - \mu \leq 3.24\sigma)$	0.999

Summary of common and useful distributions

Distribution	Probability density function f (variable; parameters)	Characteristic function $\phi(u)$	Mean	Variance
Uniform	$f(x; a, b) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{ibu} - e^{iau}}{(b-a)iu}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Binomial	$f(r; N, p) = \frac{N!}{r!(N-r)!} p^r q^{N-r}$ $r = 0, 1, 2, \dots, N; \quad 0 \leq p \leq 1; \quad q = 1 - p$	$(q + pe^{iu})^N$	Np	Npq
Multinomial	$f(r_1, \dots, r_m; N, p_1, \dots, p_m) = \frac{N!}{r_1! \dots r_m!} p_1^{r_1} \dots p_m^{r_m}$	$(\sum_{k=1}^m p_k e^{iu_k})^N$	$E[r_i] = Np_i$	$\text{cov}[r_i, r_j] = Np_i(\delta_{ij} - p_j)$
Poisson	$f(n; \nu) = \frac{\nu^n e^{-\nu}}{n!}; \quad n = 0, 1, 2, \dots; \quad \nu > 0$	$\exp[\nu(e^{iu} - 1)]$	ν	ν
Normal (Gaussian)	$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp(-(x - \mu)^2 / 2\sigma^2)$	$\exp(i\mu u - \frac{1}{2}\sigma^2 u^2)$	μ	σ^2
Multivariate Gaussian	$f(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V}) = \frac{1}{(2\pi)^{n/2} \sqrt{ \mathbf{V} }}$ $\times \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$ $-\infty < x_j < \infty; \quad -\infty < \mu_j < \infty; \quad \mathbf{V} > 0$	$\exp\left[i\boldsymbol{\mu} \cdot \mathbf{u} - \frac{1}{2}\mathbf{u}^T \mathbf{V} \mathbf{u}\right]$	\mathbf{u}	V_{jk}
Log-normal	$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \frac{1}{x} \exp(-(\ln x - \mu)^2 / 2\sigma^2)$ $0 < x < \infty; \quad -\infty < \mu < \infty; \quad \sigma > 0$	—	$\exp(\mu + \sigma^2/2)$	$\frac{\exp(2\mu + \sigma^2)}{\times [\exp(\sigma^2) - 1]}$
χ^2	$f(z; n) = \frac{z^{n/2-1} e^{-z/2}}{2^{n/2} \Gamma(n/2)}; \quad z \geq 0$	$(1 - 2iu)^{-n/2}$	n	$2n$

Parameter estimation

Parameter estimation ?

- Suppose we have a model, represented by a pdf $f(x | \theta)$
 - x is a **random variable**
 - θ represents **parameters** that affect the shape of the pdf
- Now, let us collect a sample of observed data $x=(x_1, x_2, \dots, x_N)$
- We want to say something of the parameters θ using the data
- An **estimator** is a function of the data (a.k.a 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 are born equal

X is a random variable of pdf $f(x | \theta_0)$, with θ_0 unknown. An estimator t_N of θ_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 minimum variance bound
- **optimal**: if t_N minimises the Mean Square Error (MSE):
$$\text{MSE}(t_N) = V(t_N) + b_N^2$$
- **robust**: if it does not depend on a hypothesis on the pdf

Usual method to build estimators

- **Moments method**
 - aka the sample mean !
- **Maximum likelihood method**
 - today's focus
- **Least squares method**
 - still useful in many occasions

Likelihood function

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

N independent observations of x are obtained: x_1, \dots, x_N

The joint pdf of the N observations is then:

$$P(\mathbf{X} | \theta) = \prod f(x_i | \theta)$$

The likelihood function is this pdf, evaluated with **fixed data \mathbf{X}** and regarded **as a function of the parameters θ** only:

$$L(\theta) = P(\mathbf{X} | \theta)$$

Notes:

- $L(\theta)$ is **not** a pdf for θ . 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 θ is close to the true value, then there is a high probability to get data like the observed one.

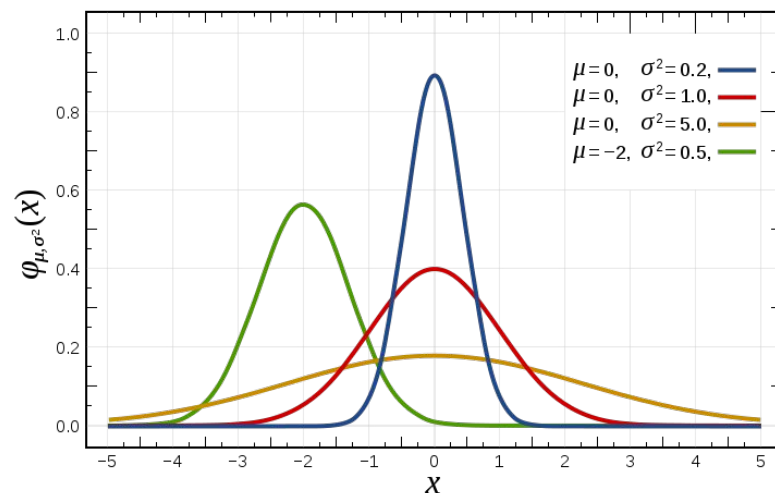
The maximum likelihood (ML) estimator(s) are defined as **the parameter value(s) for which the likelihood is maximum**

- In practice, we usually minimize **$-\ln L(\theta)$** or **$-2\ln L(\theta)$**
- ML estimators are **not** guaranteed to be always unbiased, neither optimal
- **In practice they are very good**: asymptotically unbiased, with a MLE distribution asymptotically Gaussian
- ML estimators are not robust: **the shape of the pdf must be known**

Example: estimation of Gaussian parameters

- Random process following a Gaussian law of unknown mean and variance:

- **Example:** Invariant mass distribution of $Z \rightarrow e^+e^-$
- **Parameters:** $\theta \mapsto \mu$ mean, σ standard error
- **Observables:** x_i
- **PDF:**
 $f \mapsto G(x \mid \mu, \sigma) = 1/\sqrt{2\pi\sigma^2} \exp(- (x-\mu)^2 / 2\sigma^2)$



Parameter estimation

- 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:

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

- which yields:

$$\frac{\partial -\log \mathcal{L}(x_i|\mu, \sigma)}{\partial \mu} = 0$$

$$\frac{\partial -\log \mathcal{L}(x_i|\mu, \sigma)}{\partial \sigma} = 0$$



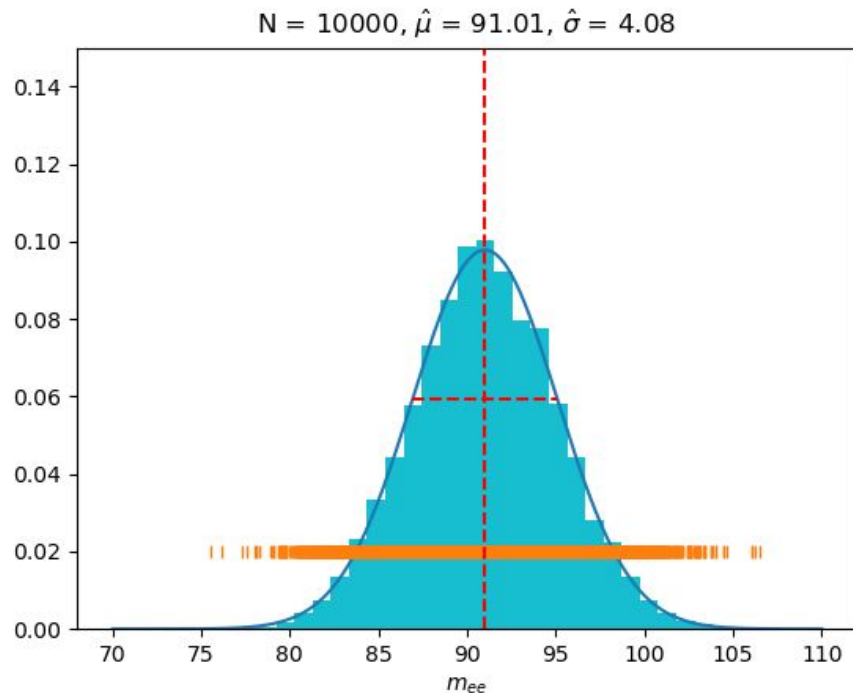
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i$$

Sample mean !

