FYST17 Lecture 8
Statistics: fitting and hypothesis testing

Thanks to T. Petersen, S. Maschiocci, G. Cowan, L. Lyons

Plan for today:

- Introduction to concepts
  - The Gaussian distribution
- Likelihood functions
- Hypothesis testing
  - Including p-values and significance
- More examples

Interpretation of probability

1. Interpretation of probability as **RELATIVE FREQUENCY**

   (frequentist approach):

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

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

   See quantum mechanics, particle scattering, radioactive decays …

2. **SUBJECTIVE PROBABILITY**

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

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

In particle physics, frequency interpretation often most useful, but subjective probability can provide a more natural treatment of non-repeatable phenomena

(systematic uncertainties, probability that higgs exists ...)

PDF = probability density function

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)

With:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Normalization
(x must be somewhere)

Note:

- $f(x) \geq 0$
- f(x) is NOT a probability ! It has dimension 1/x !

Definitions

Mean or expectation value

$$E[x] = \int x f(x) dx = \mu$$

Variance:

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

Standard deviation:

$$\sigma = \sqrt{\sigma^2}$$

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

Correlation coefficient

$$\rho_{xy} = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y}, \qquad -۱ \leq \rho_{xy} \leq +۱$$

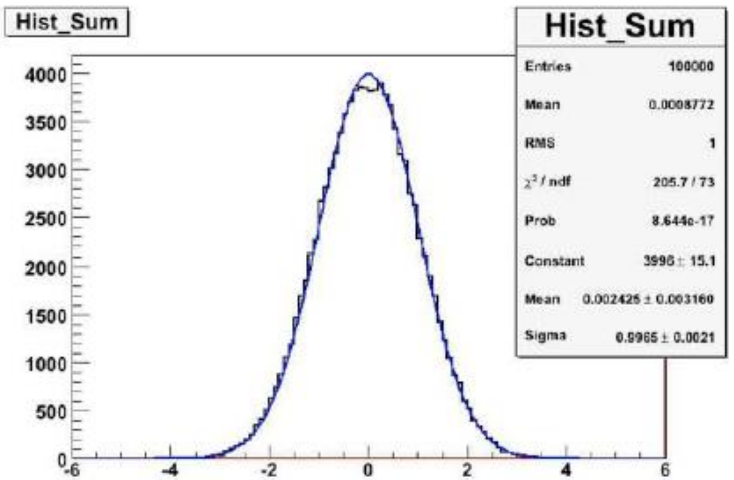
Central limit theorem

What we used already in MC studies

<u>Central Limit theorem:</u>
The sum on N *independent* continuous random variables x_i with means μ_i and variances σ_i^2 becomes a Gaussian random variable with mean $\mu = \Sigma_i \mu_i$ and variance $\sigma^2 = \Sigma_i \sigma_i^2$ in the limit that N approaches infinity

Try for yourselves!

Example: sum of 10 uniform numbers = Gaussian!



Gaussian functions play important role in applied statistics

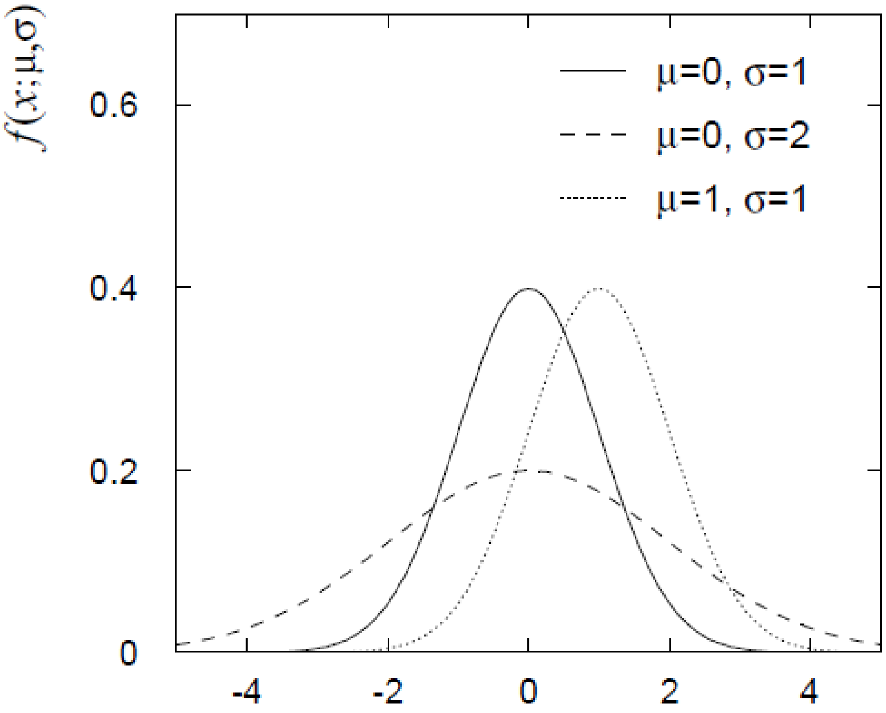
Uncertainties tend to be Gaussian!

Gaussian distribution

The Gaussian pdf is defined by

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

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



"standard Gaussian"

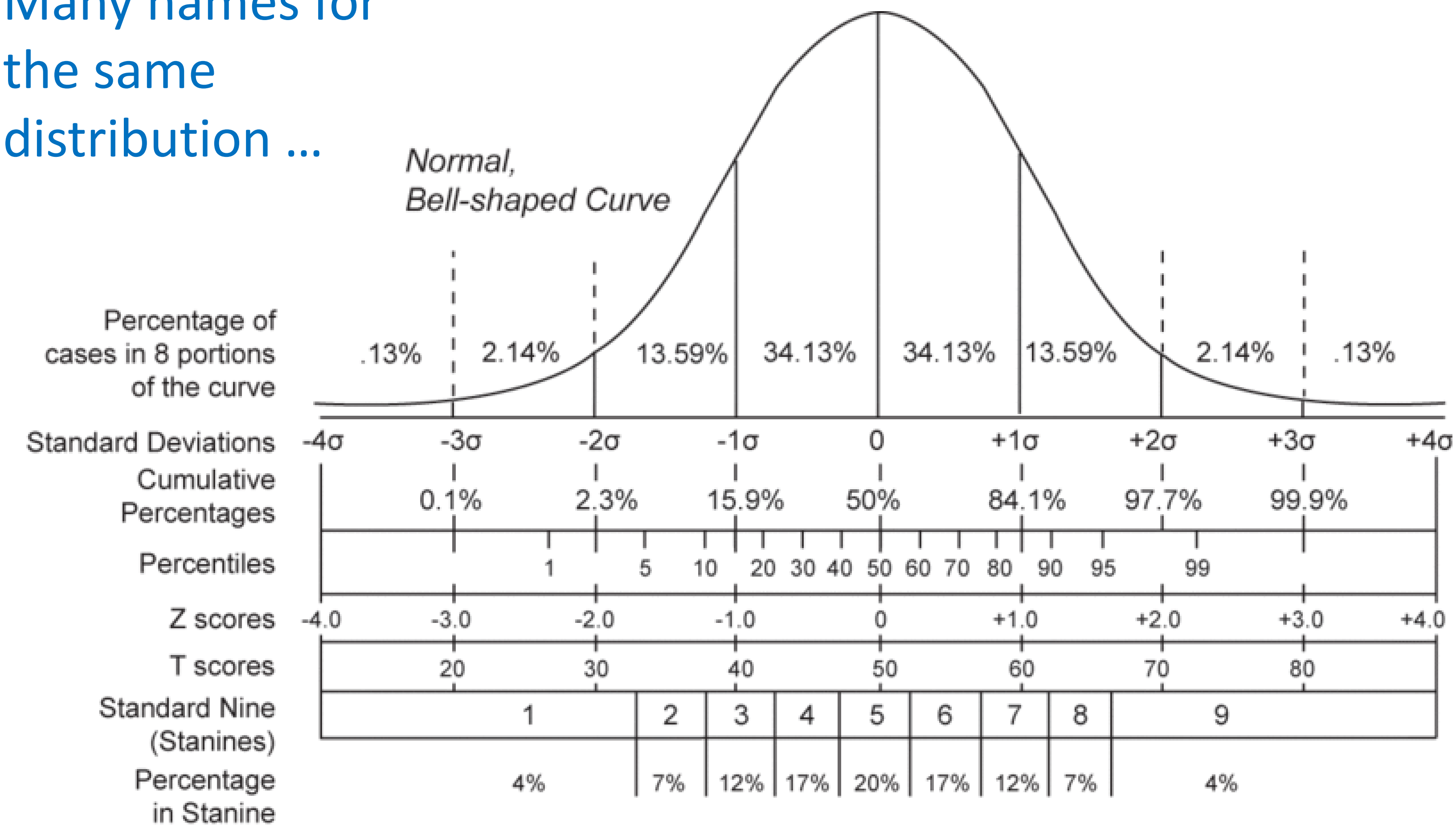
Special case: $\mu = 0,\ \sigma^2 = 1$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

If y is a Gaussian with μ, σ^2, then $x = \frac{y-\mu}{\sigma}$ follows $\varphi(x)$

