

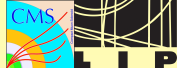
# Course on Physics at the LHC - Statistics - or “How to find answers to your questions”

Pietro Vischia<sup>1</sup>, [pietro dot vischia at cern dot ch](mailto:pietro_dot_vischia_at_cern_dot_ch)

<sup>1</sup>LIP-Lisboa, with funding by FCT grant SFRH/BD/52067/2012 (IDPASC program)

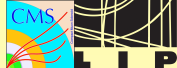
- What is the chance of obtaining a 1 when throwing a six-faced die?
- What is the chance of tomorrow being rainy?

- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?



- What is the chance of obtaining a 1 when throwing a six-faced die?
  - We can throw a dice 100 times, and count how many times we obtain 1
- What is the chance of tomorrow being rainy?
  - We can try to answer building a function of the recent weather, but we cannot repeat the experiment

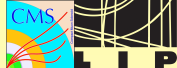
## What is a “probability”?



- $\Omega$ : set of all possible elementary (exclusive) events  $X_i$
- Exclusivity: the occurrence of one event implies that none of the others occur
- Probability then is any function that satisfies Kolmogorov axioms:
  - $P(X_i) \geq 0, \forall i$
  - $P(X_i \text{ or } X_j) = P(X_i) + P(X_j)$
  - $\sum_{\Omega} P(X_i) = 1$



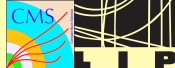
Kolmogorov, with the happy face of someone that has just created some awesome axioms



- The most familiar one: based on the possibility of repeating an experiment many times
- Consider one experiment in which a series of  $N$  events is observed.
- $n$  of those  $N$  events are of type  $X$
- Frequentist probability for any single event to be of type  $X$  is the empirical limit of the frequency ratio:

$$P(X) = \lim_{N \rightarrow \infty} \frac{n}{N}$$

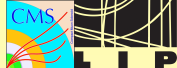
## Frequentist probability - 2



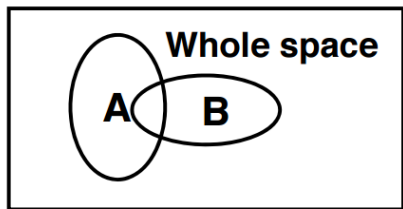
- The experiment must be repeatable in the same conditions
- The job of the physicist is making sure that all the *relevant* conditions in the experiments are the same, and to correct for the unavoidable changes.
  - Yes, *relevant* can be a somehow fuzzy concept
- In some simple cases, you can directly build the full table of frequencies (e.g. dice throws, poker)

Hand	Dist. (Hands)	Frequency	Probability	Cumulative probability	Odds	Mathematical expression of absolute frequency
<b>Royal flush</b> 	1	4	0.000154%	0.000154%	649,739 : 1	$\binom{4}{1}$
<b>Straight flush (including royal flush)</b> 	9	36	0.00139%	0.0014%	72,382 : 1	$\binom{10}{1}\binom{4}{1} - \binom{4}{1}$
<b>Four of a kind</b> 	156	624	0.0240%	0.0255%	4,264 : 1	$\binom{13}{1}\binom{12}{1}\binom{4}{1}$
<b>Full house</b> 	156	3,744	0.1443%	0.17%	689 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2}$
<b>Flush (including royal flush and straight flush)</b> 	1,277	5,108	0.1965%	0.267%	508 : 1	$\binom{13}{5}\binom{4}{1} - \binom{10}{1}\binom{4}{1}$
<b>Straight (including royal flush and straight flush)</b> 	10	10,200	0.3924%	0.76%	254 : 1	$\binom{10}{1}\binom{4}{1}^5 - \binom{10}{1}\binom{4}{1}$
<b>Three of a kind</b> 	858	64,812	2.1128%	2.87%	46.3 : 1	$\binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2$
<b>Two pair</b> 	858	129,862	4.7639%	7.62%	20.6 : 1	$\binom{13}{2}\binom{4}{2}^2\binom{11}{1}\binom{4}{1}$
<b>One pair</b> 	2,868	1,099,200	42.2669%	49.9%	1.87 : 1	$\binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$
<b>No pair / High card</b> 	1,277	1,302,540	60.1177%	100%	0.995 : 1	$\left[\binom{13}{5} - 10\right] \left[\binom{4}{1}^5 - 4\right]$
<b>Total</b>	7,462	2,598,960	100%	—	0 : 1	$\binom{52}{5}$

