Statistical data analysis

Ultra-brief introduction

Igor Boyko

Why probability?

- Physics is believed to be a precise science. Why at all do we need the concept of chance and probability?
- It is fundamental property of the nature (quantum mechanics!)
- Certain measurements are fundamentally statistical. For example, suppose you want to count the number of molecules in $1cm^3$ of air
- Last but not least: imperfections of apparatus, simplifications in the experimental methods, our own ignorance

Random variable

- Random variable x is an experimentally measured number whose value can not be predicted before doing the measurement
- Random variable may be discrete or continuous
- Distribution of the Probability Density Function (PDF):

$$p(x < X < x + dx) = f(x)dx$$

$$\int f(x)dx = 1$$

$$\sum_i P_i = 1$$

Average and variance

$$E(g(x)) = \overline{g}_f = \int g(x)\ f(x)dx$$

$$D(g(x)) = E[g(x) - E(g(x))]^2$$

- In particular, average and variance of the random variable itself:

$$\mu = E(x) = \overline{x} = \int x\ f(x)dx$$

$$\sigma^2 = D(x) = \overline{(x-\mu)^2} = \int (x-\mu)^2\ f(x)dx = \overline{x^2} - (\overline{x})^2$$

The two most important distributions

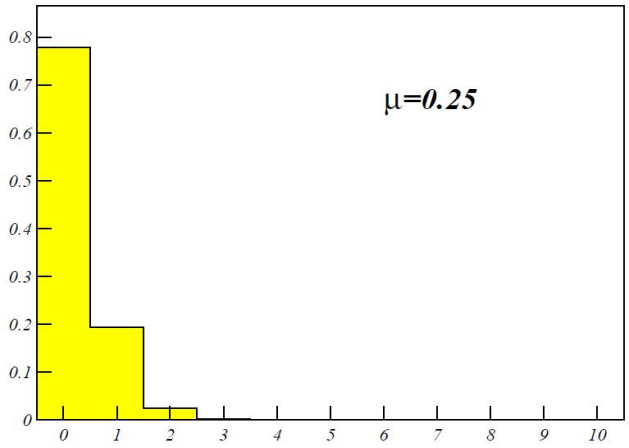
- Poisson distribution
  - Number of ***independent*** events that happen over a fixed period of time
- Normal (Gaussian) distribution
  - Result of a measurement usually has the normal distribution around the true value

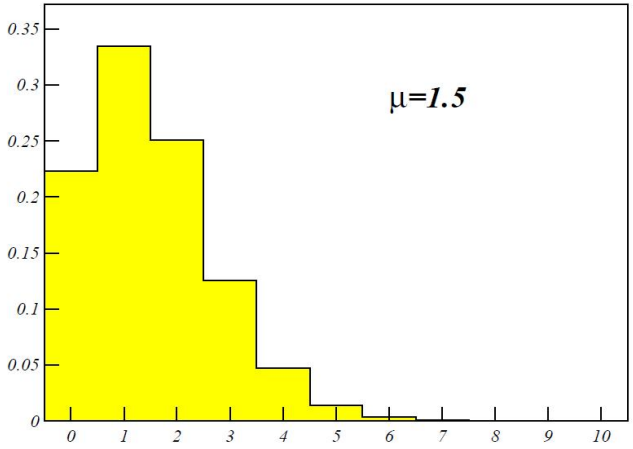
Poisson distribution

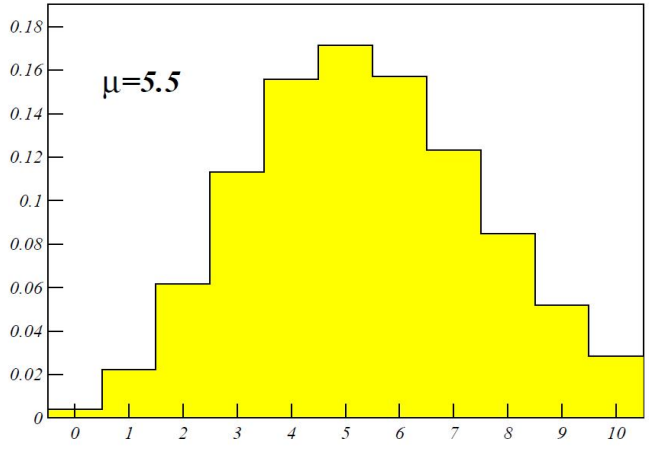
$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