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

Biased estimator !
(but asymptotically unbiased)

Live example



- When adding data
 - $\hat{\mu}$ converges to $\mu=91$
 - $\hat{\sigma}$ converges to $\sigma=4$
 - Uncertainty in the estimate decreases as well
- Maximisation of likelihood function
 - Analytical calculation here
 - Usually relying on numerical minimisers: **Minuit**

Coverage probability and confidence level

Beyond parameter estimation: **uncertainty in the parameter**

- 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 the coverage probability
 - $1 - \alpha$ is called the confidence level
 - An ensemble of statements that obeys the FP is said to cover
- Application to confidence intervals: if we report a confidence interval I and we repeat the experiment N times, then a fraction f of the intervals I will contain the true value of the parameter

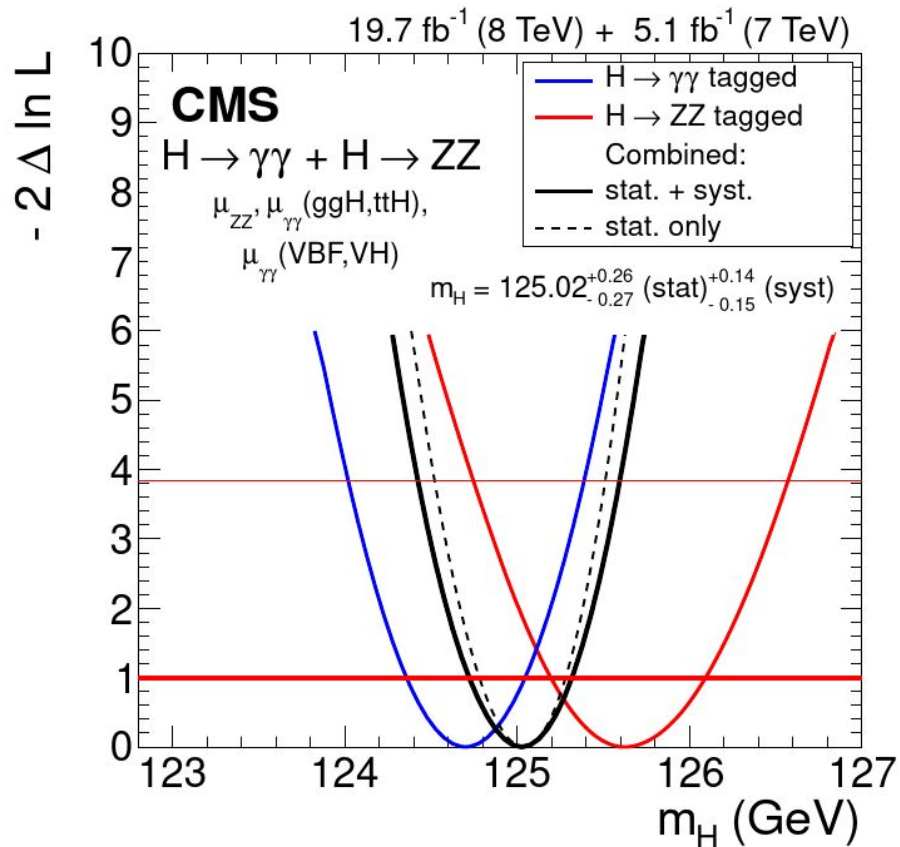
Confidence intervals for ML estimators

- Finding procedures that give correct coverage (i.e neither undercoverage nor overcoverage) is in general 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**: $-2 \ln L(\theta)$ asymptotically follows a χ^2 law with d degrees of freedom, where d is the dimensionality of θ
 - **Consequence**: Confidence intervals can be obtained from the inverse quantiles of a χ^2 distribution with d degrees of freedom: the so-called **likelihood intervals**

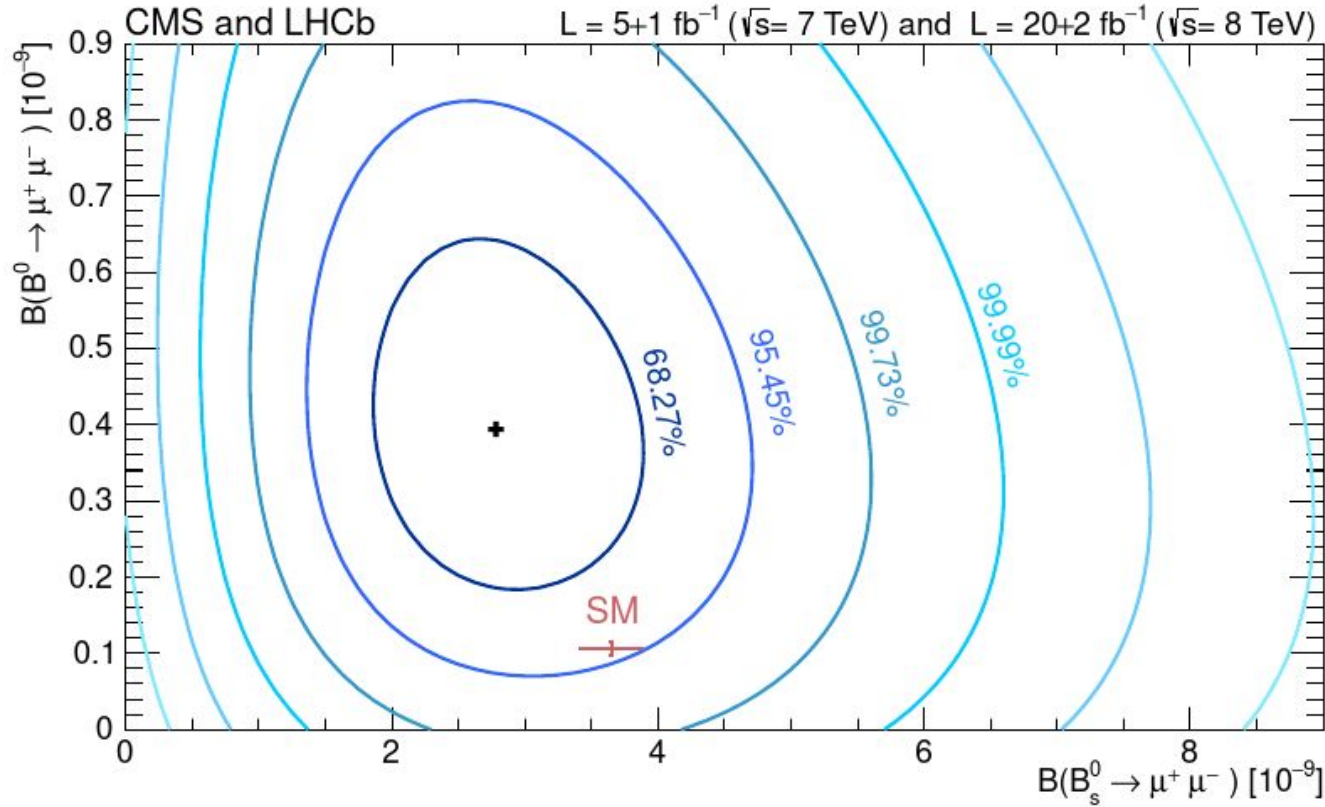
Values of $\Delta\chi^2$ or $-2\Delta \ln L$ corresponding to a coverage probability $1-\alpha$ in the large data sample limit, for d parameters.