## Conditional probabilities



- Probabilities can be combined to obtain more complex expressions



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

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

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

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

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

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

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

$$\Rightarrow P(B|A) = P(A|B) \times P(B) / P(A)$$

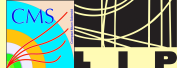


$$P(A|B) = \frac{\text{small blue oval}}{\text{large blue oval}}$$

$$P(B|A) = \frac{\text{small blue oval}}{\text{small blue oval}}$$

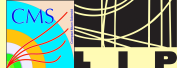
- **Conditional probabilities are not commutative!**  $P(A|B) \neq P(B|A)$
- Example from Louis Lyons:
  - $A$ : being female
  - $B$ : being pregnant
- The probability for a female to be pregnant,  $P(\text{pregnant}|\text{female})$ , is roughly 3%
- The probability for a pregnant person to be female,  $P(\text{female}|\text{pregnant})$  is unarguably  $\ggggg$  3% ☺

## Example: the Monty Hall problem



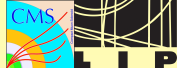
- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?

## Example: the Monty Hall problem



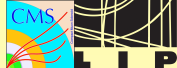
- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
- ALWAYS SWITCH, DUDE!

## Example: the Monty Hall problem

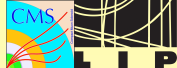


- Suppose you're on a game show, and you're given the choice of three doors
  - Behind one door is a car;
  - behind the others, goats.
- You pick a door, say No. 1, and the host, who knows what is behind the doors, opens another door, say No. 3, which has a goat.
- He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?
- ALWAYS SWITCH, DUDE!

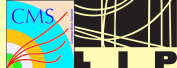
Behind door 1	Behind door 2	Behind door 3	Result if staying at door #1	Result if switching to the door offered
<b>Car</b>	Goat	Goat	<b>Wins Car</b>	Wins Goat
Goat	<b>Car</b>	Goat	Wins Goat	<b>Wins Car</b>
Goat	Goat	<b>Car</b>	Wins Goat	<b>Wins Car</b>



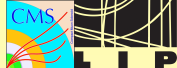
- Based on the concept of *degree of belief*
- Operational definition (by de Finetti) based on the *coherent bet*
  - Goal: determine  $P(X)$
  - Assume that if you bet on  $X$ , you win a fixed amount if  $X$  later occurs, and nothing if it doesn't
  - $P(X) = \frac{\text{largest amount you would be willing to bet}}{\text{the amount you stand to win}}$
- Surprisingly, it obeys all the Kolmogorov axioms! It is a probability.



- Is is as much a property of the observer as it is of the system being observed
- It depends on the state of the observer's *prior* knowledge, and will in general change as the observer obtains more knowledge
- This is the so-called subjective probability. There is also an *objective Bayesian probability*, but professional statisticians are not satisfied with its theory



- Which is the connection of Bayes theorem with Bayesian probability?
- Bayes theorem,  $P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$ , holds for any generic probability
- In frequentist probability,  $A$  and  $B$  are sets of events
- In bayesian probability,  $A$  or  $B$  are a set of hypotheses  $\theta_i$ 
  - $P(\theta_i)$  represents the *degree of belief* in hypothesis  $\theta_i$
  - Hypotheses are not random variables, so frequentist approach is not possible!
  - Bayes theorem involving hypotheses can only be applied in Bayesian framework!