$$E(n) = \mu \qquad D(n) = \mu$$

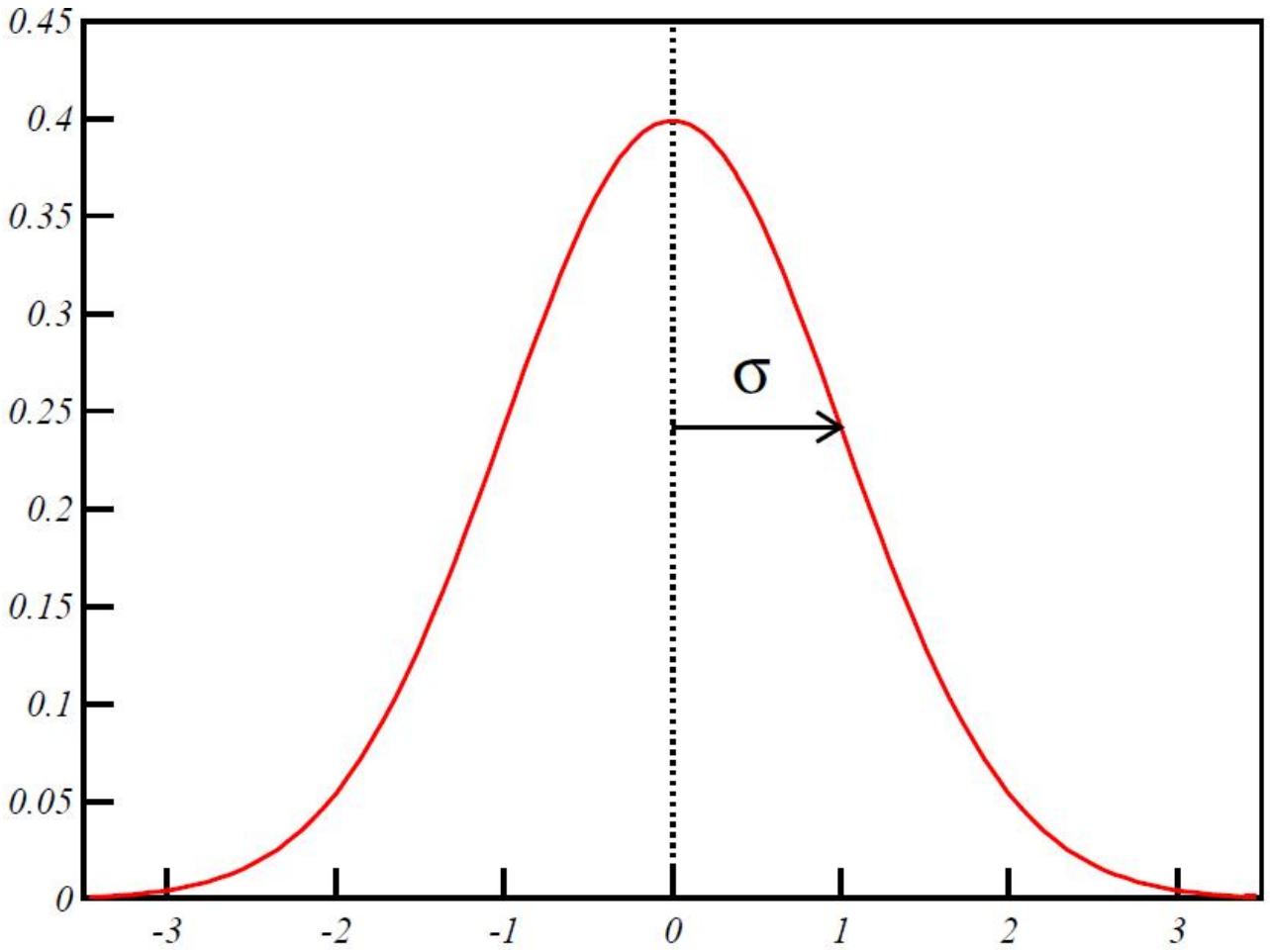
$$\hat{\mu} = n \pm \sqrt{n}$$







Normal distribution



Normal distribution

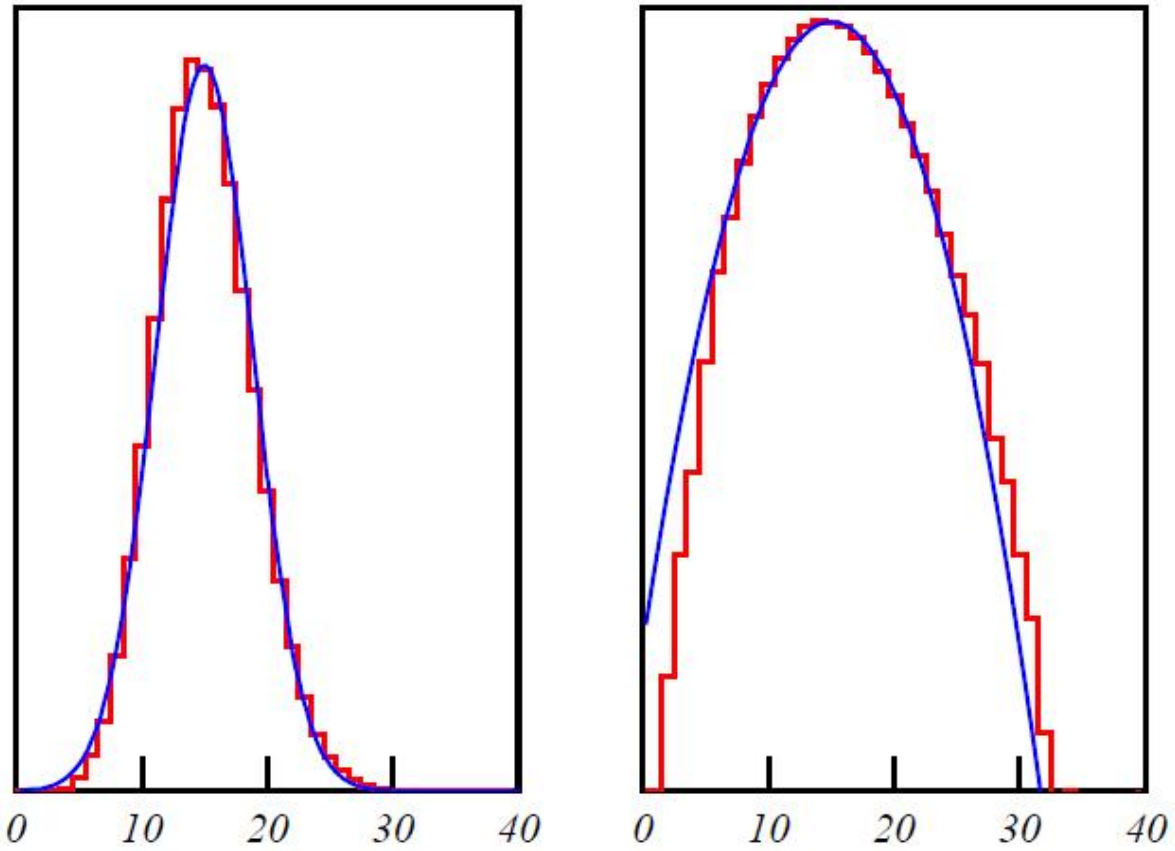
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = N(\mu, \sigma^2)$$