1- α (%)	$\Delta\chi^2$		
	$d = 1$	$d = 2$	$d = 3$
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



Example in 2D: measurement of $B_s \rightarrow \mu\mu$ and $B_0 \rightarrow \mu\mu$



Contours: $\Delta \ln(\mathcal{L}) = 2.30, 6.18, 11.83, \dots$

Building a likelihood

Likelihoods in collider experiments

- Observables are **numbers of events**
 - After selections, in categories, bins...
 - Due to the nature of collisions (independent), they obey **Poisson laws**

- Simplest case: **number counting experiment**

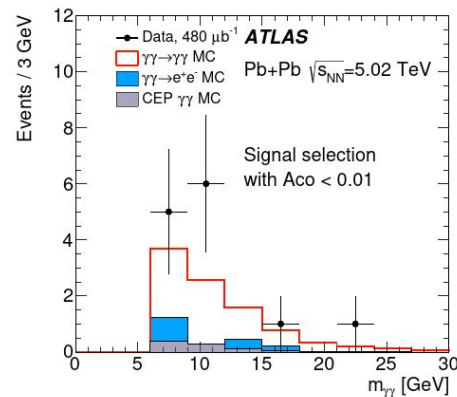
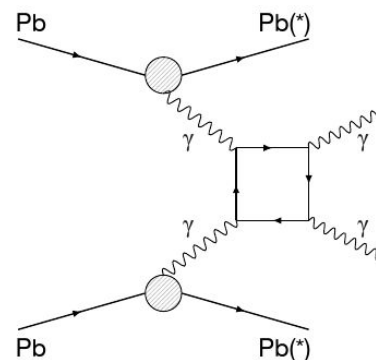
- D observed events
- s expected signal events (parameter of interest)
- b expected background events (known)

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

- What is the MLE of s ?

Example: light-by-light scattering

- D = 13
- $b = 2.6 \pm 0.7$
- $s = 10.4$



Nature Phys. 13 (2017) no. 9, 852-858

Likelihoods in collider experiments

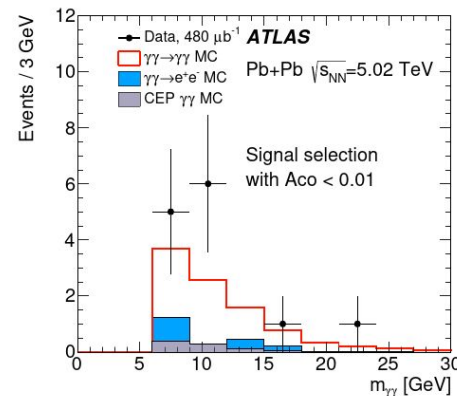
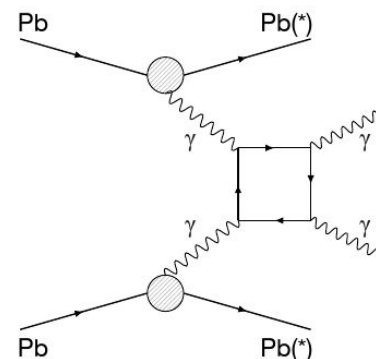
- Observables are **numbers of events**
 - After selections, in categories, bins...
 - Due to the nature of collisions (independent), they obey **Poisson laws**
- Simplest case: **number counting experiment**
 - D observed events
 - s expected signal events (parameter of interest)
 - b expected background events (known)

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

- What is the MLE of s ?
 - $L(s) = p(D|s, b)$
 - $dL/ds = 0 \Rightarrow s = D - b$

Example: light-by-light scattering

- D = 13
- $b = 2.6 \pm 0.7$
- $s = 10.4$



Nature Phys. 13 (2017) no. 9, 852-858

Extension: multiple analysis regions, multiple bins

In practice, almost all analyses have more than one observable

- **Signal strength μ** : often used as the main **parameter of interest**

- $\mu = \sigma/\sigma_{SM}$

- Likelihood is a product of Poisson:

$$L(n_1, \dots, n_{Nbins} | \mu) = \prod_{i=1}^{Nbins} P(n_i | \mu s_i + b_i)$$

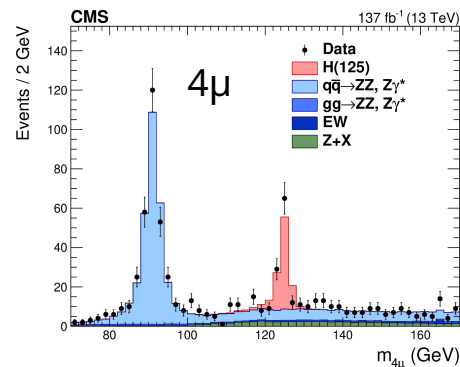
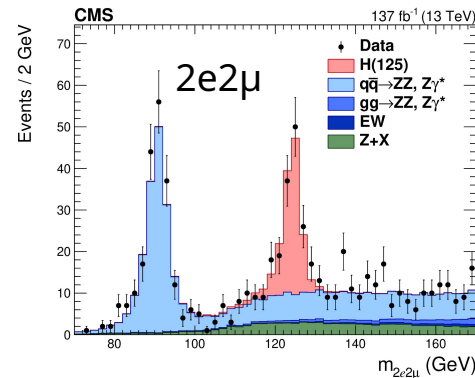
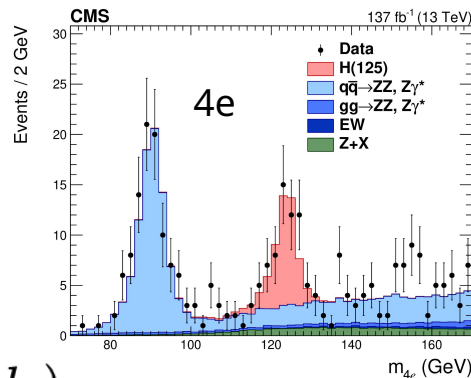
expected numbers of events in bin i

- Special cases:

- $N=1$: counting experiment
- $N=\infty$: unbinned analysis (s_i and b_i become pdf values)