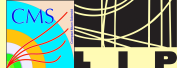


- Bayesian case:  $P(\theta_i|X^o) = \frac{P(X^o|\theta_i) \times P(\theta_i)}{P(X^o)}$ 
  - $\mathbf{P}(\theta_i|\mathbf{X}^o)$ : *posterior probability* for hypothesis  $\theta_i$ , given the observed data
  - $\mathbf{P}(\mathbf{X}^o|\theta_i)$ : probability of obtaining the observed data given the hypothesis  $\theta_i$ . It must be known (it is essentially a description of the behaviour of the experimental apparatus)
  - $\mathbf{P}(\theta_i)$ : *prior probability*, representing the knowledge or degree of belief in different hypotheses before the experiment is performed
  - $\mathbf{P}(\mathbf{X}^o)$ : it can be seen as a normalization constant (since the sum over  $i$  of the left side must be unity if the hypotheses form a complete and exclusive set)



- Frequentists are restricted to statements related to
  - $P(data|theory)$  (kind of deductive reasoning)
  - The data is considered random
  - Each point in the “theory” phase space is treated independently (no notion of probability in the “theory” space)
  - Repeatable experiments
- Bayesians can address questions in the form
  - $P(theory|data) \propto P(data|theory) \times P(theory)$  (it is intuitively what we normally would like to know)
  - It requires a prior on the theory
  - Huge battle on subjectiveness in the choice of the prior goes here - see §7.5 of James book

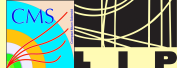
## A joke to (hopefully) keep you awake...



- Frequentists use impeccable logic to deal with an issue of no interest to anyone
- Bayesians address the question everyone is interested in, by using assumptions no-one believes

P. G. Hamer

## Speaking of questions everyone is interested in...



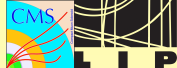
- Diagnostic example (Michael Goldstein)
- There is a deadly illness
  - D: you are diseased
  - H: you are healthy
- There is a diagnostic test
  - +: you are positive
  - -: you are negative
- It catches almost all the sick people:  $P(+|D) = 0.99$
- It take in a small number of false positives:  $P(+|H) = 0.01$
- You test is positive. Are you fucked up?

- Diagnostic example (Michael Goldstein)
- There is a deadly illness
  - D: you are diseased
  - H: you are healthy
- There is a diagnostic test
  - +: you are positive
  - -: you are negative
- It catches almost all the sick people:  $P(+|D) = 0.99$
- It take in a small number of false positives:  $P(+|H) = 0.01$
- You test is positive. Are you fucked up?
- You need to know that the incidence of the disease is 1 out of 1000 people!  $P(D) = 0.001$

- Diagnostic example (Michael Goldstein)
- There is a deadly illness
  - D: you are diseased
  - H: you are healthy
- There is a diagnostic test
  - +: you are positive
  - -: you are negative
- It catches almost all the sick people:  $P(+|D) = 0.99$
- It take in a small number of false positives:  $P(+|H) = 0.01$
- You test is positive. Are you fucked up?
- You need to know that the incidence of the disease is 1 out of 1000 people!  $P(D) = 0.001$
- Then, Bayes theorem says:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{[P(+|D)P(D)+P(+|H)P(H)]} = 0.09$$

## On reporting the result of a measurement - 1



- Let's measure the acceleration  $g$  due to gravity, that we know from super-pro experiments to be  $9.81\text{ m/s}^2$ !
  - We can do that by using a pendulum
- Say we obtain, as a result, 9.70
- Are we happy? Why?