$$E(x) = \mu \qquad\qquad D(x) = \sigma^2$$

- Standard normal distribution: $\mu = 0$, $\sigma = 1$

$$N(0,1) = g(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}$$

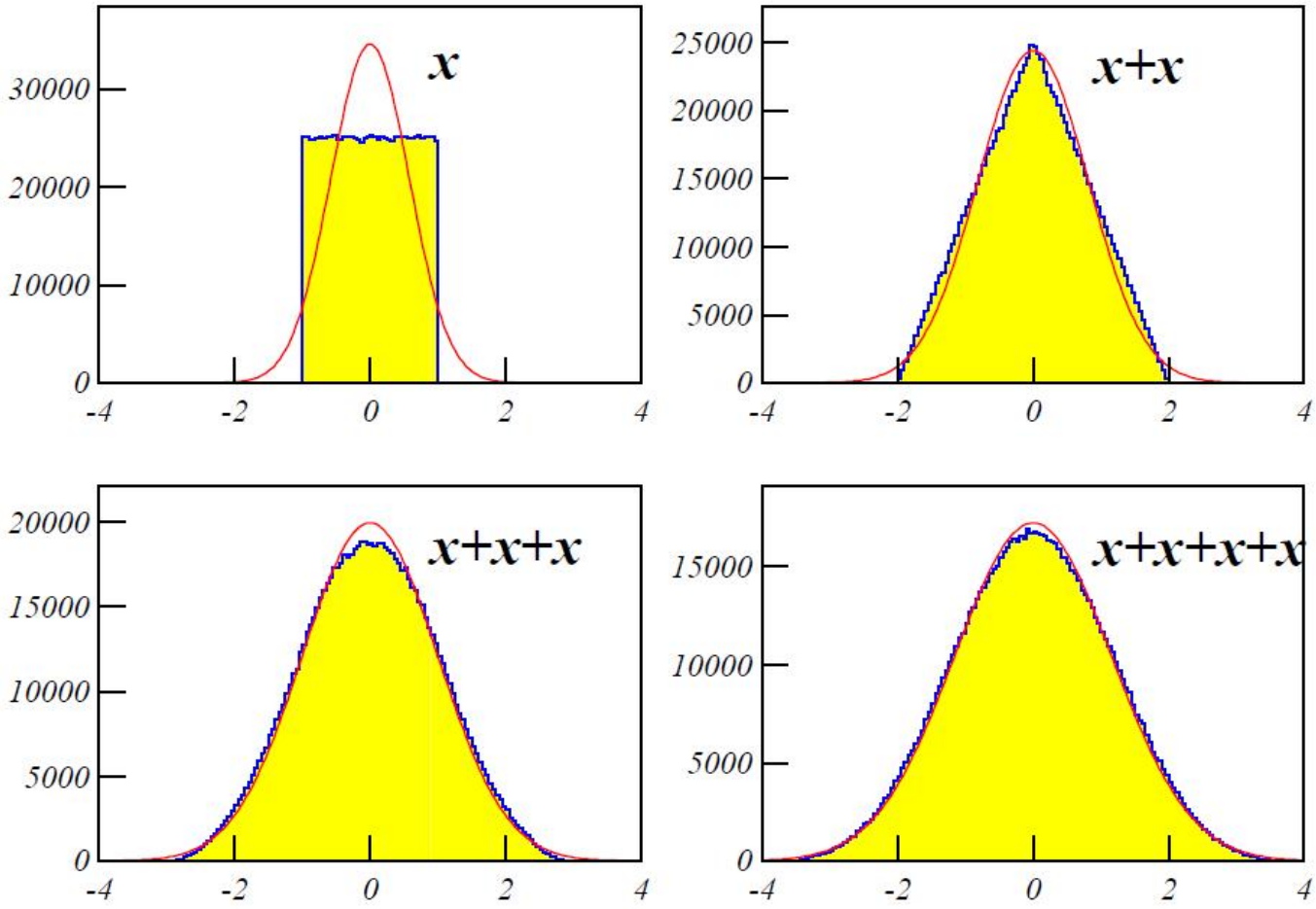
Comparison of Poisson (μ=15) and normal distribution



Central limits theorem

- Distribution of the sum of many random variables is normal in the limit of infinite summation.
  - Attention – this formulation is simplified and approximate!
- Do not confuse sum of distributions $f(x)+g(x)+\ldots$ and distribution of the sum $f(x+y+\ldots)$!!!
- Typically, the experimental error is a sum of many random contributions. That's why the experimental uncertainty is usually described by the Gaussian distribution

Illustration of Central Limits



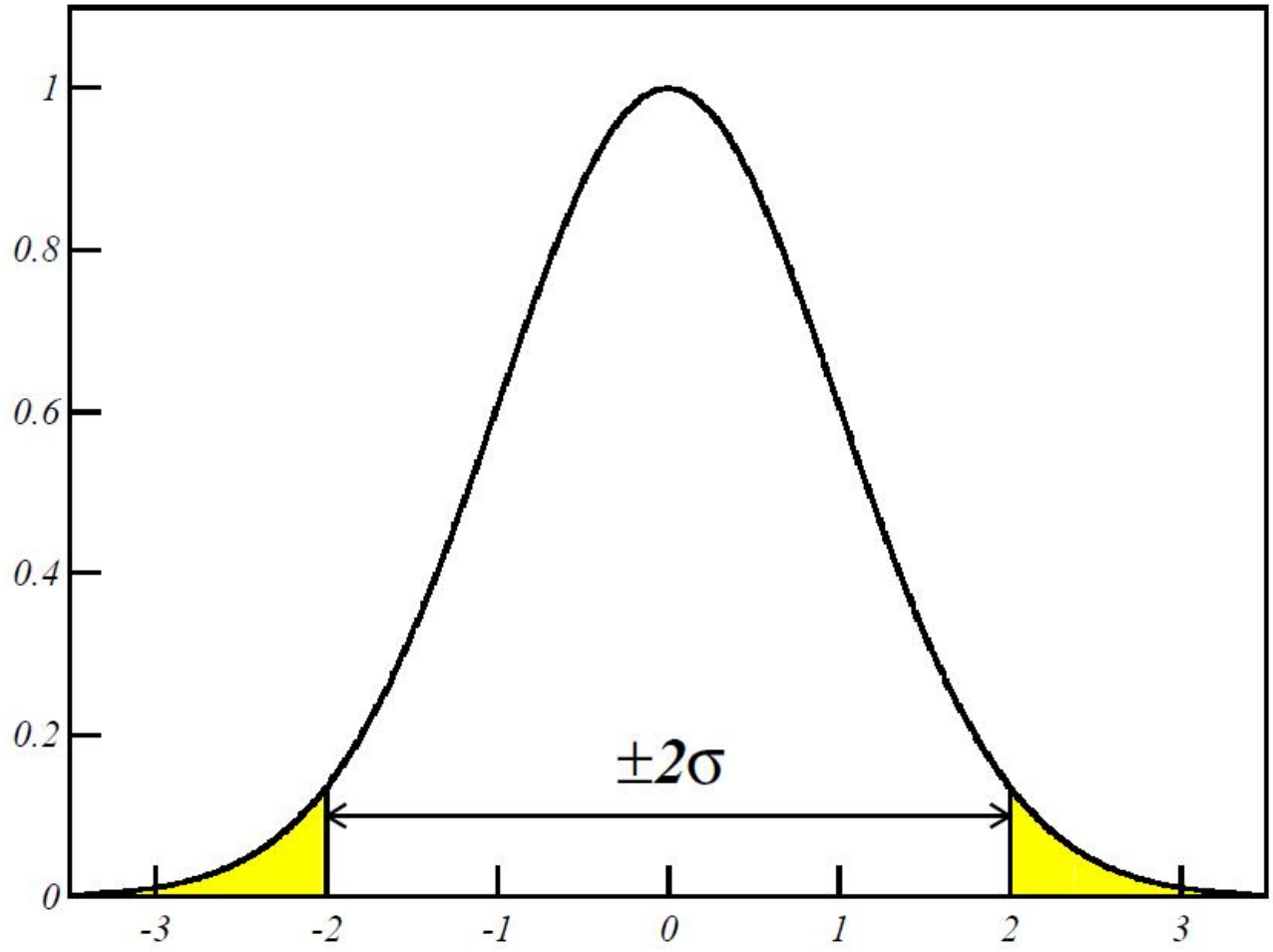
Quantile

(Квантиль)