- In practice, binned pdfs are often used

- Software to do it easily: **HistFactory**



Dealing with uncertainties: nuisance parameters

- In a realistic model, the expectations for s_i and b_i are uncertain
 - Affected by systematic uncertainties
- This uncertainty can be added to our likelihood model as new parameters affecting the shape of the pdfs
 - $b_i \rightarrow b_i(\theta)$, $s_i \rightarrow s_i(\theta)$
 - They are new parameters of the likelihood: $L(\mu) \rightarrow L(\mu, \theta)$
 - But they are of no interest for our measurement: **nuisance parameters (NP)**
- Often we do have additional knowledge on these parameters
 - Ex: Background estimation performed in a dedicated control region (with some uncertainty)
 - Ex: Luminosity calibrated in a dedicated measurement (with some uncertainty)
 - This knowledge should be incorporated into the likelihood
 - Factorizable: $\mathbf{p}(\mathbf{n}_i | \mu) \rightarrow \mathbf{p}(\mathbf{n}_i, \mathbf{y}_j | \mu, \theta) = \text{Poiss}(\mathbf{n}_i | \mu, \theta) \times \text{pdf}(\mathbf{y}_j | \theta)$
 - For the likelihood: $\mathbf{L}(\mu, \theta) = \mathbf{L}_{\text{meas}}(\mu, \theta) \times \mathbf{C}(\theta)$
 - ↑ data bins
 - ↑ constraint terms
 - ↑ auxiliary measurements

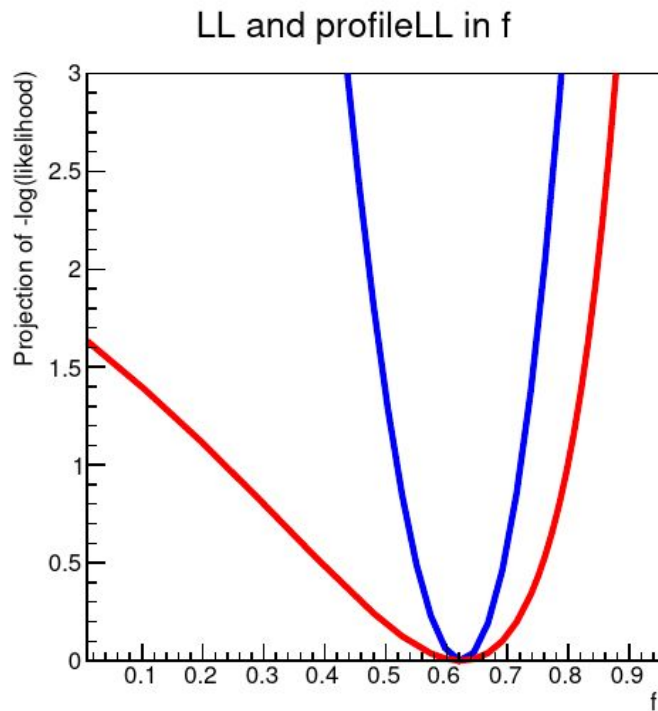
Constraint terms are most often **Normal (Gaussian)**, but other distributions sometimes used (log-normal)

Profile likelihood

- Nuisance parameters are not of interest for our measurement
- **Profile likelihood:**

$$L_p(\mu) = L(\mu, \hat{\theta})$$

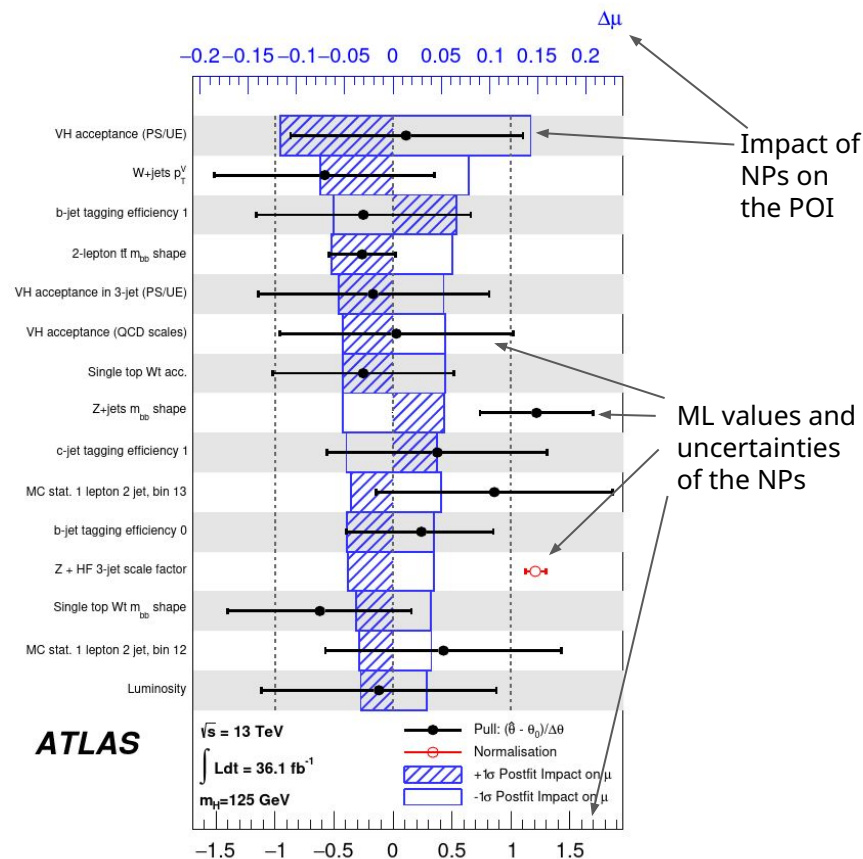
- $\hat{\theta}$ are the values of the parameters θ that maximise the likelihood for a given value μ of the parameters of interest
- $L_p(\mu)$ is a function of the POI only: for each μ , new values of $\hat{\theta}$ are obtained.
- It reduces the dimensionality of L to that of μ (usually 1-d)



Working with profile likelihoods I

Profile likelihoods are complex objects !

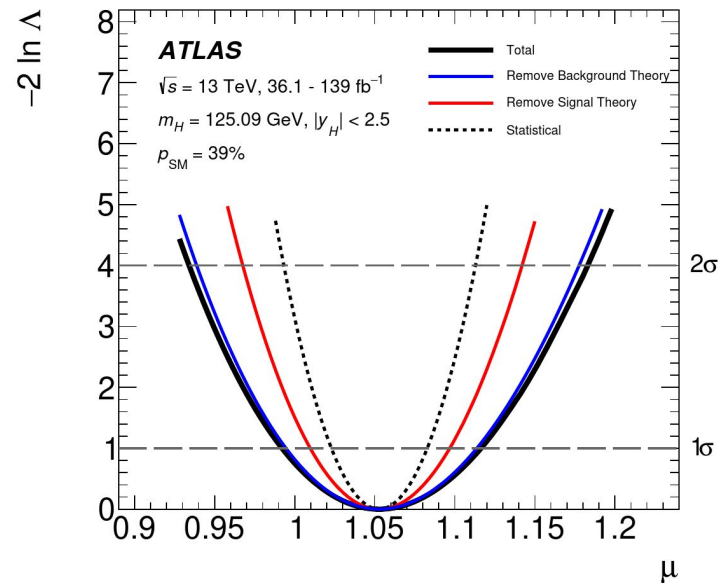
- Working with >200 NPs is commonplace
- ML estimators make sense and have good properties, **assuming the pdf** (i.e the model) **is correct** !
- Great care should be taken to ensure this is the case
 - Goodness-of-fit tests
 - Do the ML fitted values for the NPs make sense (**pulls**) ?
 - Do the ML uncertainties for the NPs make sense (**constraints**) ?
- Keep track of how NPs affect the estimated POI (**impact**)