- Let's measure the acceleration  $g$  due to gravity, that we know from super-pro experiments to be  $9.81 m/s^2$ 
  - We can do that by using a pendulum
- Say we obtain, as a result, 9.70
- Are we happy? Why?
  - We are not happy. Any measurement is subject to uncertainties (detector inefficiencies, effects due to the limited amount of events available, and so on) that limit its accuracy. We want to quote an experimental error on the quantity of interest, as an expression of the accuracy of our measurement

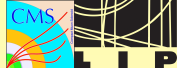
- Let's measure the acceleration due to gravity, that we know from super-pro experiments to be  $9.81 m/s^2$ !
- What can we say in the different cases?
- Say we obtain, as a result,  $9.70 \pm 0.15$
  
- Say we obtain, as a result,  $9.70 \pm 0.01$
  
- Say we obtain, as a result,  $9.70 \pm 5$



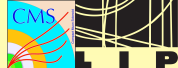
- Let's measure the acceleration due to gravity, that we know from super-pro experiments to be  $9.81 m/s^2$ !
- What can we say in the different cases?
- Say we obtain, as a result,  $9.70 \pm 0.15$ 
  - It is compatible with the “known” value, assumed not affected by error (or with negligible error)
- Say we obtain, as a result,  $9.70 \pm 0.01$
  
- Say we obtain, as a result,  $9.70 \pm 5$

- Let's measure the acceleration due to gravity, that we know from super-pro experiments to be  $9.81\text{ m/s}^2$ !
- What can we say in the different cases?
- Say we obtain, as a result,  $9.70 \pm 0.15$ 
  - It is compatible with the “known” value, assumed not affected by error (or with negligible error)
- Say we obtain, as a result,  $9.70 \pm 0.01$ 
  - It is incompatible with the “known” value. If our experiment is OK, we have made an earth-shattering discovery!
- Say we obtain, as a result,  $9.70 \pm 5$

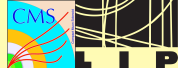
- Let's measure the acceleration due to gravity, that we know from super-pro experiments to be  $9.81 m/s^2$ !
- What can we say in the different cases?
- Say we obtain, as a result,  $9.70 \pm 0.15$ 
  - It is compatible with the “known” value, assumed not affected by error (or with negligible error)
- Say we obtain, as a result,  $9.70 \pm 0.01$ 
  - It is incompatible with the “known” value. If our experiment is OK, we have made an earth-shattering discovery!
- Say we obtain, as a result,  $9.70 \pm 5$ 
  - It is compatible with the “known” value, and with too many other values. We should set up a better experiment!



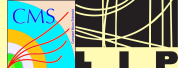
- You got a value incompatible with previous prestigious experiments. What now?
  - “The result is wrong, please repeat experiment until you get the correct result”
  - “The result might be wrong, please recheck the full procedure, possibly coming up with alternative measurements. If no issue is found, quote the current value. After all, the previous experiments might have issues”
  - “The result is correct, let’s just publish it and claim previous experiments have issues”



- You got a value incompatible with previous prestigious experiments. What now?
  - “The result is wrong, please repeat experiment until you get the correct result”
    - **WRONG ATTITUDE, MATE.** That’s the worst you can do (more on that when speaking about coverage)
  - “The result might be wrong, please recheck the full procedure, possibly coming up with alternative measurements. If no issue is found, quote the current value. After all, the previous experiments might have issues”
  - “The result is correct, let’s just publish it and claim previous experiments have issues”

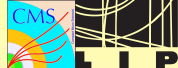


- You got a value incompatible with previous prestigious experiments. What now?
  - “The result is wrong, please repeat experiment until you get the correct result”
    - **WRONG ATTITUDE, MATE.** That’s the worst you can do (more on that when speaking about coverage)
  - “The result might be wrong, please recheck the full procedure, possibly coming up with alternative measurements. If no issue is found, quote the current value. After all, the previous experiments might have issues”
    - Healthy attitude
  - “The result is correct, let’s just publish it and claim previous experiments have issues”