- Quantile of level α – is a value which is not exceeded by the random variable with the probability α.
- Example: if 90% of people own less than 1M$, this means that 1M$ is the 90%-quantile of people's wealth
- For the normal distribution it is convenient to use 2-sided quantiles, i.e. the distance from the average which is not exceeded with probability α:

$$\alpha = P(|x-\mu| < k\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-k\sigma}^{\mu+k\sigma} \mathrm{e}^{-(x-\mu)^2/2\sigma^2} dx$$

Example of 2-sided quantile



Quantiles of the normal distribution

Quantile $\lvert x-\mu\rvert<k\sigma$	Level α
1 σ	68.27%
2 σ	95.42%
3 σ	99.73%
1.645 σ	90%
1.960 σ	95%
2.57 σ	99%
3.29 σ	99.9%

“So many sigma”

- If the measurement uncertainty has the normal distribution, then the measurement deviates from truth:
  - within $\pm 1\sigma$ – 68% probability
  - within $\pm 2\sigma$ – 95% probability
  - within $\pm 3\sigma$ – 99.7% probability
- Even in the case of non-Gaussian distributions, the probability is often converted into the quantile of the normal distribution. We say for example: “the effect is at 2.5σ level”.

Disproving a theory

- Suppose your experiment is inconsistent with the Energy Conservation law at the level of 3σ.
- Does it mean that the energy conservation is disproven with the probability 99.7%?
- To answer this question, we need a deeper understanding of the concept of probability

There are (at least) two interpretations of probability

- Frequentist interpretation
  - Do an experiment; repeat it many times; count the number of successful results; divide it by the number of trials
  - **p = Success/Total**
- But what if the experiment can not be repeated?
  - For example: what is the probability that tomorrow it will be rain? that Barcelona will win the next Champions League? that…

Bayesian probability

(байесовская вероятность)

- The degree of belief of someone (person, expert committee, the mankind) that certain event will happen.
- Numerically can be defined via an imaginary bet.
  - Are you ready to stake $100 against $10 that tomorrow it will be rain? OK, and what about staking $10 against $100?
- Bayesian probability is based on the full knowledge of the person who estimates it.
  - Therefore it is subjective – different for different people

Back to our example

- Your experiment deviates by 3σ (99.7%) from energy conservation.
- Are you ready to stake $1000 against $3 that Energy Conservation is disproven? I think, not.
- Your decision is based on your subjective knowledge that energy conservation is a robust and fundamental law. Most likely, the problem is with your experiment, rather than with the theory.

Bayes theorem

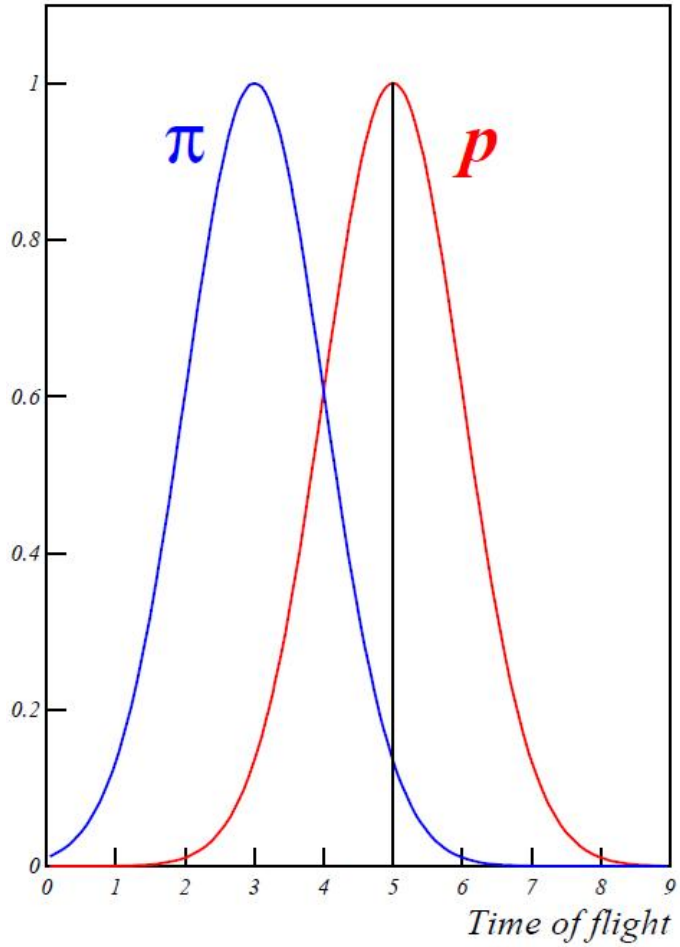
- For the events A and B :
- P(A&B) = P(A|B)P(B) = P(B|A)P(A)
- P(A|B) – *conditional* probability that A will happen if B has already happened
- The Bayes theorem:
- P(A|B) = P(B|A)P(A)/P(B)

P(A|B) = P(B|A)P(A)/P(B)