The ”normal” distribution - Gaussian

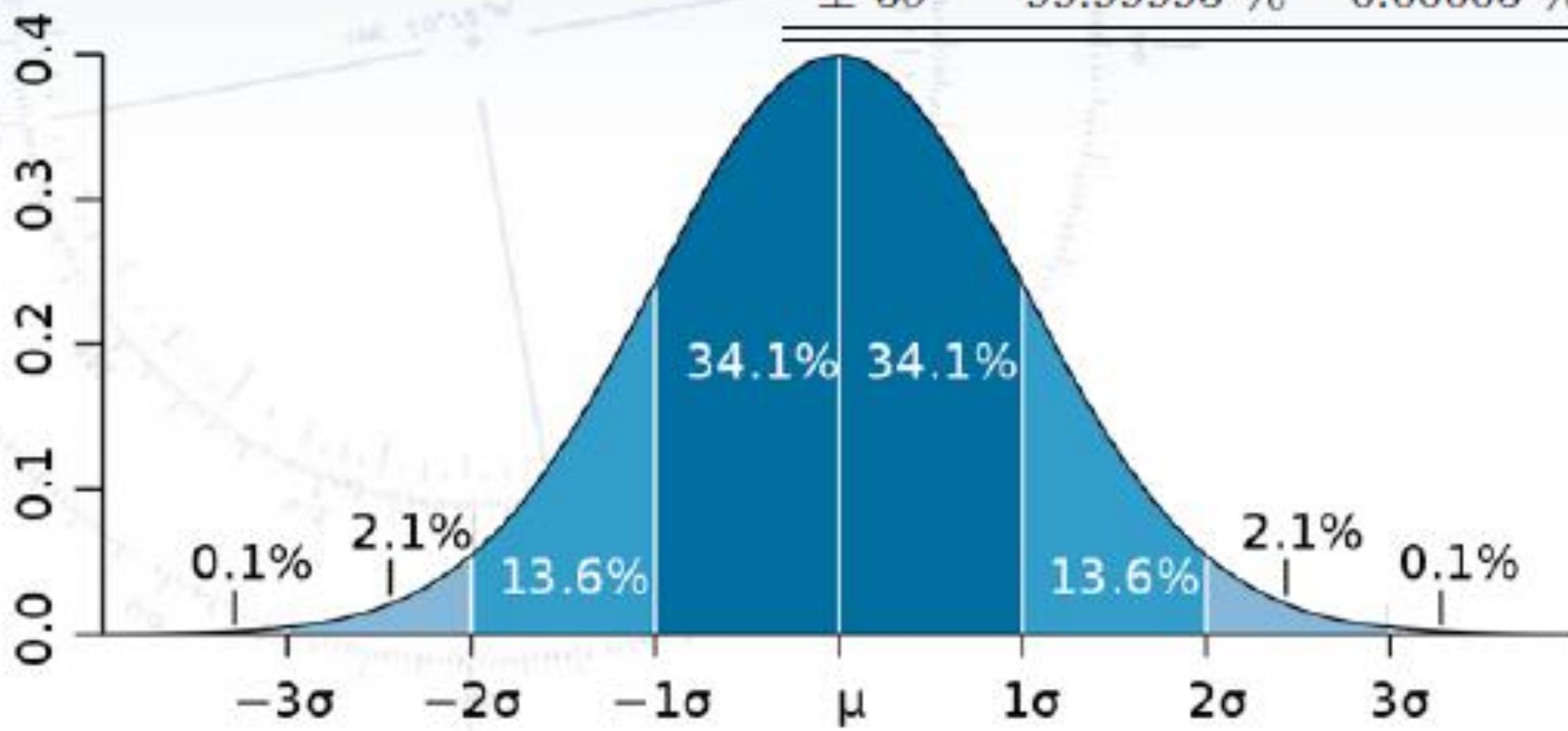
Many names for the same distribution …



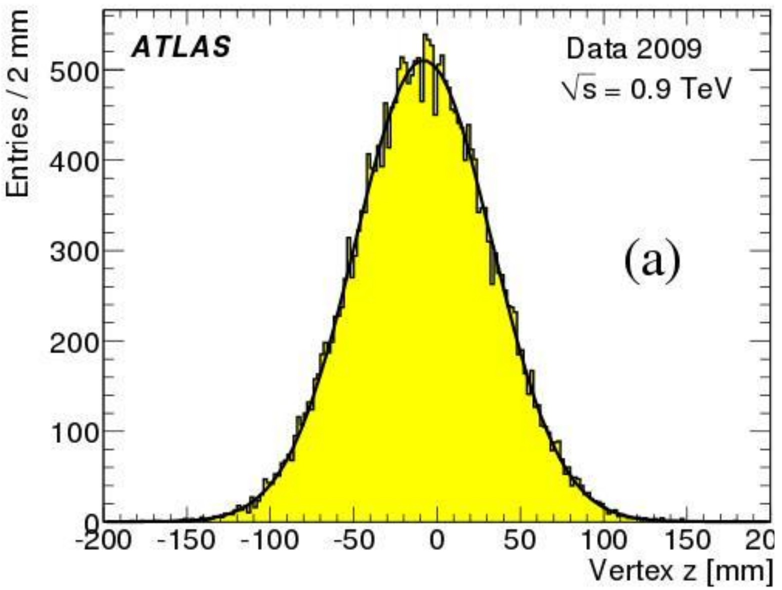
The Gaussian distribution

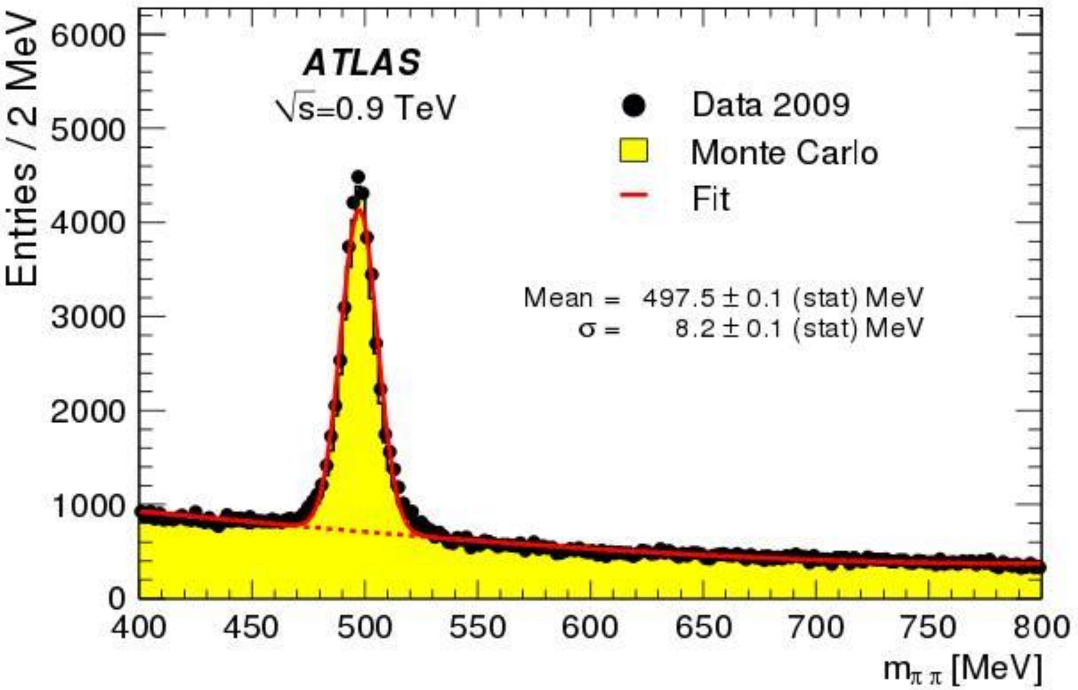
It is useful to know just a few of the most common Gaussian integrals:

Range	Inside	Outside
$\pm 1\sigma$	**68** %	32 %
$\pm 2\sigma$	**95** %	5 %
$\pm 3\sigma$	**99.7** %	0.3 %
$\pm 5\sigma$	99.99995 %	0.00005 %



ATLAS examples of Gaussian distributions





Distribution of vertex z coordinate for tracks

Invariant mass for $K^0{}_s$ peak fitted with a Gaussian (!)