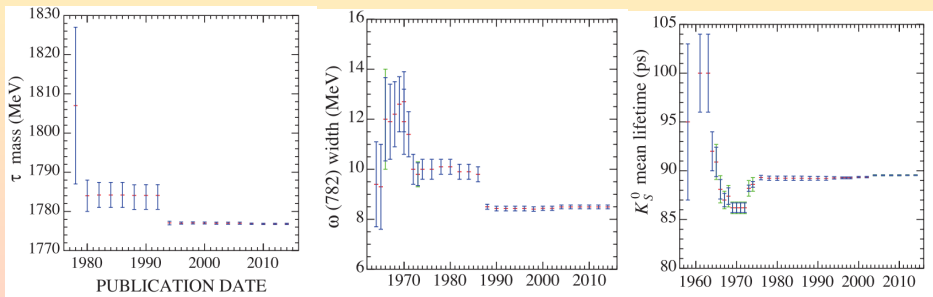


- You got a value incompatible with previous prestigious experiments. What now?
  - “The result is wrong, please repeat experiment until you get the correct result”
    - **WRONG ATTITUDE, MATE.** That’s the worst you can do (more on that when speaking about coverage)
  - “The result might be wrong, please recheck the full procedure, possibly coming up with alternative measurements. If no issue is found, quote the current value. After all, the previous experiments might have issues”
    - Healthy attitude
  - “The result is correct, let’s just publish it and claim previous experiments have issues”
    - Borderline, but not advisable. You should really crosscheck your result. Reporting the value can be done without claiming that previous measurements have issues (you cannot be certain of that, so you cannot claim it)

## On the dangers of “compatibility with previous values” - 2



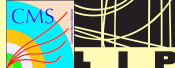
- A bit of history shows that biases in measurements are pretty common in history
  - Some of those are just the effect of high-precision experiments popping up, or older data being discarded, beware



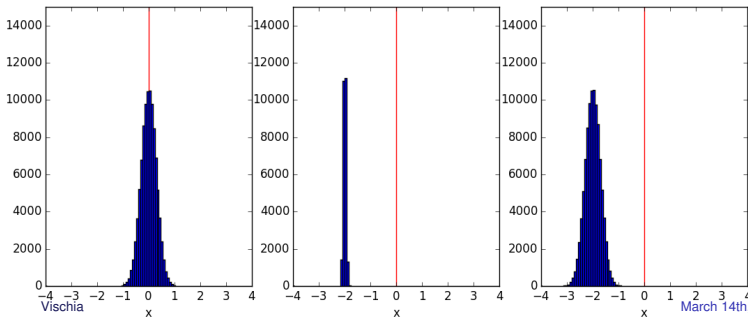
Plots from <http://pdg.lbl.gov/2015/reviews/rpp2015-rev-history-plots.pdf>



## So, what about these “errors”?

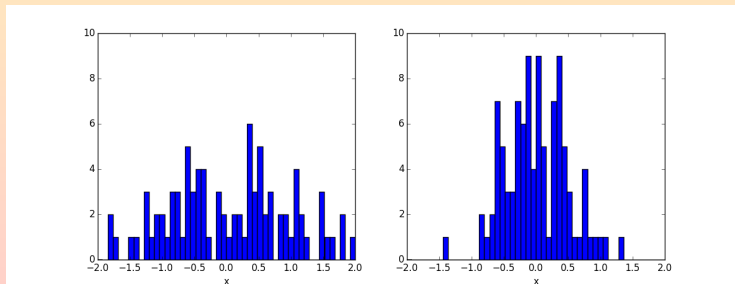


- Two fundamentally different kinds of error:
  - **Random (statistical) errors**
    - Inability of any measuring device (and scientist) to give infinitely accurate answers
    - Even for integral quantities (e.g. counting experiments), fluctuations occur in observations on a small sample drawn from a large population
    - They manifest as spread of answers scattered around the true value
  - **Systematic errors**
    - They result in measurements that are simply wrong, for some reason
    - They manifest usually as offset from the true value, even if all the individual results can be consistent with each other



- An approach based on repeated measurements will not work: systematic errors are sneaky
  - If you measure a resistance with an ohmeter reading in kOhms while we thing he reads in Ohms, you will fuckup of a factor 1000 every single time, yet everything will seem consistent
- If you suspect something might be off, you can devise cross-checks (e.g. measure a known resistor with your ohmeter)
- There is no general prescription on how to deal with systematic errors
  - To a large extent, it requires common sense plus experience
- Normally once a source of systematic error is identified, one can just correct for it
  - Sometimes (often, in HEP) the correction factor itself is affected by systematic and/or random error, and ultimately the final estimate will have to have a systematic error associated to it