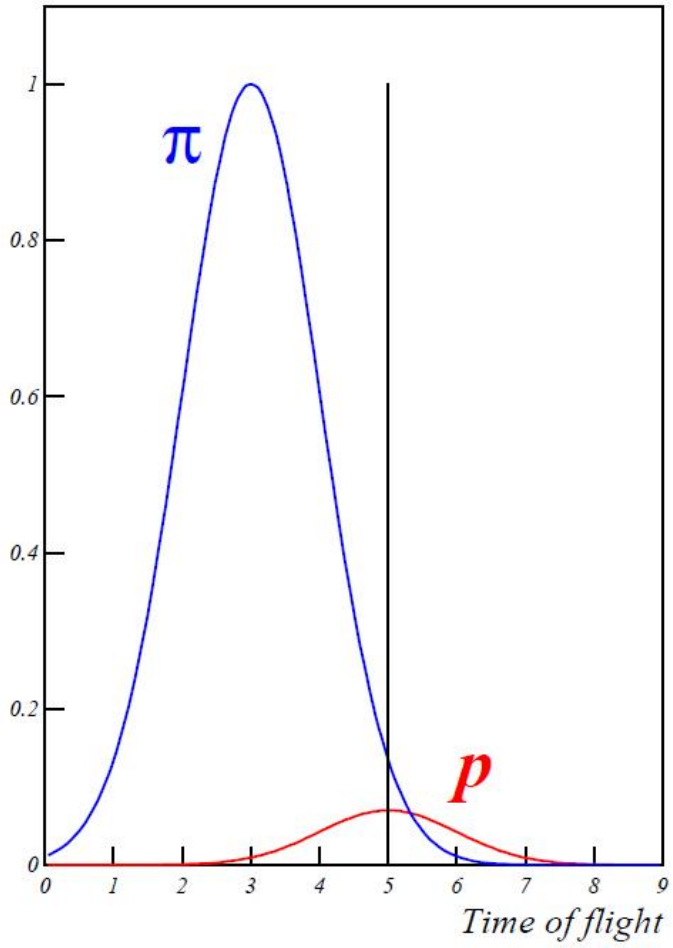
- Let A be some theory, B is an experiment which is testing the validity of this theory.
  - P(A) – the probability (our belief) ***before the experiment*** that A is correct.
  - P(B|A) – probability of the observed experimental result in the case ***if*** A is indeed correct.
  - P(A|B) – our belief in the correctness of A ***after*** the experiment.
- ***After*** the experiment the probability of the theory is proportional to [its probability before the experiment (prior probability)] ***times*** [the probability to observe what was observed in the experiment if our theory is indeed correct]

Practical example

- You detected a particle, and you are trying to identify it: is it a pion or a proton?
- The measured time-of-flight perfectly agrees with the expectation for proton. For a pion such TOF could be measured with just 10% probability.
- At first look, it is certainly a proton!



But what if there are MANY pions?



- Must take into account both prior probability (initial fractions of π and p) and the outcome of the experiment (measurement of the TOF)
- The posterior probability:

$$P_T(\pi) = \frac{N_\pi \cdot P_\pi(T)}{N_\pi \cdot P_\pi(T) \; + \; N_p \cdot P_p(T)}$$

A dramatic example

- Suppose there is a test for the Ebola decease, which never misses a real case of decease. Unfortunately, in 2% cases it gives positive result for a healthy person.
- You undergo the test, the result is positive. What is the probability that you are really sick? 98%?
- Correct answer: the provided input is insufficient! The average level of Ebola decease in you country must be specified.
- Homework: calculate probability that you are sick if 1% of the country population is infected.

(Slide from G.Cowan)

probability of the data assuming hypothesis H (the likelihood)

prior probability, i.e., before seeing the data

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

posterior probability, i.e., after seeing the data

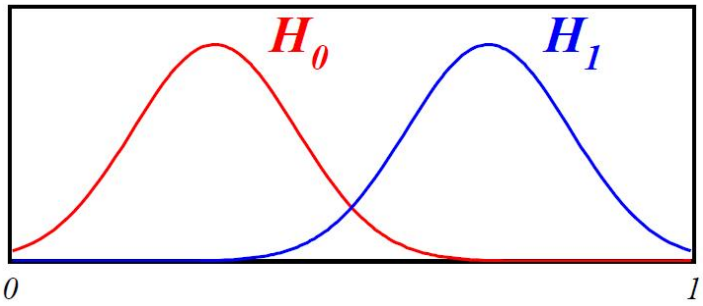
normalization involves sum over all possible hypotheses

Bayes' theorem has an "if-then" character: If your prior probabilities were $\pi(H)$, then it says how these probabilities should change in the light of the data.

No unique prescription for priors (subjective!)

Hypothesis testing

- Hypothesis testing is a rule or an algorithm by which we either accept a hypothesis H_0 or reject it in favor of an alternative hypothesis H_1
- A hypothesis can not be “rejected in general”, only in favor of some alternative hypothesis!



Test statistics

- In particle physics typically H_0 denotes “signal does exist!”, H_1 is for “only background is present”.
- Usually, for hypothesis testing we use not the whole sample of measurements (millions of events!), rather we use the ***test statistics*** which is an “extraction” from measurements, typically one or several values that are most sensitive to the hypothesis.
- Example of a simple test statistics: the number of events that passed the selection criteria.