Likelihood functions

Given a set of measurements x_i and parameter(s) θ, the likelihood function is defined as:

$$\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = \prod_i f(x_i, \theta)$$

The **principle of maximum likelihood** for parameter estimation consists of maximizing the likelihood of parameter(s) (here θ) given some data (here x)

The likelihood function plays a central role in statistics, as it can shown to be:

- ✓ Consistent (converges to the right value!)
- ✓ Asymptotically normal (converges with Gaussian errors)

Efficient and "optimal" if it can be applied in practice

Computational: often easier to minimize log likelihood:

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \theta}\right|_{\theta=\bar{\theta}} = 0$$

In problems with Gaussian errors boils down to a χ^2

Two versions, in practice:

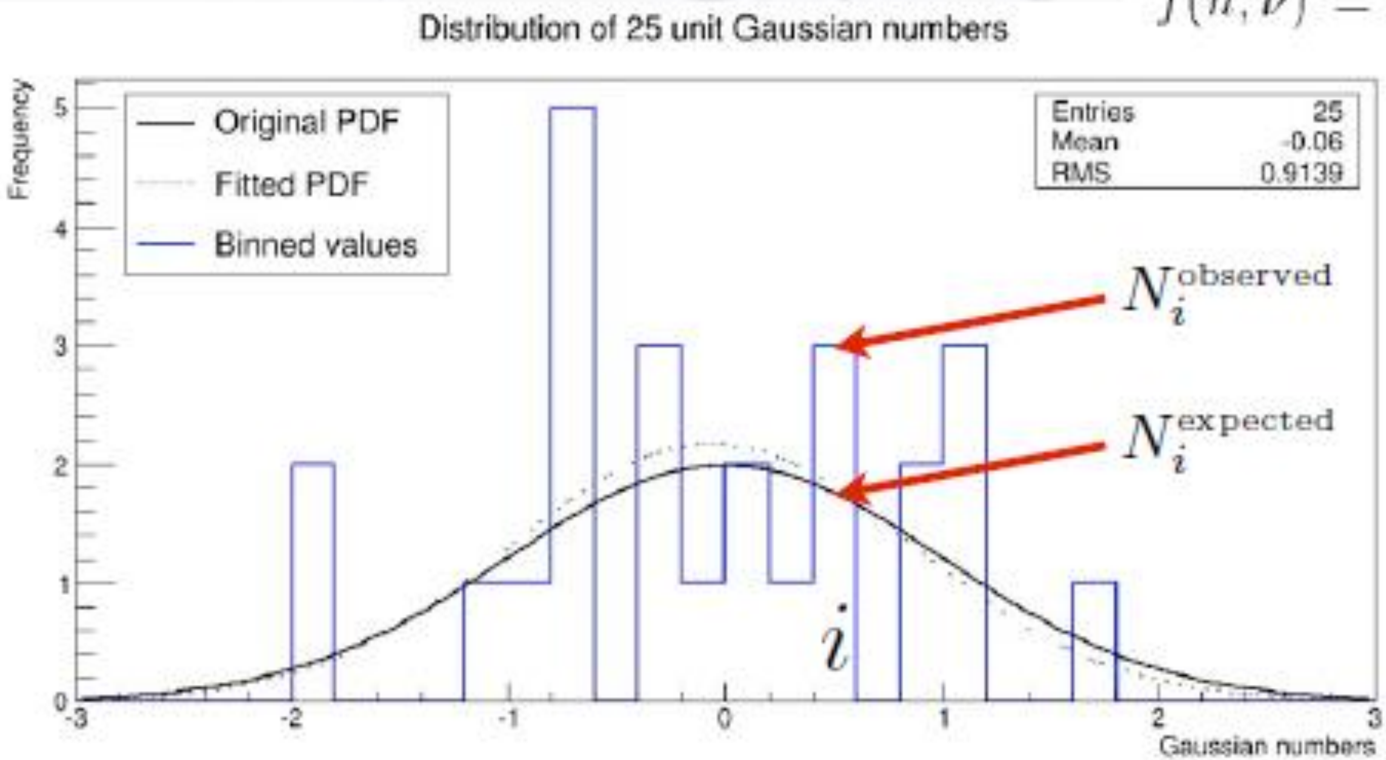
- Binned likelihood
- Unbinned likelihood

Binned likelihood

The binned likelihood is a sum over bins in a histogram:

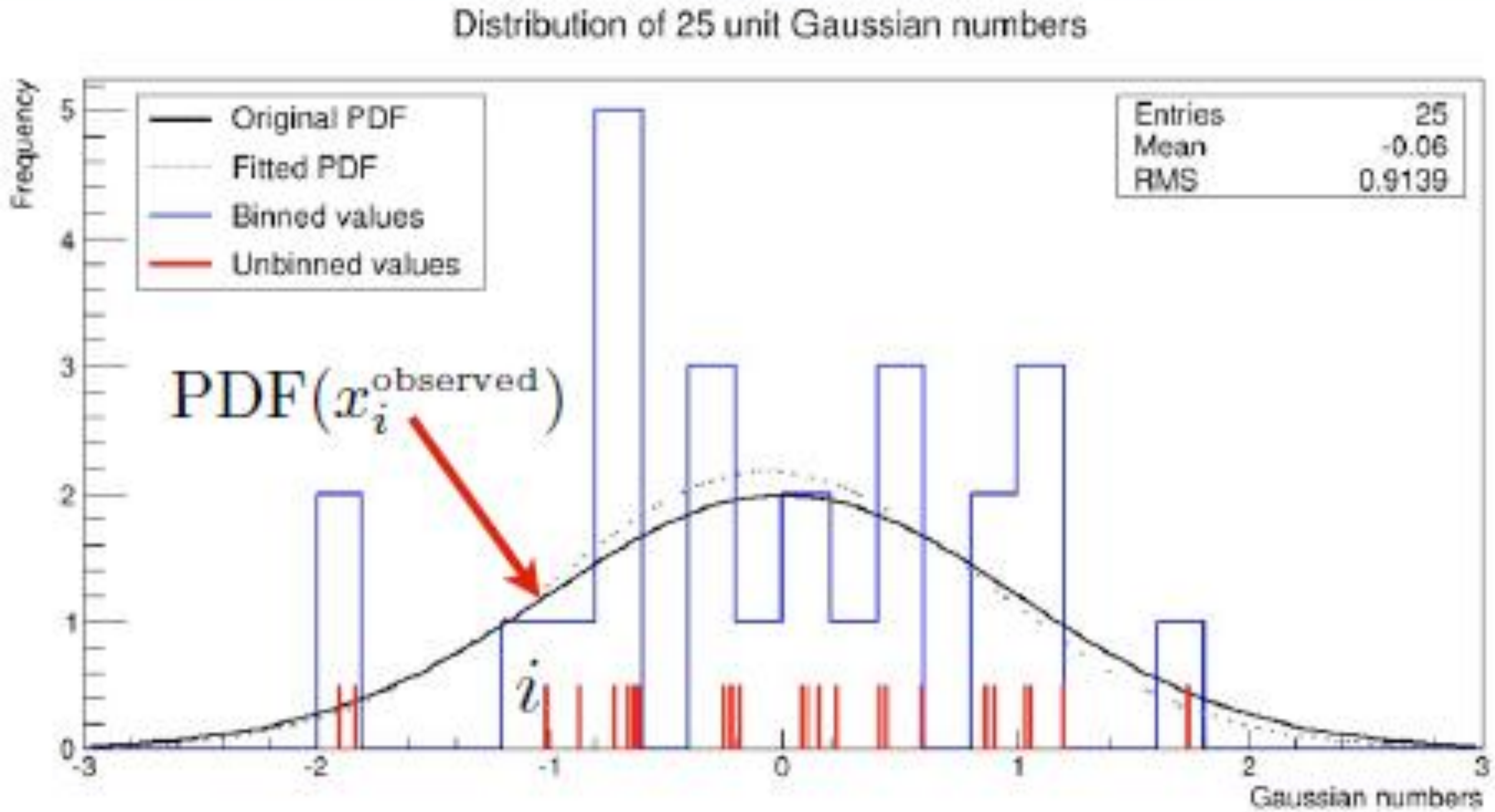
$$\mathcal{L}(\theta)_{\text{binned}} = \prod_{i}^{N_{\text{bins}}} \text{Poisson}(N_i^{\text{expected}}, N_i^{\text{observed}})$$

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



Unbinned likelihood

$$\mathcal{L}(\theta)_{\text{unbinned}} = \prod_{i}^{N_{\text{meas.}}} \text{PDF}(x_i^{\text{observed}})$$



Hypothesis testing

Hypotheses and acceptance/rejection regions

Goal is to make some statement based on the observed data x, as to the validity of the possible hypotheses.

A test of hypothesis H_0 is defined by specifying a **critical region** W (also called **rejection** region) of the data space S, such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there:

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

If x is observed there, reject H_0.

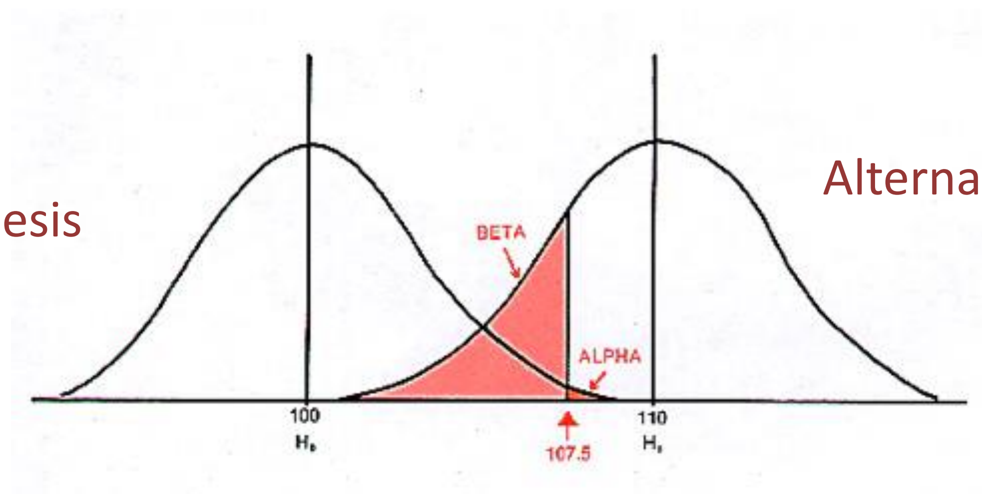
α is called the **size or significance level** of the test.

The complementary region is called **acceptance region**.