- The notion of random error is related to the concept of spread of values obtained from a set of repeated experiments
- A *distribution*  $n(x)$  describes how often a value of the variable  $x$  occurs in a defined sample
- We need a way of synthesizing the information given by the distribution, typically into
  - The value at which the distribution is centered
  - How widely spread are the values around that central value



- As a physicist you may often think that nature can be described by a continuous probability distribution  $f(X)$ 
  - In this view, our distributions are discrete because of the finiteness of our scan of the variable  $X$
- One can write  $f(X) = \lim_{\Delta X \rightarrow 0} \frac{P(X)}{\Delta X}$ 
  - From dimensional arguments,  $f(X)$  is a *probability density function* (p.d.f.)
  - Can be extended to an arbitrary number of independent variables  $f(X, Y, Z, \dots)$

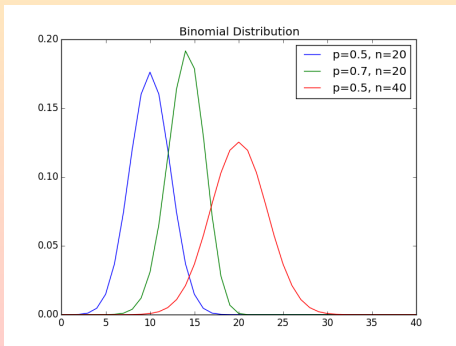
- Take any given function  $g(X)$  of a random variable with p.d.f.  $f(X)$
- You can obtain the average value of the function simply by weighting it by the p.d.f.!
- **Expectation value:**  $E(g) = \int_{\Omega} g(X)f(X)dX$ 
  - $g(X) = X$  is a legitimate function!
  - The expectation of  $X$  is called *mean of the density  $f(X)$* , or *expected value of  $X$* , denoted by  $\mu = \int_{\Omega} Xf(X)dX$
  - The expectation value of  $(X - \mu)^2$  is called *variance  $V(X)$  of the density  $f(X)$* , denoted often by  $\sigma^2$
  - $\sigma$  is called “standard deviation”, but at this level this is **just a definition**
  - Any connection between probability content and standard deviations require a large discussion on confidence intervals
- Not all distributions have a mean (e.g., the Cauchy distribution does not have a mean)

- Let's assume that there is an underlying population of infinite events, characterized by a continuous p.d.f.
- Our set of values for  $X$  can be seen as *sampling* from that population
- We want to obtain estimates for our sample mean and variance
- Sample mean  $\hat{X} = \frac{1}{N} \sum_{i=1}^N X_i$ 
  - $E(\hat{X}) = E(X)$ : unbiased!
  - $\sigma^2(\hat{X}) = \frac{1}{N} \sigma^2(X)$ : more data improve the accuracy of the estimate of the population mean!
- Sample variance would be :  $\hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$
- However,  $\mu$  is the unknown population mean: we have only sample-wise quantities available
- Sample variance:  $\hat{\sigma}_X^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x})^2$ 
  - $E(\hat{\sigma}_X^2) = \frac{N-1}{N} \sigma^2(X)$
  - Is this OK? Why?

- The sample variance is an estimator of the variance of the population
- The variance of the population is obviously independent on the size of our sample
- To obtain an unbiased estimate of the variance of the population, it is enough to change normalization factor
- Sample variance:  $\hat{\sigma}_X^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \hat{x})^2$ 
  - $E(\hat{\sigma}_X^2) = \frac{N-1}{N-1} \sigma^2(X) = \sigma^2(X)$ : unbiased!