Working with profile likelihoods II

Error decomposition:

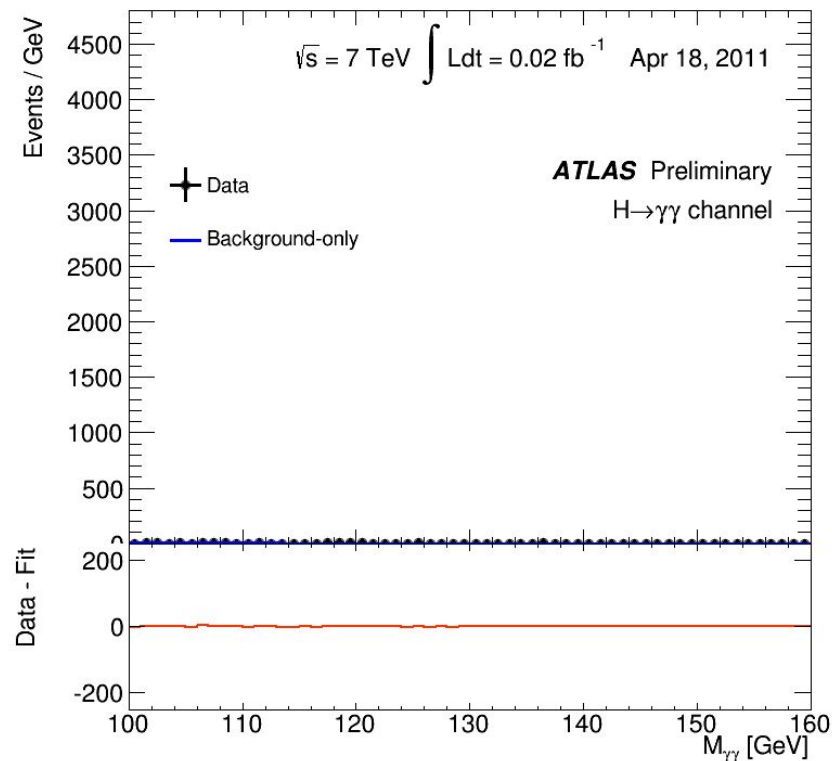
- **Total uncertainty** comes from the profile likelihood scan: $L_p(\mu) = L(\mu, \hat{\theta})$
- **Statistical** comes from a scan where all NPs are set to their best fit value $\hat{\theta} = \hat{\theta}(\hat{\mu})$:
$$L(\mu, \hat{\theta}) = L(\mu, \hat{\theta}(\hat{\mu}))$$
- Other curves are intermediate cases where some NP are profiled while others are set to their best fit value in the scan
 - Allows to estimate the fraction of the total uncertainty coming from some NPs



Hypothesis testing

Motivation: July 2012

- Analysis ready: take data and wait
- We “see” a bump at 125 GeV:
 - Is that a discovery ?
 - How do we quantify it ?



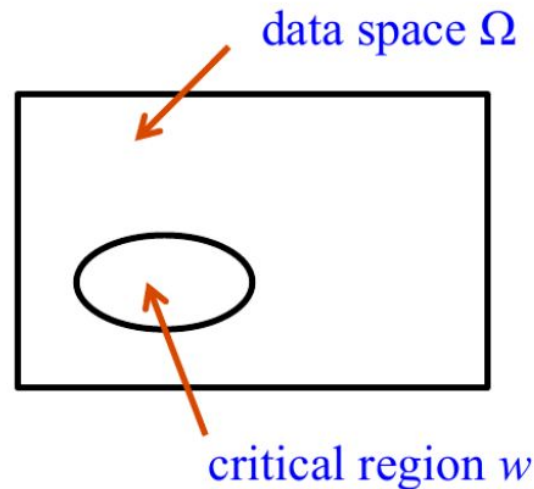
Hypothesis testing

Types of hypotheses

- **Simple hypothesis**: fully specified, including parameter values
 - eg: $H_0 = \text{Higgs is } 0^+$ vs $H_1 = \text{Higgs is } 0^-$
- **Composite hypothesis**: ensemble of simple hypotheses
- **Nested hypotheses**: most common case for searches
 - $\mu = 0$ (background-only) vs $\mu > 0$ (new physics signal !)
 - $\mu = 1$ (SM expectation) vs $\mu \neq 1$ (SM is broken !)

Two ingredients for a hypothesis test

- A **test statistic** $t(x)$
- A **critical region** w such the hypothesis H_0 is false (with a given probability) if t in w
 - The alternative hypothesis is named H_1



Power and size of a test

- **Size of the test (level of significance):**

$$\alpha = P(t \in w | H_0 \text{ true})$$

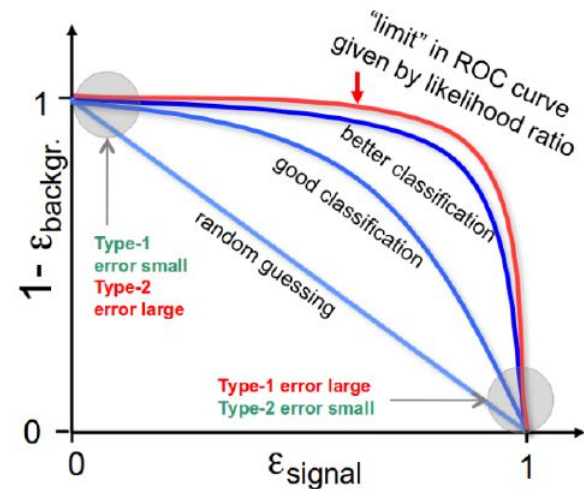
- Also called error of the first kind (Type I error)
- “false discovery claim”: probability of rejecting H_0 when it is true

- **Power of the test: $1 - \beta$ with $\beta = P(t \notin w | H_1)$**

- Also called error of the second kind (Type II error)
- Probability of not claiming a discovery when there is one

	$P(t \notin w)$	$P(t \in w)$
H_0 true	$1 - \alpha$	α
H_1 true	β	$1 - \beta$

There is a **tradeoff** between Type I and Type II errors



Example: muon ID experiment

A muon detection experiment measures:

- $P(\text{muon ID} \mid \text{muon})$, i.e., efficiency for tagging muons
- $P(\text{muon ID} \mid \text{not a muon})$, i.e., efficiency for background
- $P(\text{no muon ID} \mid \text{muon}) = 1 - P(\text{muon ID} \mid \text{muon})$
- $P(\text{no muon ID} \mid \text{not a muon}) = 1 - P(\text{muon ID} \mid \text{not a muon})$

Hypotheses:

- H_0 : not a muon
- H_1 : muon

Then:

- Size of the test $\alpha = P(\text{muon ID} \mid \text{not a muon})$
- Power of the test $\beta = P(\text{no muon ID} \mid \text{muon})$