Test statistics

1. State hypothesis (null and alternative)
2. Set criteria for decision, select test statistics, select a significance level
3. Compute the value of the test statistics and from that the probability of observation under null-hypothesis (p-value)
4. Make the decision! Reject null hypothesis if p-value is below significance level

Null hypothesis



Alternative hypothesis

Test statistics

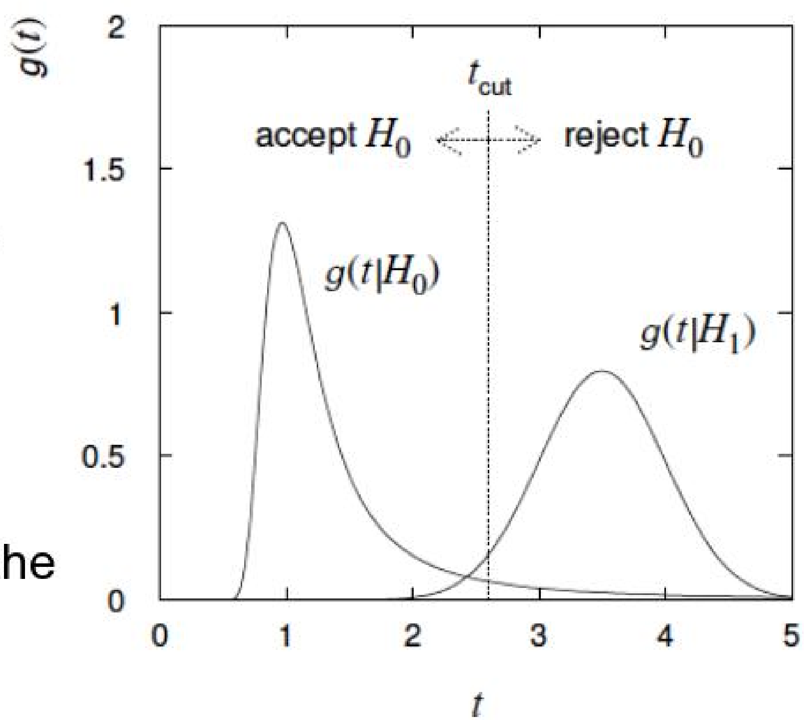
The decision boundary can be defined by an equation of the form:

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

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

We can work out the pdf's:

$$g(t|H_0), \;\; g(t|H_1)$$

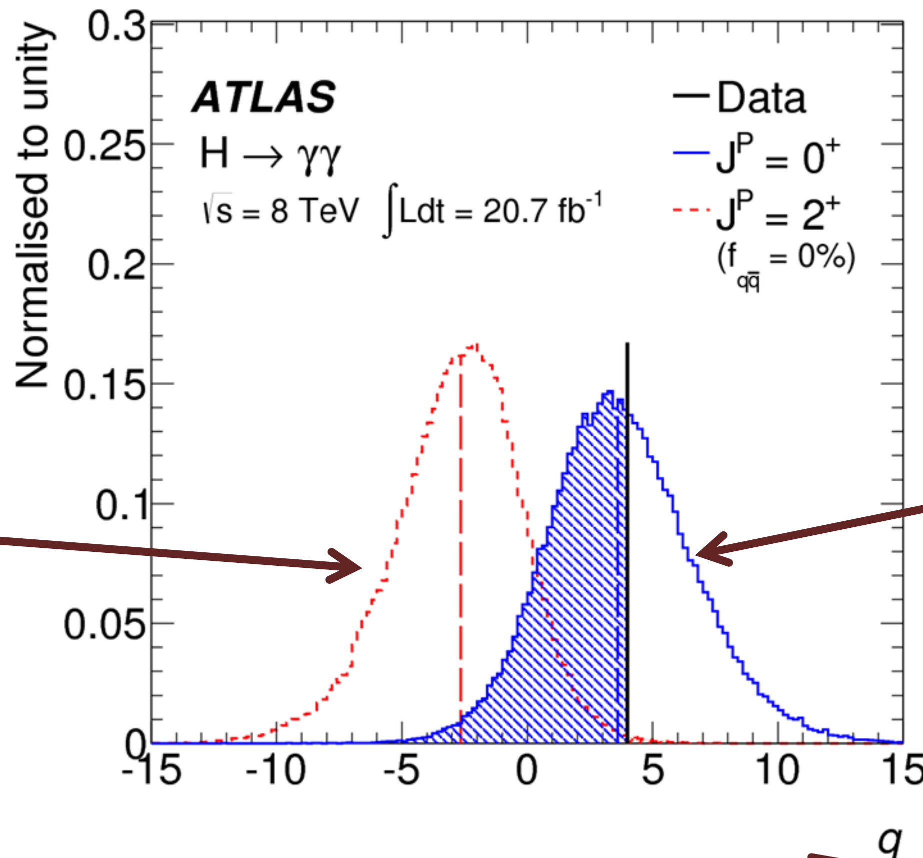


Decision boundary is now a single 'cut' on t, which divides the space into the critical (rejection region) and the acceptance region.

This defines a TEST: if the data fall in the critical region, we reject H_0

Example of hypothesis test

The spin of the newly discovered Higgs-like particle (spin 0 or 2?)



PDF of spin 2 hypothesis

PDF of spin 0 hypothesis

Test statistics (Likelihood ratio[decay angles])

Selection

We have a data sample with two kinds of events, corresponding to hypotheses H_0 (background) and H_1 (signal).

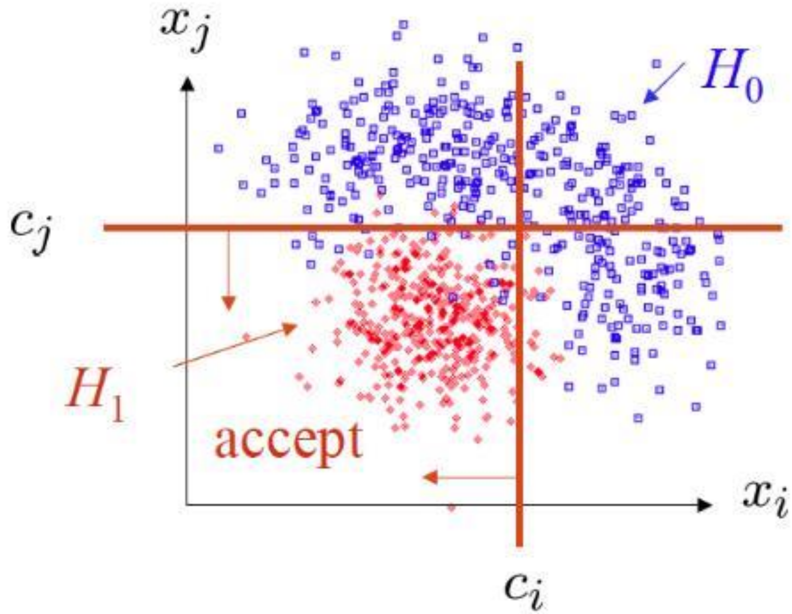
We want to select those of type H_1.

Each event is a point in $\vec{x}$ space (n dimensions).

What 'decision boundary' should we use to accept/reject events as belonging to event types H_0 or H_1?

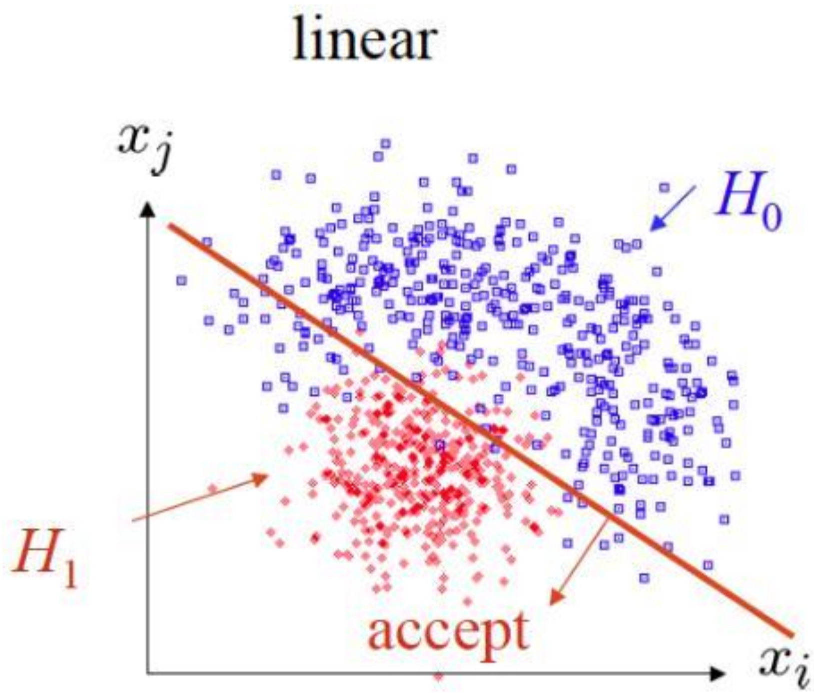
One possibility is to select events with several 'cuts': e.g.