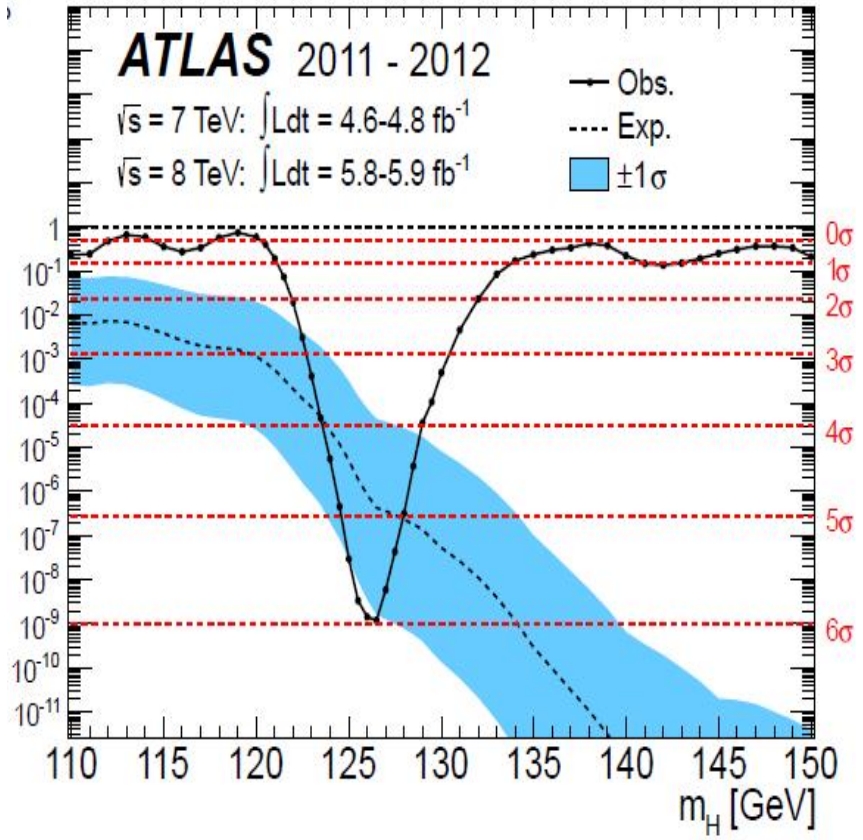
p-value

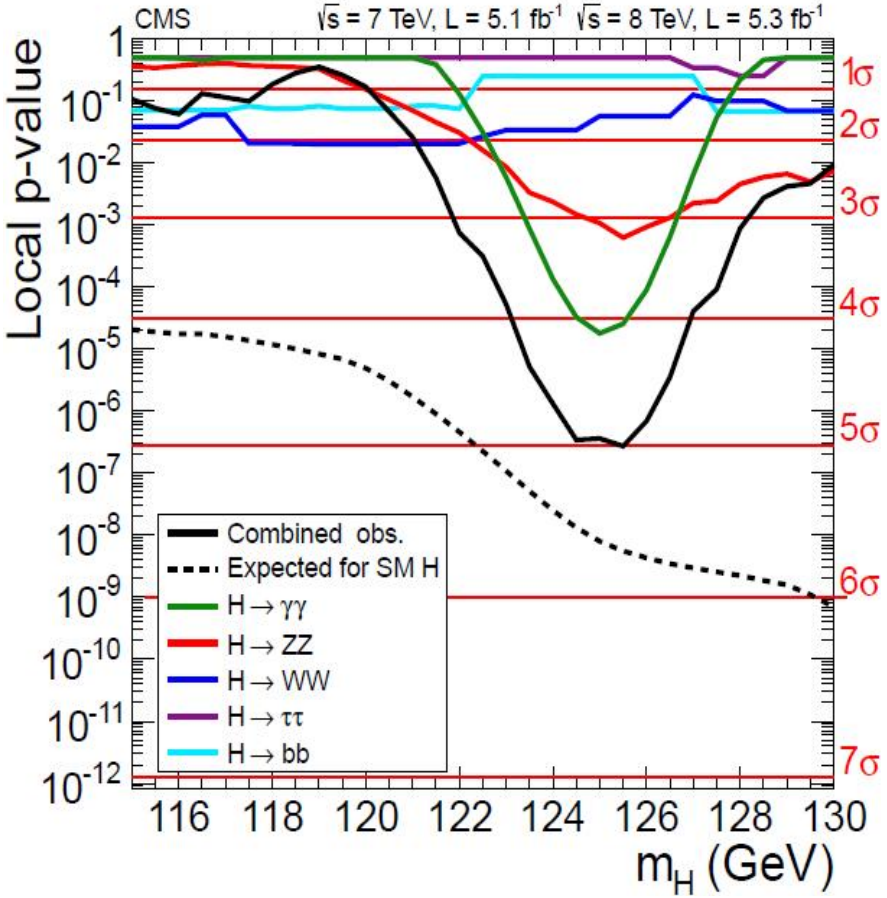
- Sometimes we would like to check how well H_0 agrees with data – regardless of any alternative H_1
- Let's introduce a test statistics t with PDF distribution $f(t|H_0)$ – such that (for example) large t correspond to poor agreement of hypothesis with data. From our measurements we observed the value t_{obs}
- Introduce the **p-value**:

$$p = \int_{t_{obs}}^{\infty} f(t|H_0)\, dt$$

- p-value is **not** a probability of some hypothesys!! It is simply the ***probability to observe such or worse (dis)agreement*** between the data and the hypothesis
- For the true hypothesis the p-value distribution is uniform between 0 and 1 (from definition!)

2012: p-value of the hypothesis “Higgs boson of given mass does not exist”