Neyman-Pearson lemma

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

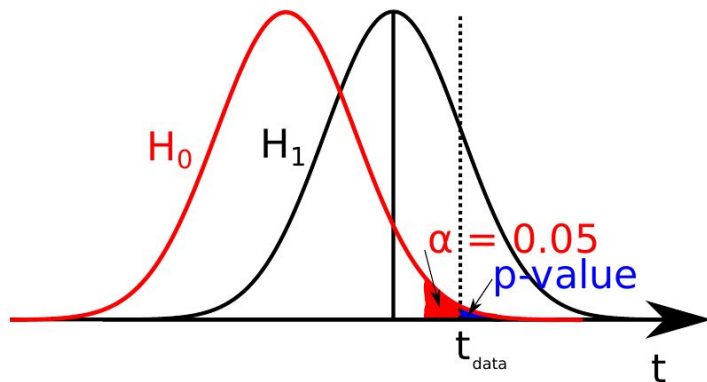
$$t(x) = \frac{L(x|H_1)}{L(x|H_0)}$$

Notes:

- **Optimal**: minimizes Type II error for a given Type I level of significance
- Valid for any **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 extremely well for our nested hypotheses**

Procedure for hypothesis testing

- Specify the null and the alternate hypotheses
 - Ex: H_0 = SM background, H_1 = BSM signal
- Build a test statistic $t(x)$ using e.g Neyman-Pearson lemma
- Specify the significance of the test (what we accept as a false discovery rate)
 - Ex: $2.9 \cdot 10^{-7}$ (5σ) for discovery
 - 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: $p_0 = p(t \geq t_{\text{obs}} \mid H_0)$
 - Significance level of the test α : chosen prior to look at the data
 - p-value: interesting quantity to compute when looking at the data

- Interpretation:

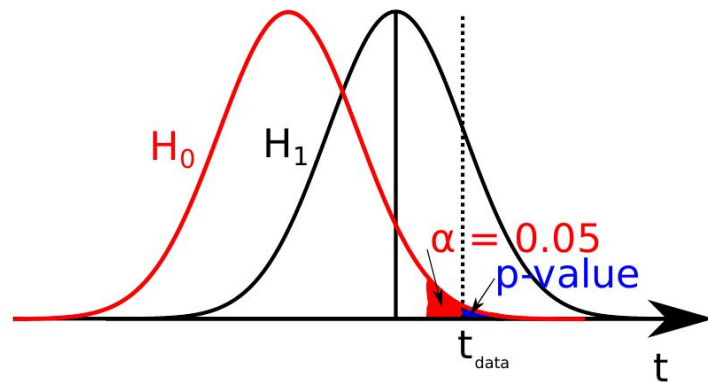
- probability for the test statistic t to be larger than the observed one t_{obs} , under the null hypothesis H_0
- **NOT** “the probability that H_0 is true”

- “Significance” in number of sigmas:

- translation of the p-value using the integral in one tail of a Gaussian

$$p_0 = \int_Z^{\infty} G(x|0, 1)dx = 1 - \Phi(Z)$$

- **Convention: 3σ is evidence, 5σ is discovery**



z-value (σ)	p-value
1.0	0.159
2.0	0.0228
3.0	0.00135
5.0	2.87×10^{-7}

Profile likelihood ratio and asymptotics formulae

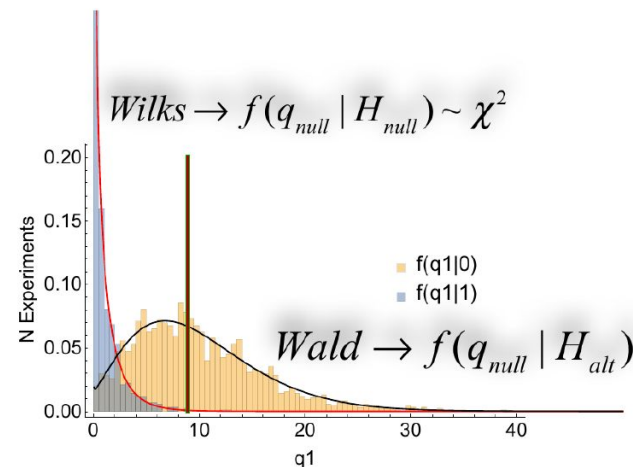
- At the LHC, to deal with systematics, the basis of test statistics used for hypothesis testing is the **Profile Likelihood Ratio** (PLR):

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})}$$

- Then the test statistic for discovery is:

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

- Asymptotics properties** of the PLR make it easy to work with:
 - Wald's approximation, Wilks' theorem
 - Cowan, Cranmer, Gross and Vitells, EPJC 71 (2011) 1554
 - Median expected properties from the **Asimov dataset**
 - No need for CPU intensive toys !
 - $Z = \sqrt{-q_0} = \sqrt{-2 \ln \lambda(0)}$



p-value in counting experiments

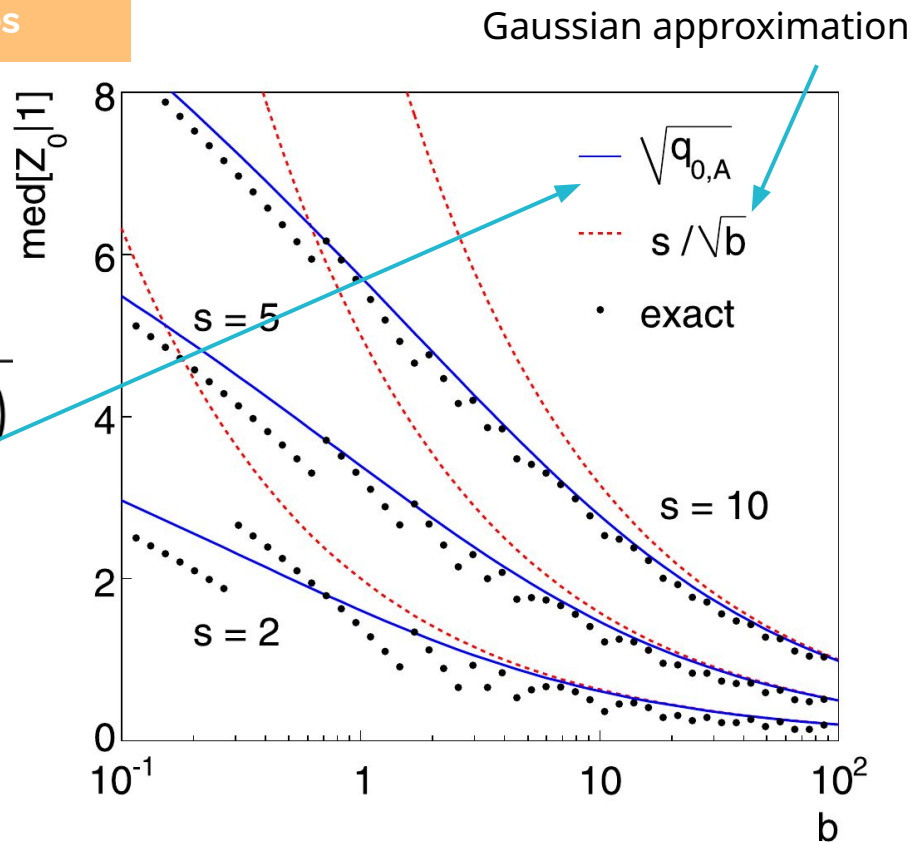
Very useful formula for e.g optimization studies