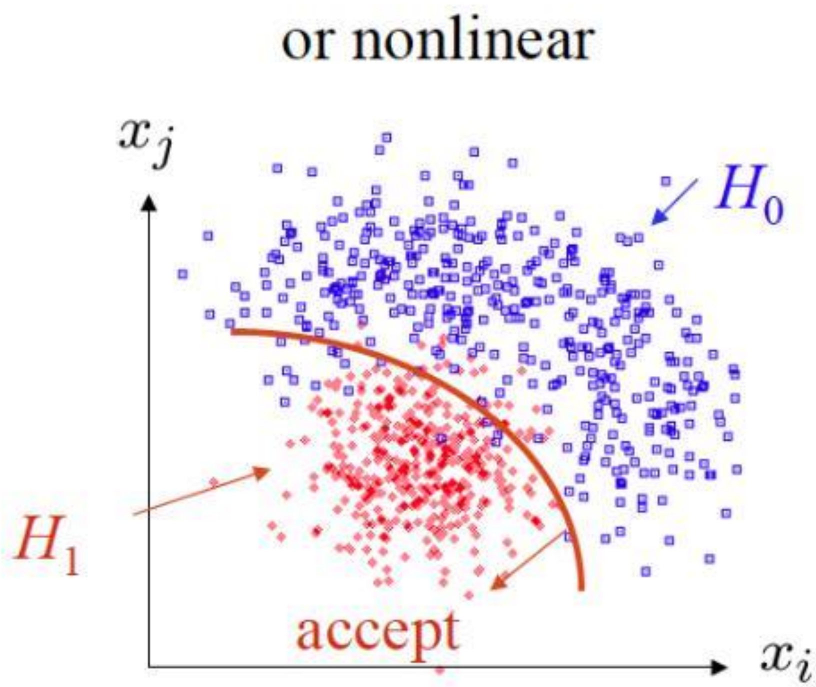
$$x_i < c_i$$
$$x_j < c_j$$



Other selection options

But we can also use some other sort of decision boundary !!



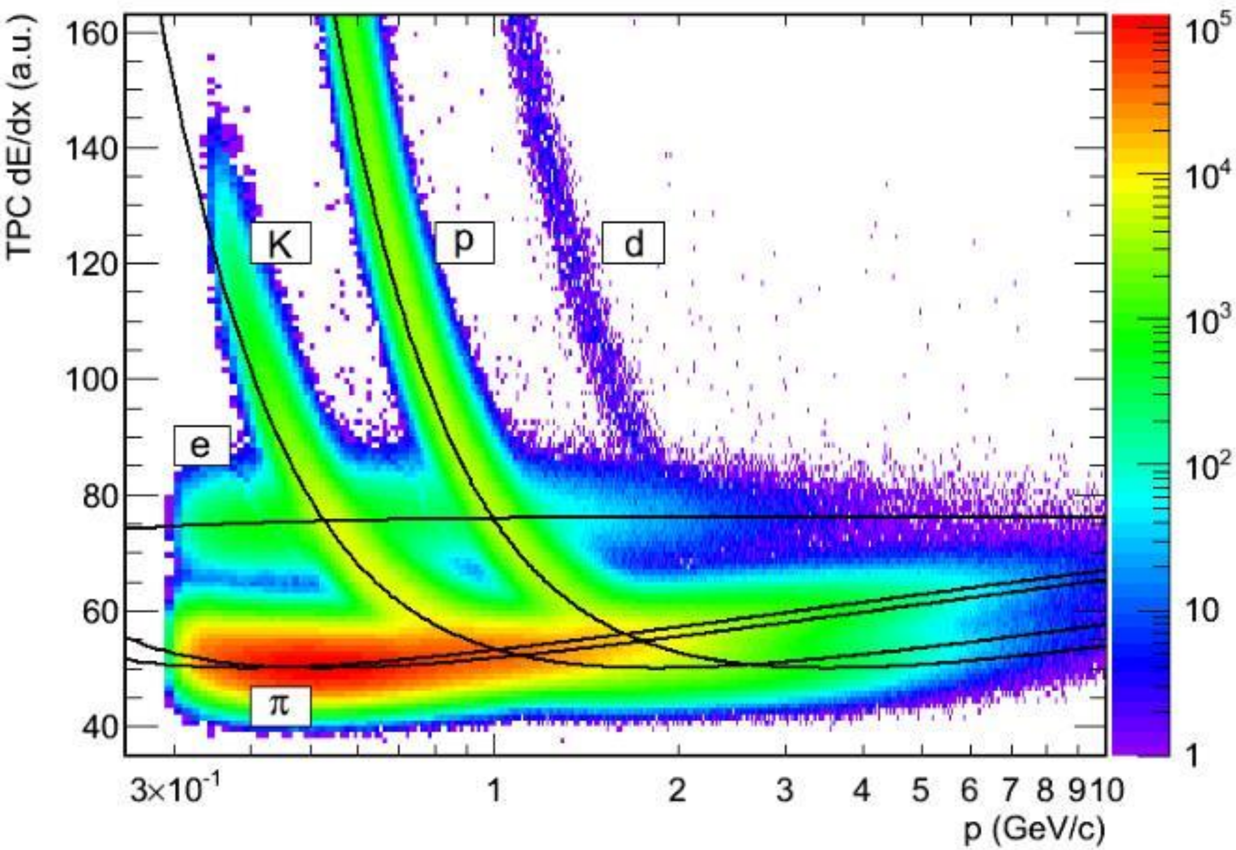


How can we formalize this to choose the boundary in an 'optimal' way?

ALICE example

Use the ALICE Time Projection Chamber to identify the particle species: electron, muon, pion, kaon, proton, deuteron

"x" = particle momentum (p), specific energy loss in TPC (dE/dx) (and more)



Example:
I want to select electrons (hypothesis H_1) from all other particles (hypothesis H_0)

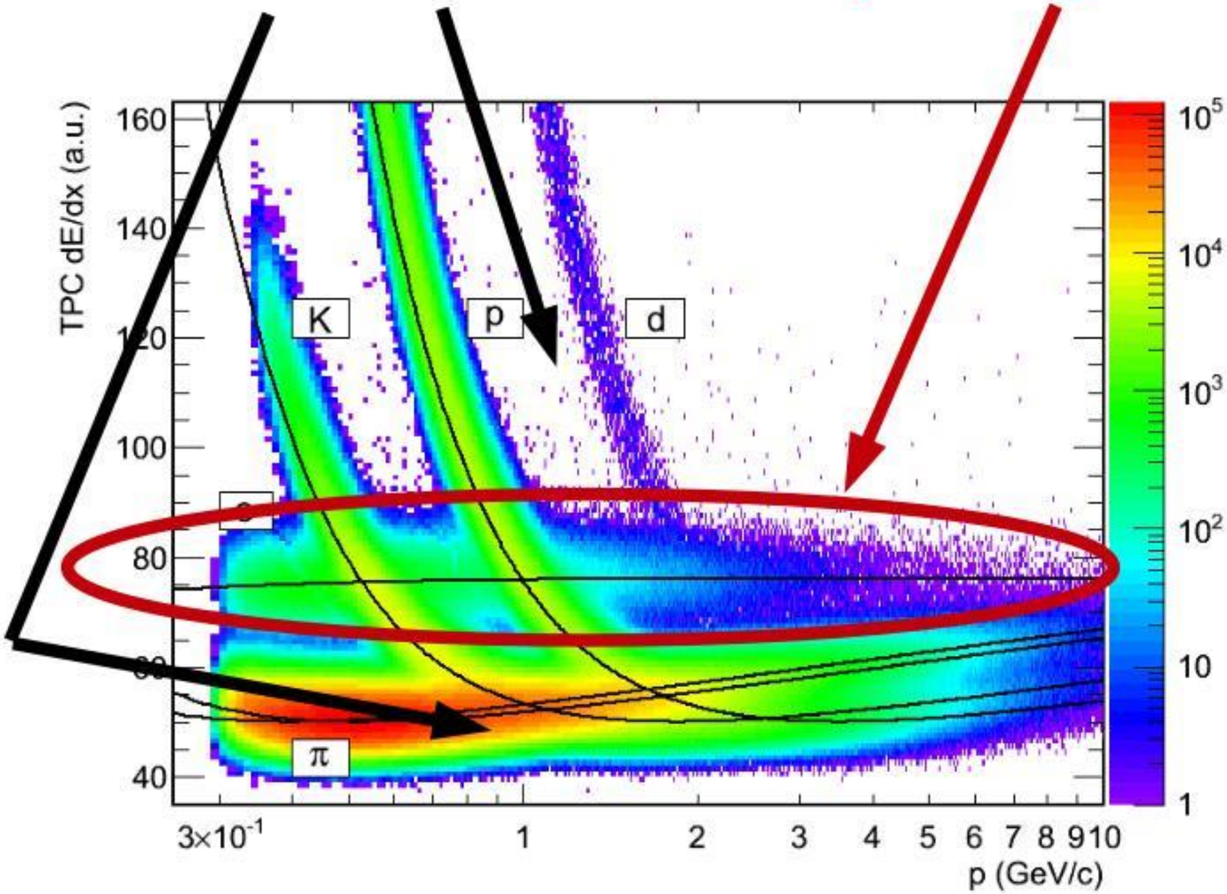
In Bayesian approach:
Can add prior hypotheses on the relative particle abundances (e.g. you see that pions are many more!)