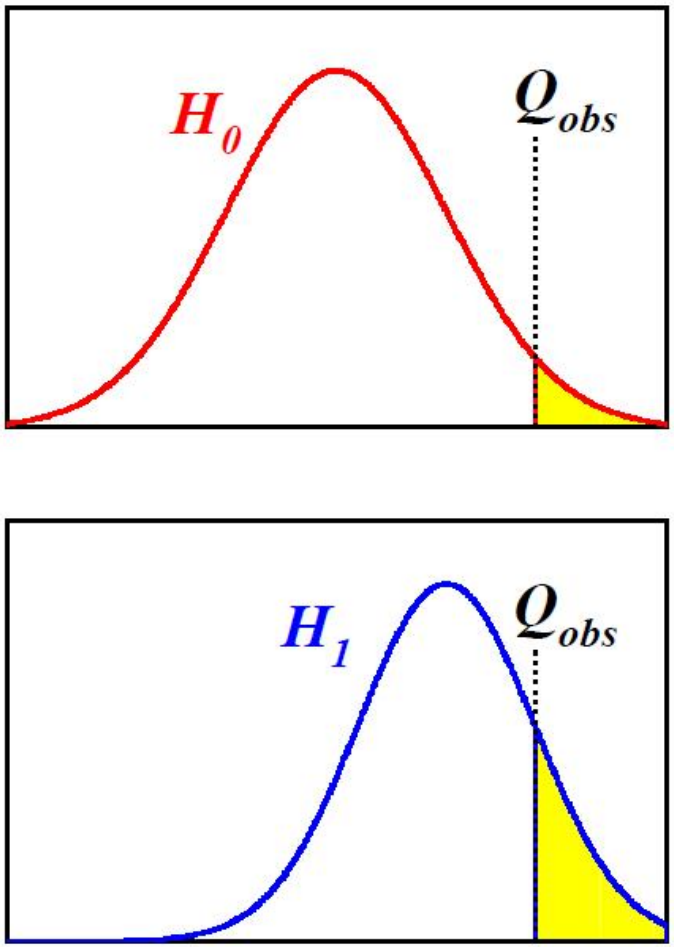




ATLAS: 6.0σ $(3 \cdot 10^{-9})$ **CMS: 5.0σ $(3 \cdot 10^{-7})$**

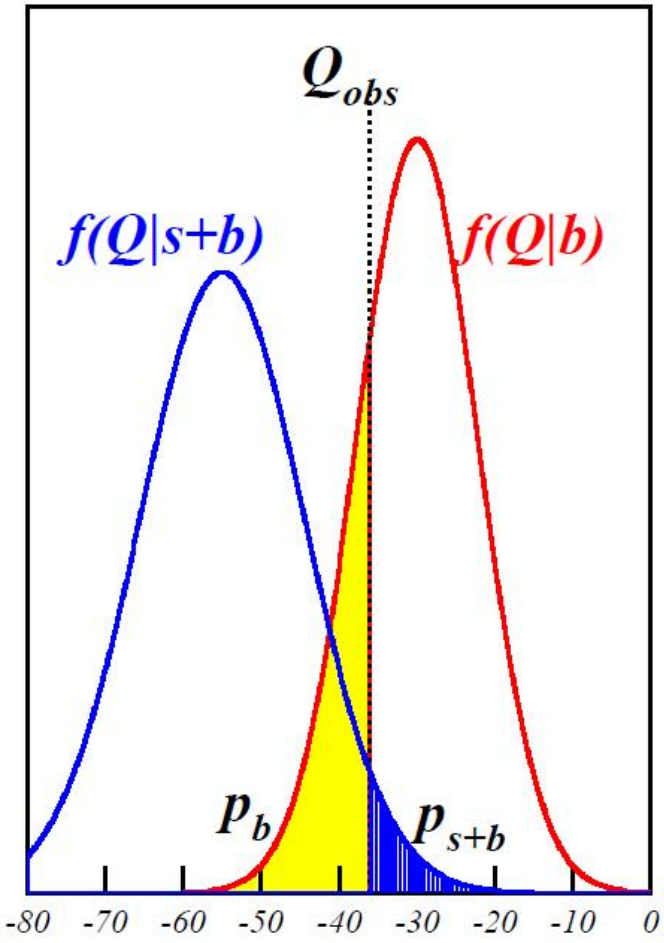
Rejection of a hypothesis

- Suppose, p-value of the hypothesis H_0 is small (poor agreement with data).
- So, we reject this hypothesis??
- But: what if the data disagree not only H_0, but also the alternative H_1?
- This means that our data simple have poor sensitivity, they do not allow to exclude any of the hypotheses.
- The poor agreement with both hypotheses is (most likely) due to a fluctuation



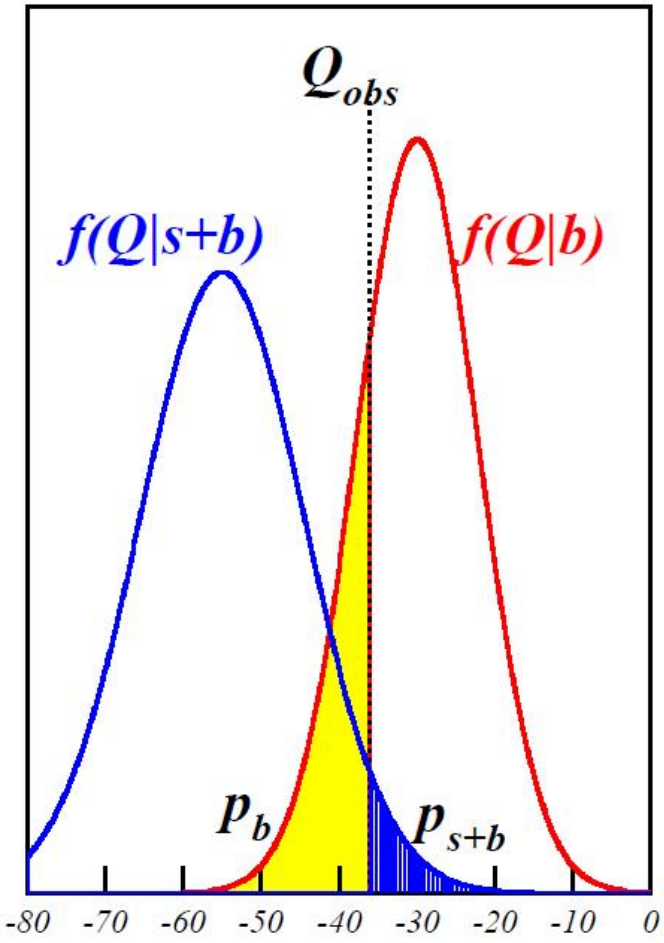
CL_s Method

- Suppose we have 2 hypotheses: “b” (background-only) and “s+b” (signal exists, together with the background).
- We introduce the test statistics $Q=-2\ln(L_{s+b}/L_b)$.
- L is the probability to observe exactly what we did observe, calculated under each hypothesis
- The value Q is expected to be large for background and small for signal