- n observed events, b background
 - $n = \mu \cdot s + b$

$$L(\mu) = \frac{(\mu s + b)^n e^{-(\mu s + b)}}{n!}$$

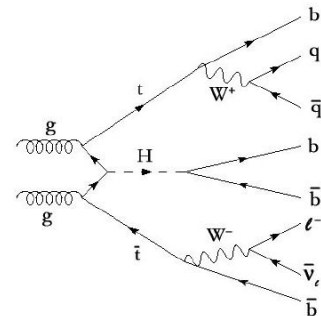
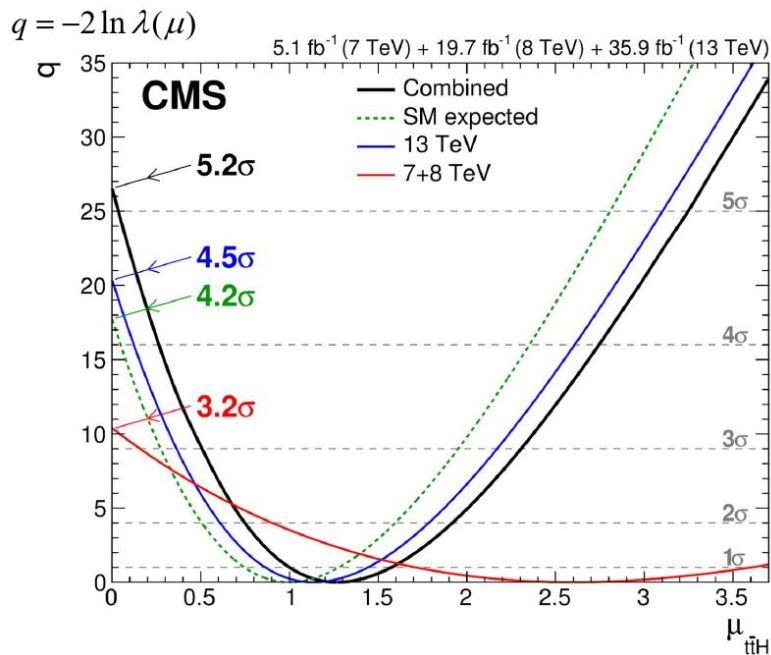
$$Z = \sqrt{-2 \ln \frac{L(0)}{L(\hat{\mu})}} = \sqrt{2 \left(n \ln \left(1 + \frac{\hat{\mu} s}{b} \right) - \hat{\mu} s \right)}$$

$$Z = \sqrt{2 \left(n \ln \left(\frac{n}{b} \right) + b - n \right)}$$



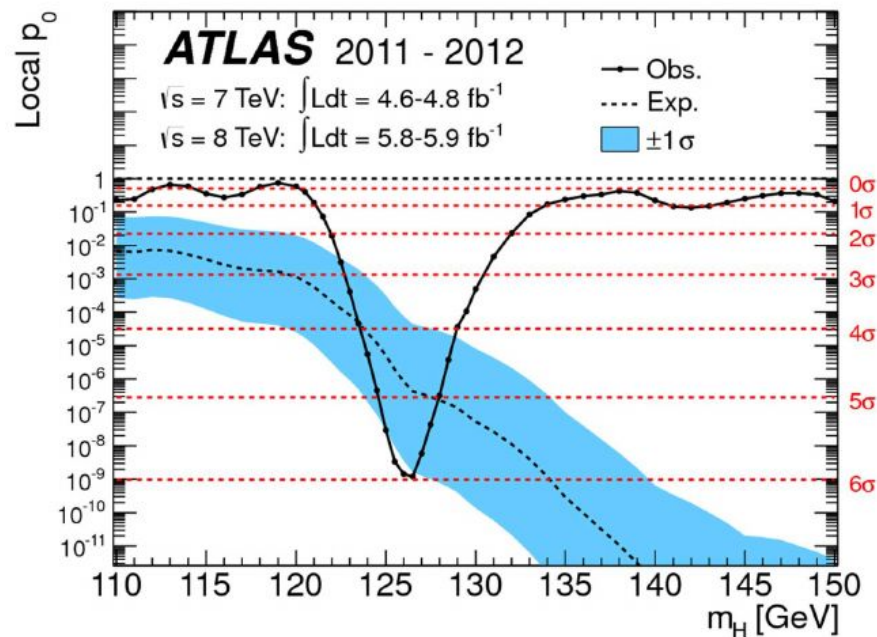
Example: ttH observation

p-value (transformed as a significance) can be directly read on the y axis:
 $q_0 = q(\mu=0)$



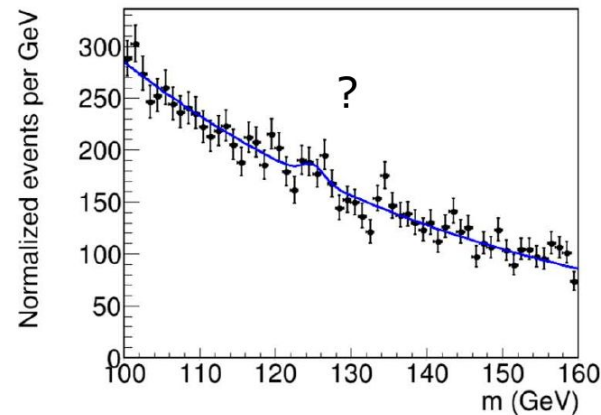
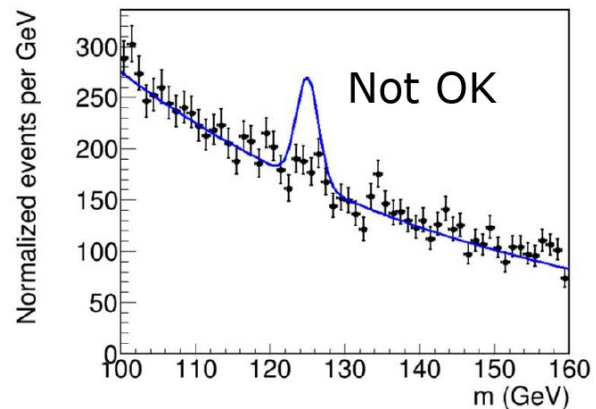
Example: Higgs boson discovery

- Each Higgs mass hypothesis is scanned independently
- For each mass:
 - **Observed**: p-value observed in data
 - **Expected**: median of the p-value expected in the presence of the SM Higgs boson
 - Blue band: interval containing 68% of the p-values under SM Higgs hypothesis
- “Local” p_0
 - Many mass points scanned
 - **Look-elsewhere effect**: global p_0 to correct for number of trials



Exclusion limits

- Similar procedure to discovery case, but hypotheses are inverted:
 - H_0 : signal + background hypothesis
 - H_1 : background-only hypothesis
- Goal: disprove H_0 by estimating the probability of downward fluctuation of s+b
- Size of test less stringent: $\alpha = 0.05$
 - 95% CL limits
- Upper limit: minimal signal strength for which H_0 can be excluded at 95% CL