ALICE example

Looking for electrons: null hypothesis H_0 is to be a hadron

Acceptance region here:

Rejection region here (for bkg = hadrons)



Type I / Type II errors:

Trying to select signal events:
(i.e. try to disprove the null-hypothesis stating it were "only" a background event)

accept as: / truly is:	signal	back-ground
signal	☺	Type II error
back-ground	Type I error	☺

Rejecting the hypothesis H_0 when it is true is a Type-I error.
The maximum probability for this is the **size of the test**:

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

But we might also accept H_0 when it is false and an alternative H_1 is true.
This is called Type-II error, and occurs with probability:

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

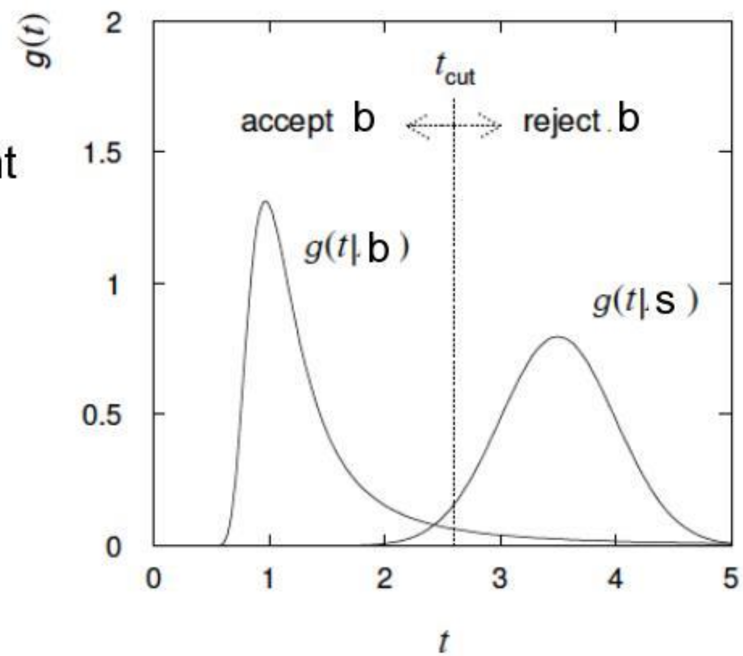
One minus this is called the power of the test with respect to the alternative hypothesis H_1:

Power = $1 - \beta$

Signal/background efficiency

The probability to reject background hypothesis for a background event (background efficiency) is:

$$\epsilon_b = \int_{t_{cut}}^{\infty} g(t|b)\, dt = \alpha$$



The probability to accept a signal event as signal (signal efficiency) is:

$$\epsilon_s = \int_{t_{cut}}^{\infty} g(t|s)\, dt = 1-\beta$$

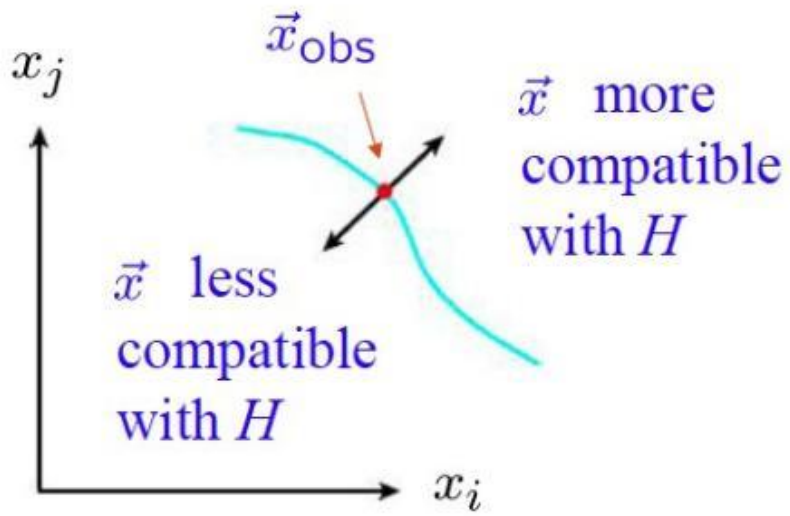
Significance tests/goodness of fit

Suppose hypothesis *H* predicts pdf $f(\vec{x}|H)$ for a set of observations $\vec{x} = (x_{1,} \ldots, x_n)$

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

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

Decide what part of the data space represents less compatibility with *H* than does the point $\vec{x}_{obs}$
(Not unique!)



p-values

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

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

NOTE! This is NOT the probability that H is true!

In frequentist statistics we don't talk about P(H) (unless H represents a repeatable observation).

In Bayesian statistics we do. Use Bayes' theorem to obtain

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

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

For now stick with the frequentist approach.

Neyman-Pearson's lemma

The Neyman–Pearson lemma states: to get the highest purity for a given efficiency, (i.e. highest power for a given significance level), choose the acceptance region such that

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c,$$

where $c =$ constant which determines the efficiency.

This even gives that the likelihood ratio, $-2\ln\frac{\mathcal{L}_0}{\mathcal{L}_1}$, is the most powerful test

Significance of an observed signal

Suppose we observe n events. These can consist of:

n_b events from known processes (background)

n_s events from a new process (signal)

If n_s, n_b are Poisson random variables with means s, b, then $n=n_s+n_b$ is also Poisson, with mean s+b

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

Suppose b=0.5, and we observe nobs=5. Should we claim evidence for a new discovery?

Significance of an observed signal

Suppose we observe n events. These can consist of:

n_b events from known processes (background)

n_s events from a new process (signal)

If n_s, n_b are Poisson random variables with means s, b, then $n=n_s+n_b$ is also Poisson, with mean s+b

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

Suppose b=0.5, and we observe nobs=5. Should we claim evidence for a new discovery?

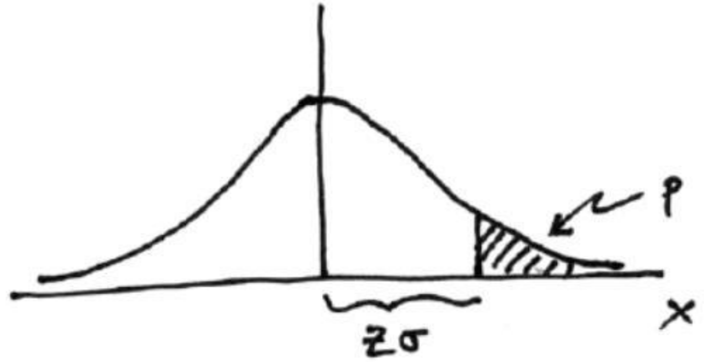
Give p-value for hypothesis s=0:

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

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

Significance vs 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



Small p = unexpected

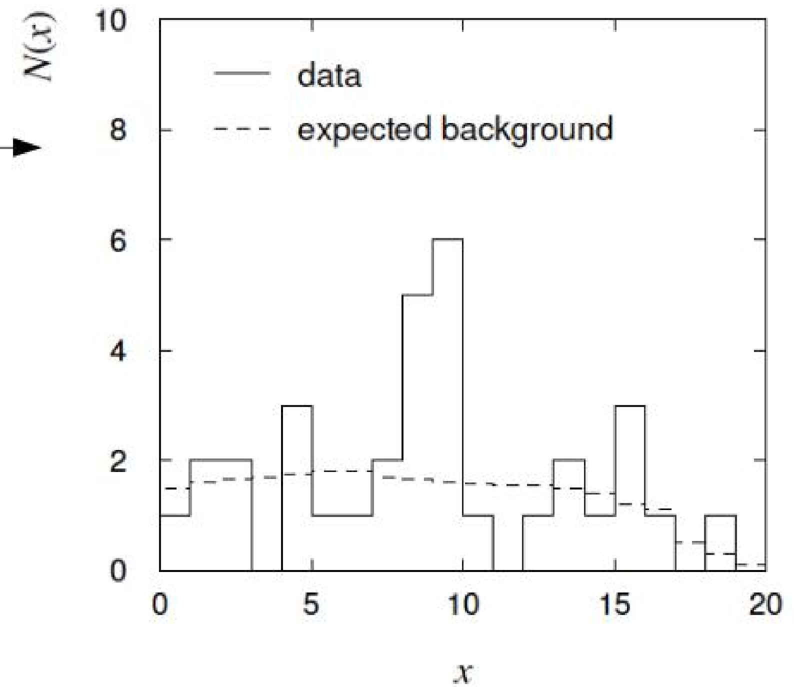
$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

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

$(x_0-\mu)/\sigma$	1	2	3	4	5
p	16%	2.3%	0.13%	0. 003%	$0.3*10^{-6}$

Significance of a peak

Suppose we measure a value x for each event and find:



Each bin (observed) is a Poisson r.v., means are given by the dashed line

In the two bins with the peak, 11 entries found with b = 3.2
The p-value for the s=0 hypothesis is:

$$P(\, n \geq 11;\, b=3.2,\, s=0) = 5.0 \times 10^{-4}$$

Significance of a peak

But … did we know where to look for the peak?

→ give $P(n \geq 11)$ in any 2 adjacent bins

Is the observed width consistent with the expected x resolution?

→ take x window several times the expected resolution

How many bins x distributions have we looked at?

→ look at a thousand of them, you'll find a 10^{-3} effect

Did we adjust the cuts to "enhance" the peak?

→ freeze cuts, repeat analysis with new data

How about the bins to the sides of the peak … (too low!)

Should we publish ??

How many σ's?

HEP folklore is to claim discovery when $p = 2.9 \times 10^{-7}$, corresponding to a significance $Z=5$.

This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.