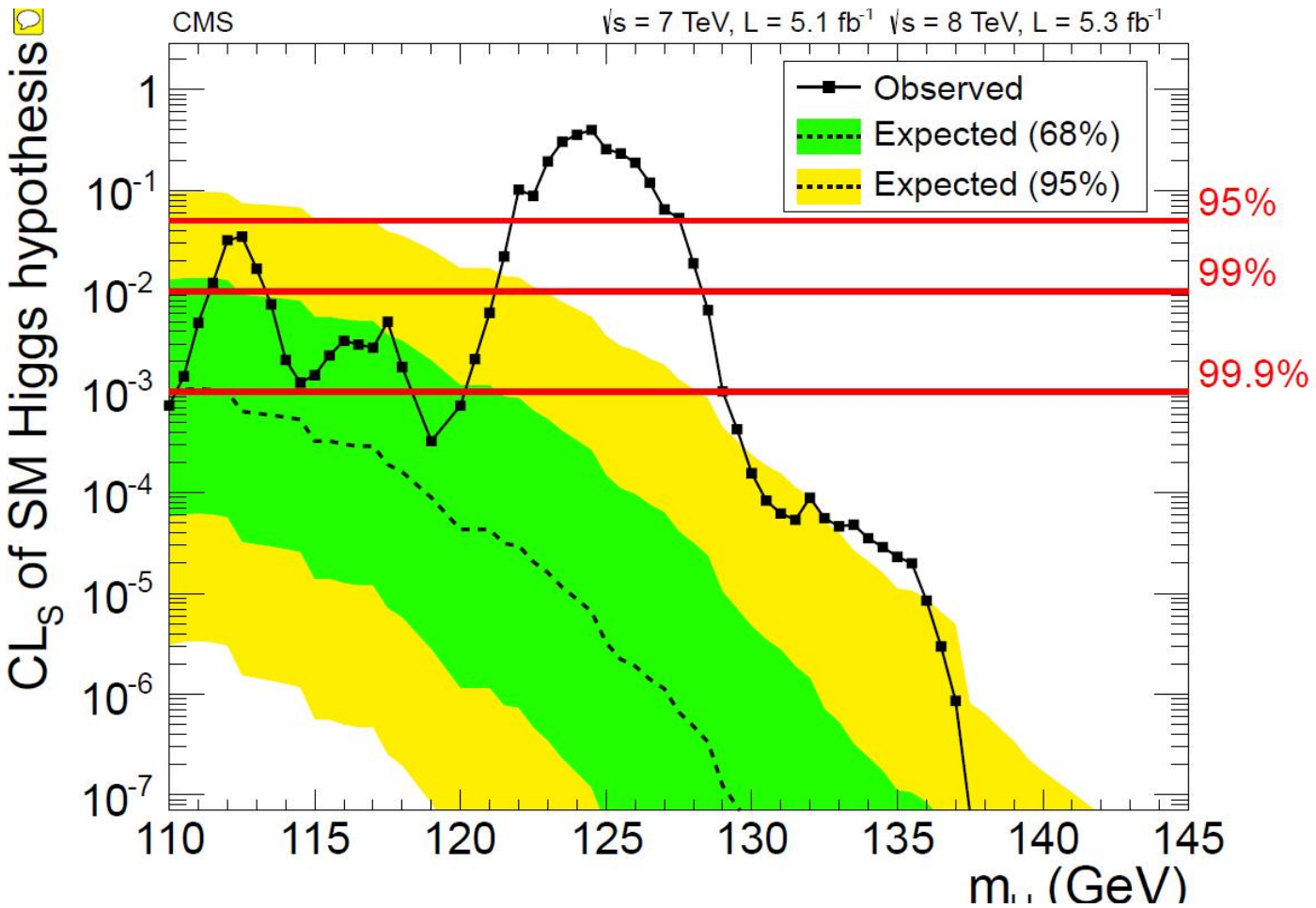


CL_s Method

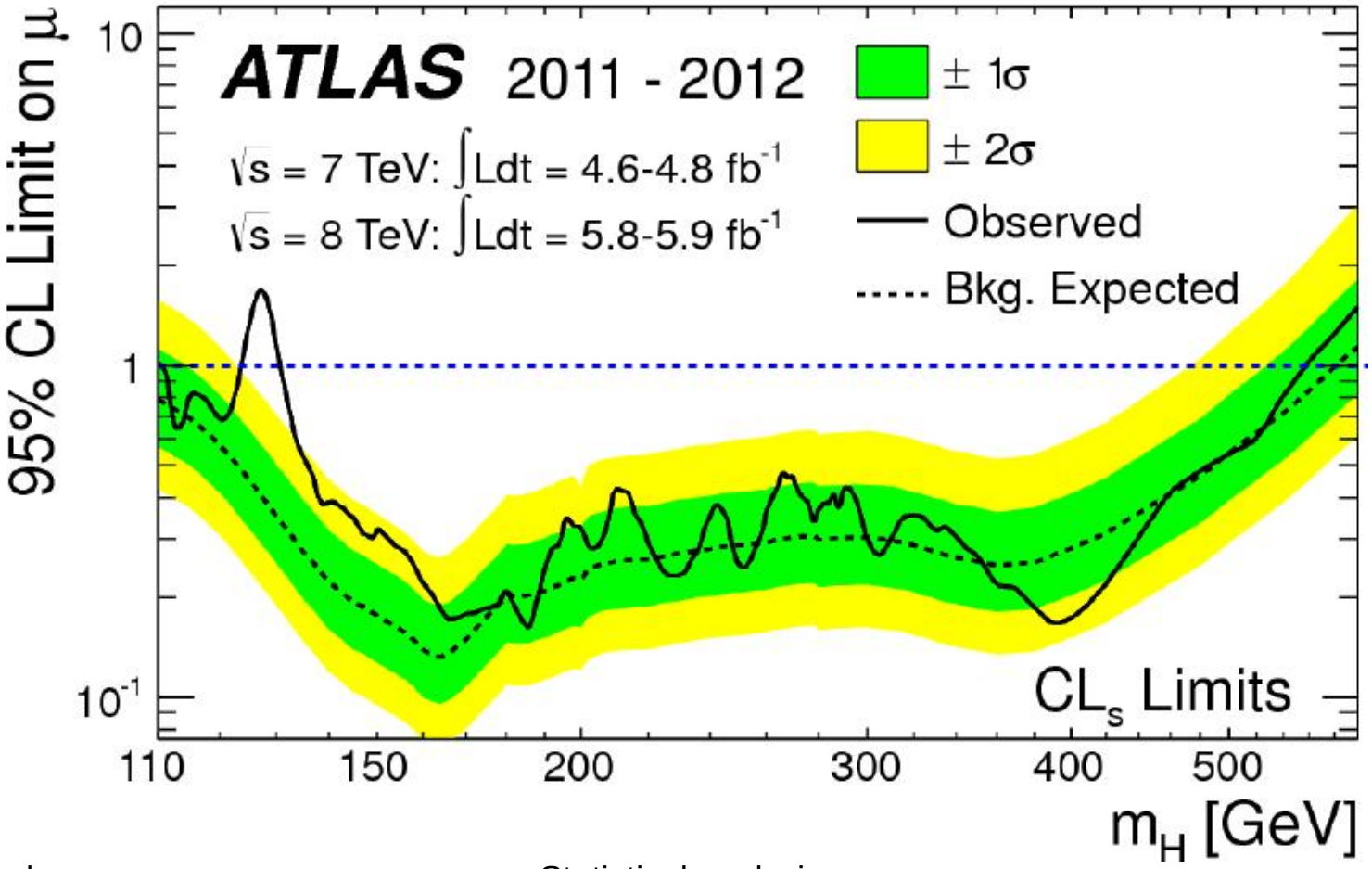
- For the measured Q_{obs} we define:
  - $CL_b=P(Q>Q_{obs}|b)=1-p_b$
  - $CL_{s+b}=P(Q>Q_{obs}|s+b)=p_{s+b}$
- Now calculate **CL_s** = $CL_{s+b}/CL_b = p_{s+b}/(1-p_b)$
- If $CL_s<\alpha$, then the hypothesis s+b is rejected “at the confidence level $1-\alpha$”



Higgs boson search



- In Higgs searches we allow that the theoretically predicted signal might be corrected by a factor μ: instead of s+b we would observe μs+b. Here the curve shows the upper limit on μ at 95% Confidence Level.



What CL to use?

- When we publish a paper, should we quote results at 95% (2σ) Confidence Level? 99.7% (3σ)? $1 - 3\times10^{-7}$ (5σ)?
- The issue is psychological, not physical.
- If you claim a fundamental discovery which turns to be wrong, you are in a very bad position.
- If you say there is no signal but it actually exists – it's a pity, but not a major problem.
- Therefore the standard CL are: 5σ if you claim a *discovery*, and 2σ if you report an *absence* of a signal