Test statistics for exclusion

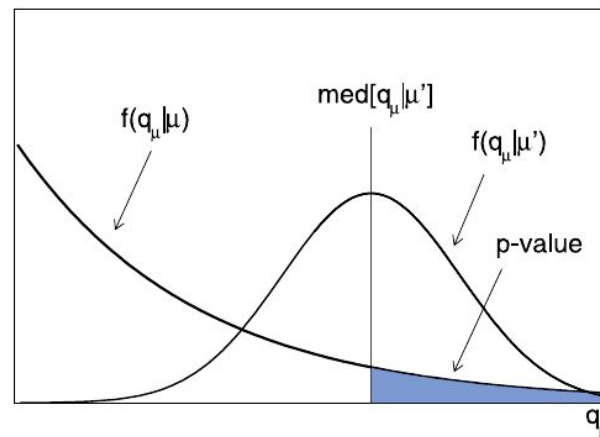
- Still using PLR-based test statistic:

$$q_{\mu} = \begin{cases} -2 \ln \lambda(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases}$$

- NB: one does not regard an upwards fluctuation of the data as representing incompatibility with the hypothesized μ

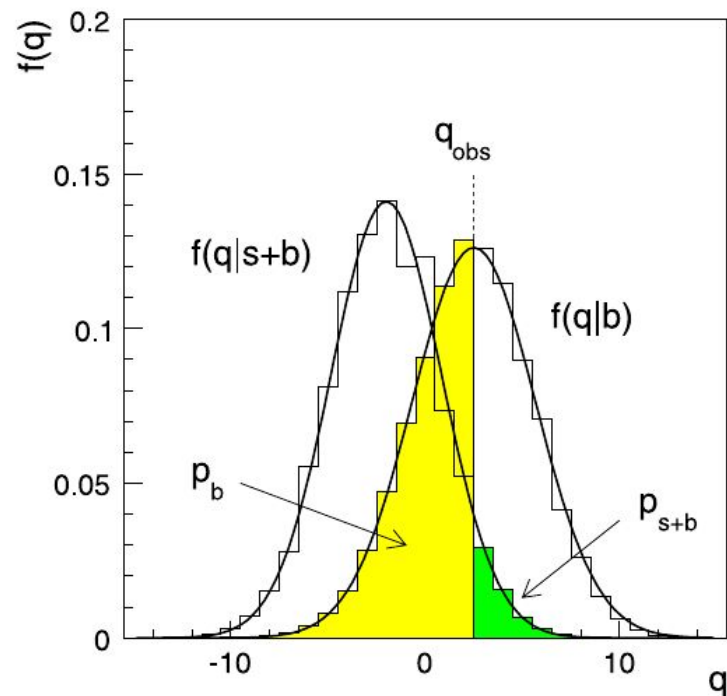
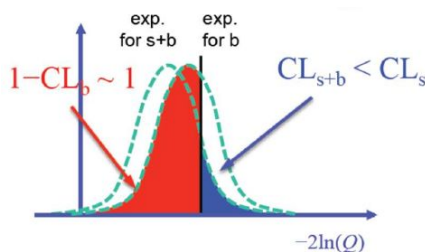
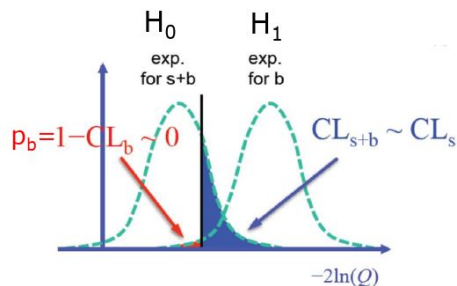
$$p_{\mu} = \int_{q_{\mu, \text{obs}}}^{\infty} f(q_{\mu} | \mu) dq_{\mu}$$

- In this case as well asymptotics formulae exist for the different distributions
 - Fairly quick computation of limits



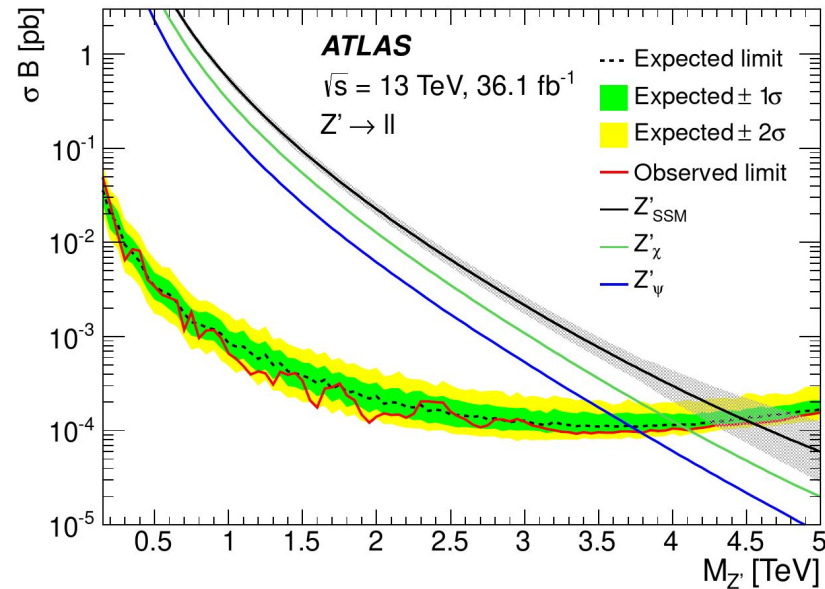
CL_s correction

- Potential problem when setting limits: spurious exclusion when there is a downward fluctuation of the data even wrt the background-only hypothesis
- Definition: $CL_s = \frac{p_\mu}{1-p_b} = \frac{CL_{s+b}}{CL_b}$
 - If the two distributions are well separated, small change wrt CL_{s+b}
 - If distributions are close, prevents spurious exclusion



Example: search for high mass dilepton resonances

- For each hypothesized $M_{Z'}$ value, compute:
 - Expected limit:** **median** value of upper limit under bkg-only hypo.
 - Expected $\pm 1\sigma$:** interval containing 68% of the upper limit values under bkg-only
 - Expected $\pm 2\sigma$:** interval containing 95% of the upper limit values under bkg-only
 - Observed limit:** upper limit obtained using data actually observed
 - Theoretical curves:** often superimposed. Crossing point gives lower limit on the Z' mass for the given model
- All expected limits can be obtained with asymptotics formulae
 - At very high masses, very low number of events: good practice to cross-check limit with toys



Conclusions

- **Statistical analyses rely on likelihood functions:**
 - Parameters of interest we want to measure (cross-section, mass...)
 - Other parameters of the model are called nuisance parameters
- **Parameter estimation uses maximum likelihood values as estimators**
 - Asymptotic properties of the likelihoods allows to set easily confidence intervals
- **Hypothesis testing is used to claim discovery or to set limits**
 - Use Profile likelihood ratio-based test statistics
 - Null and alternative hypotheses have to be set appropriately
 - Significance of the test: 0.05 for exclusion, 5σ for discovery, etc...
 - Asymptotic formulas allow to compute limits and significances without the need for massive amount of toy data.

Bibliography

- G. Cowan, Statistical data analysis (Oxford University Press)
- G. Cowan, Statistics for searches at the LHC, arxiv:1307.2487
- G. Cowan, Foundations of statistics, CERN Summer school
- N. Berger, Statistical analysis methods in HEP, LAL Winter lecture, <https://indico.lal.in2p3.fr/event/4738/>
- E. Chapon, Statistics lectures, X/ETHZ Master
- Particle Data Group, Review on statistics (G. Cowan)