Phenomenon	Reasonable p-value for discovery
D^0D^0 mixing	~0.05
Higgs	$\sim 10^{-7}$
Life on Mars	$\sim 10^{-10}$
Astrology	$\sim 10^{-20}$

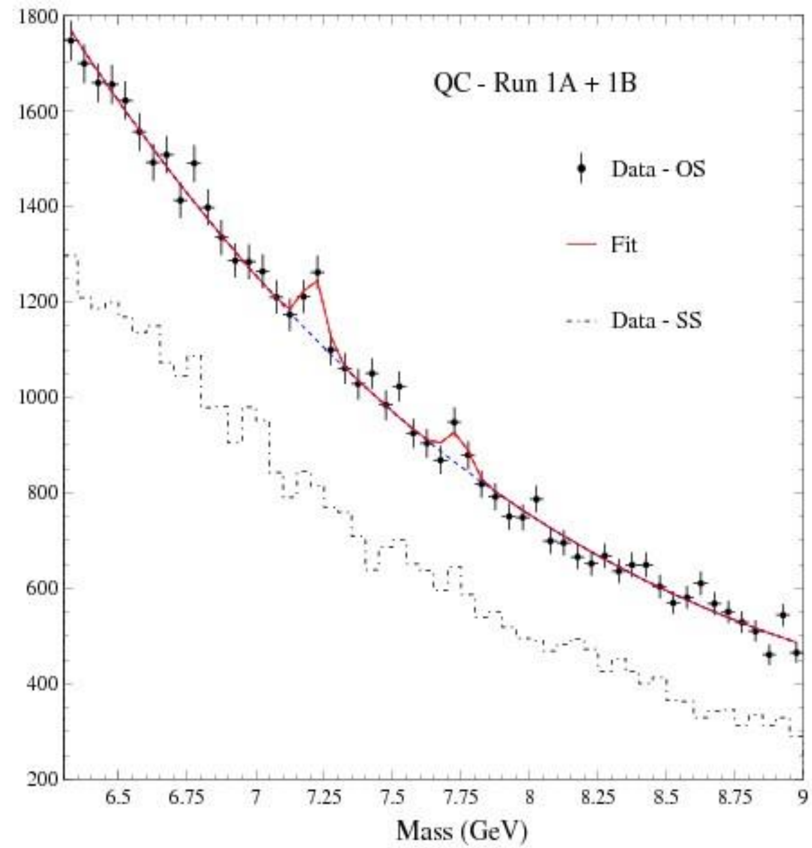
One should also consider the degree to which the data are compatible with the new phenomenon, not only the level of disagreement with the null-hypothesis: p-value is only the first step !!!

Look-elsewhere-effect (LLE)

Example from CDF: **Is there a bump at 7.2 GeV ?** *(and even 7.75 GeV?!)*

Excess has significance but when we take into account that the bump(s) could have been anywhere in the spectrum (the look-elsewhere-effect) significance is reduced:

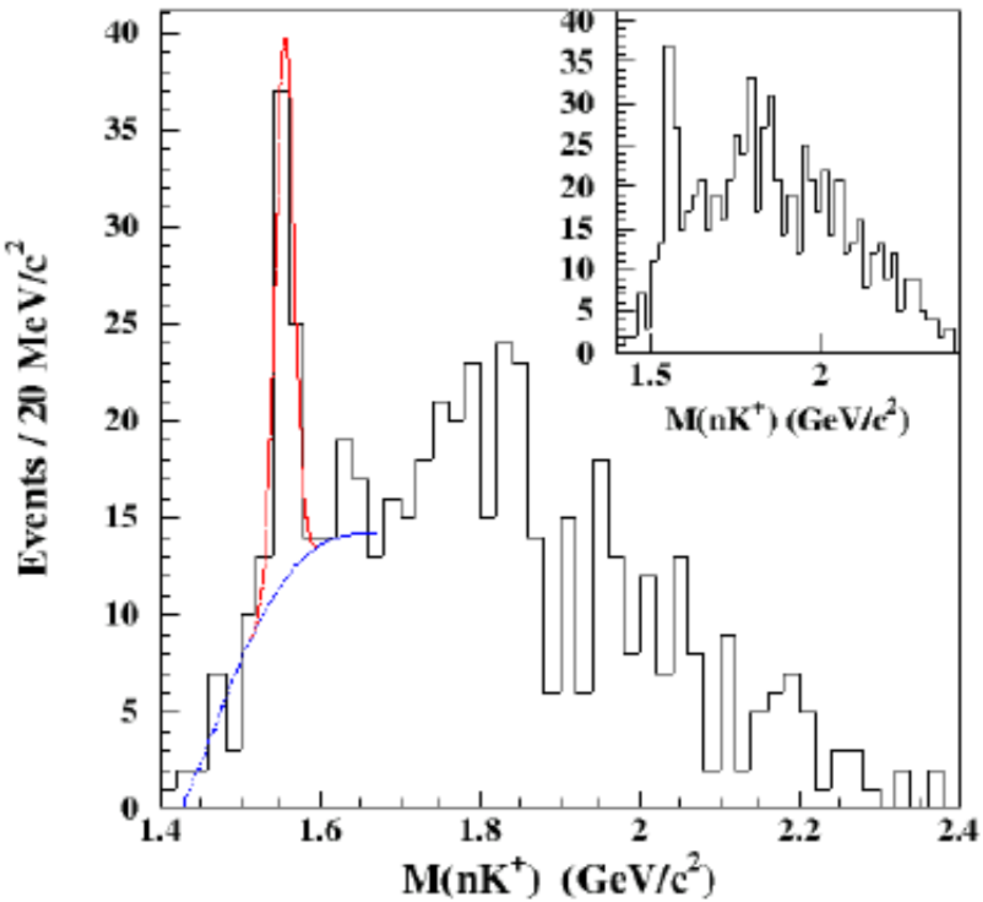
p-value(corr) = p-value $\times$ (number of places it might have been spotted in spectrum)



In this case ~ mass interval / width of bump

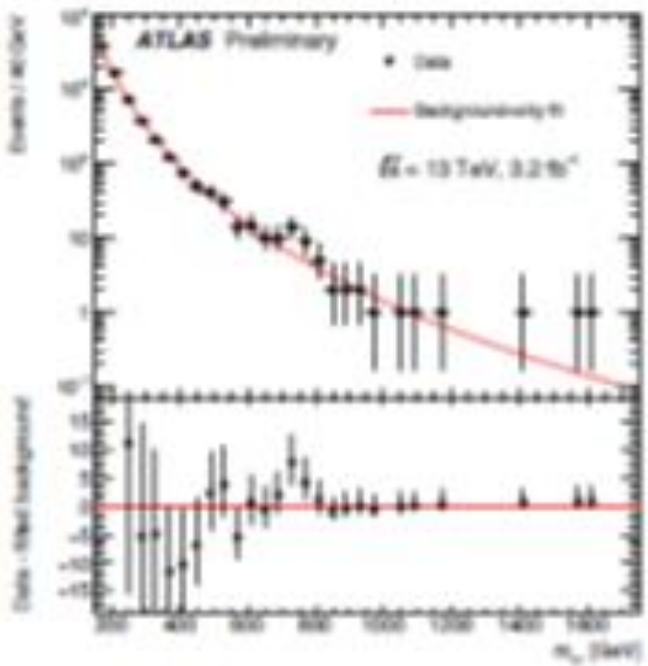
Results in low significance
Never saw these again

Remember the penta-quark …

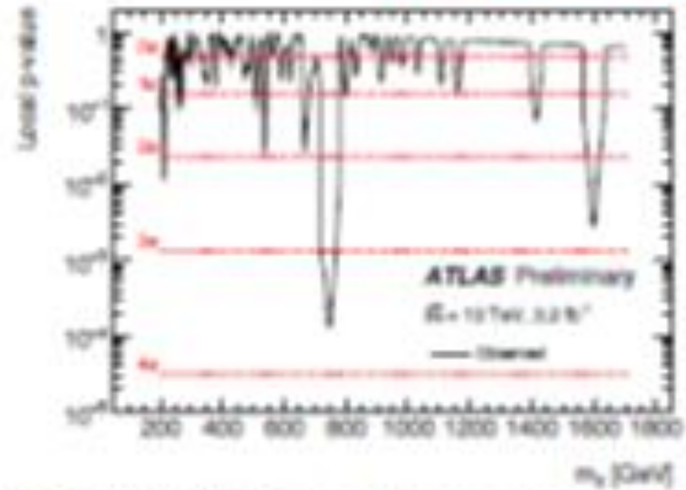


Search for a Two Photons Resonance (II) **ATLAS results**

Results: Events with mass in excess of 200 GeV are included in **unbinned fit**



- **In the NWA search, an excess of 3.6σ (local) is observed at a mass hypothesis of minimal p_0 of 750 GeV**
- **Taking a LEE in a mass range (fixed before unblinding) of 200 GeV to 2.0 TeV the global significance of the excess is 2.0σ**



In the **NWA** fit the resolution uncertainty is profiled in the **NWA** fit and is pulled by 1.5σ

The data was then fit under a **LW hypothesis** yielding a width of approximately 45 GeV (Approx. 6% of the best fit mass of approximately 750 GeV)

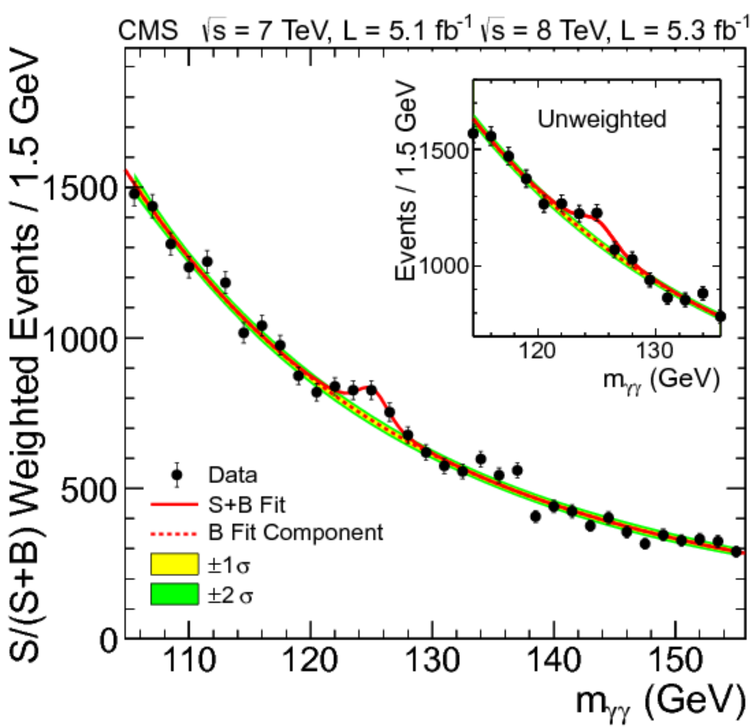
- As expected the local significance increases to 3.9σ
- Taking into account a LEE in mass and width of up to 10% of the mass hypothesis of 2.3σ (Note: upper range in resolution fixed after unblinding)

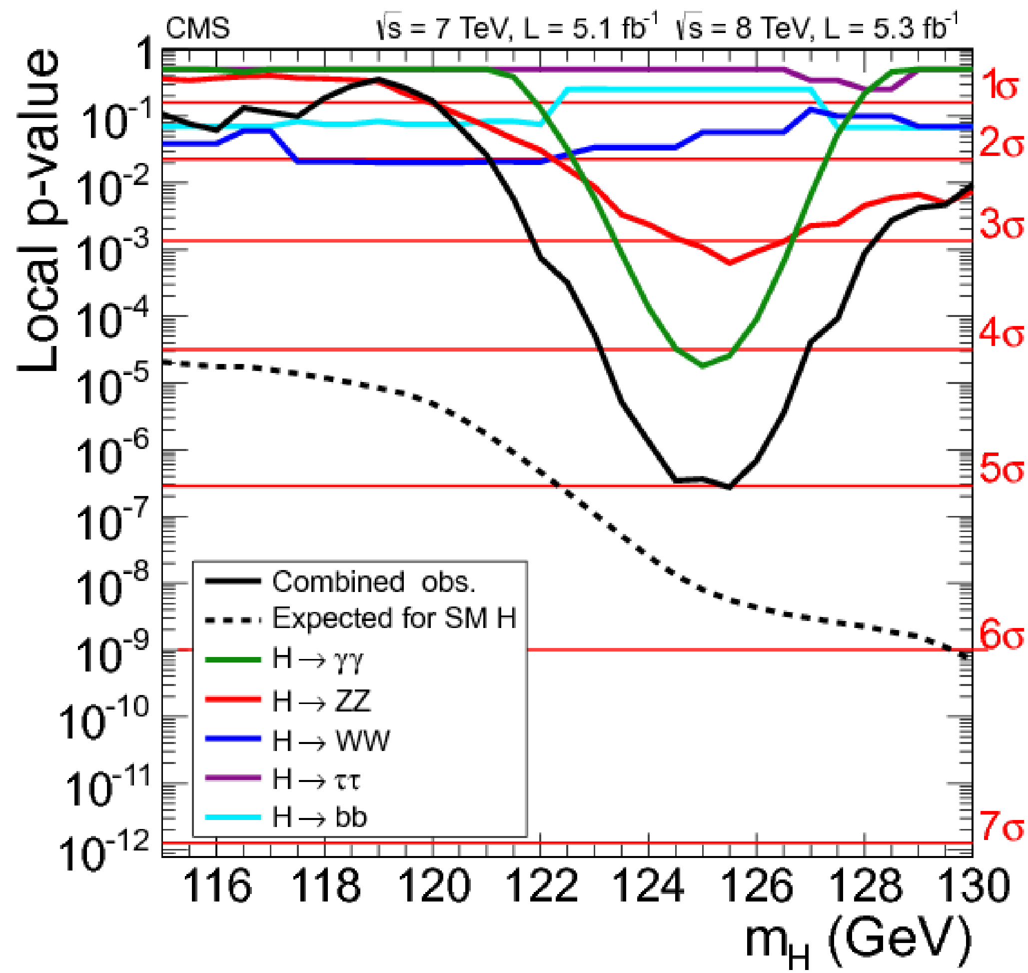
31

(example 3.2)

Limit setting

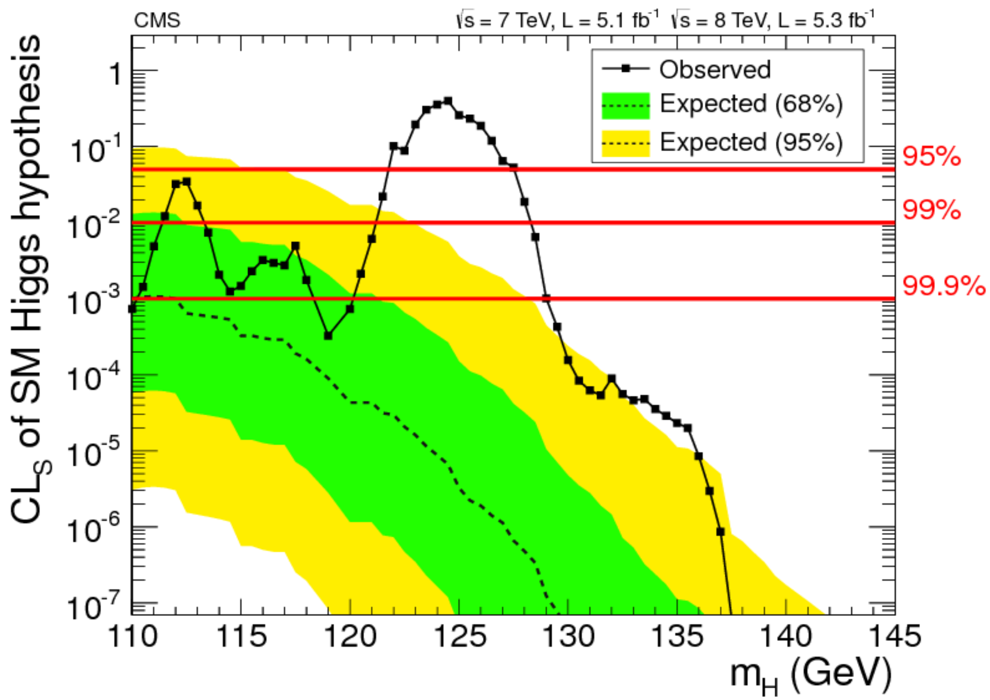
CMS 2012 Higgs result



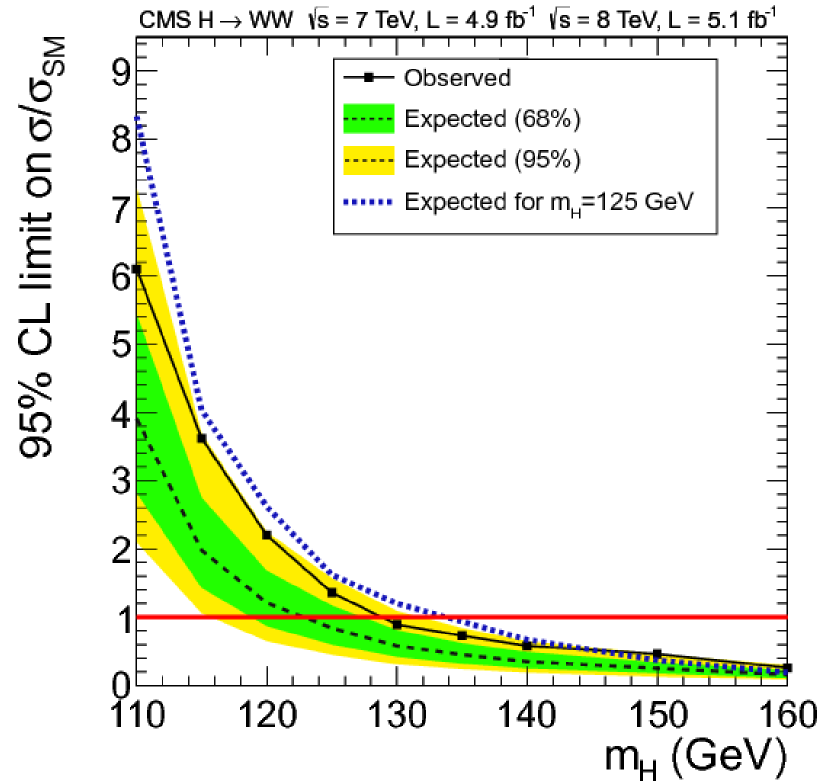


(example 3.2)

Confidence intervals



Upper limt on signal cross section/ SM Higgs cross section in $H \rightarrow WW$ channel



Summary/ outlook

- Gaussian distribution very useful
  - Errors tend to be gaussian
- To check a New Physics hypothesis against the Standard Model
  - Define test statistics
  - Define level of significance
  - Remember the look elsewhere effect
- P-values gives P(data|null hypothesis)
  - It does not say whether the hypothesis is true!