## • Binomial

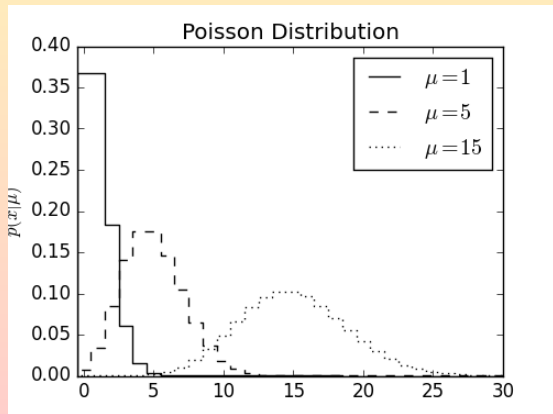
- Discrete variable:  $r$ , positive integer  $\leq N$
- Parameters:
  - $N$ , positive integer
  - $p$ ,  $0 \leq p \leq 1$
- Probability function:  $P(r) = \binom{N}{r} p^r (1-p)^{N-r}$ ,  $r = 0, 1, \dots, N$
- $E(r) = Np$ ,  $V(r) = Np(1-p)$
- Usage: probability of finding exactly  $r$  successes in  $N$  trials. The distribution of the number of events in a single bin of a histogram is binomial (if the bin contents are independent)





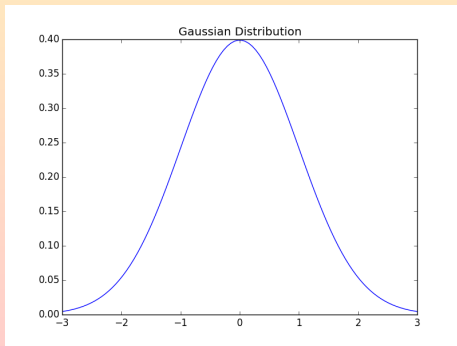
### ● Poisson

- Discrete variable:  $r$ , positive integer
- Parameter:  $\mu$ , positive real number
- Probability function:  $P(r) = \frac{\mu^r e^{-\mu}}{r!}$
- $E(r) = \mu$ ,  $V(r) = \mu$
- Usage: probability of finding exactly  $r$  events in a given amount of time, if events occur at a constant rate.

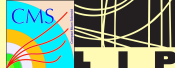


### ● Gaussian

- Variable:  $X$ , real number
- Parameters:
  - $\mu$ , real number
  - $\sigma$ , positive real number
- Probability function:  $f(X) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2} \frac{(X-\mu)^2}{\sigma^2}\right]$
- $E(X) = \mu$ ,  $V(X) = \sigma^2$
- Usage: describes the distribution of independent random variables. It is also the high-something limit for many other distributions

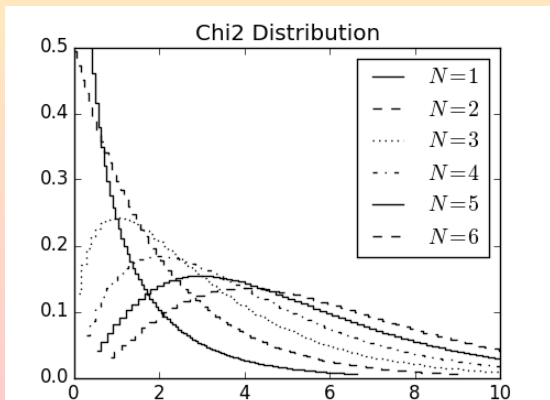


## A few notable distributions - 4

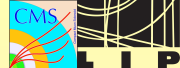


### • $\chi^2$ (Chi2)

- Variable:  $X$ , positive real number
- Parameters:
  - $N$ , positive integer (“degrees of freedom”)
- Probability function:  $f(X) = \frac{1}{2} \left(\frac{X}{2}\right)^{N/2-1} e^{-X/2} \frac{1}{\Gamma\left(\frac{N}{2}\right)}$
- $E(X) = N$ ,  $V(X) = 2N$
- Usage: describes the distribution of the sum of squares of a random variable,  $\sum_{i=1}^N X_i^2$



## Some relationship among distributions



- It is often convenient to know the asymptotic properties of the various distributions

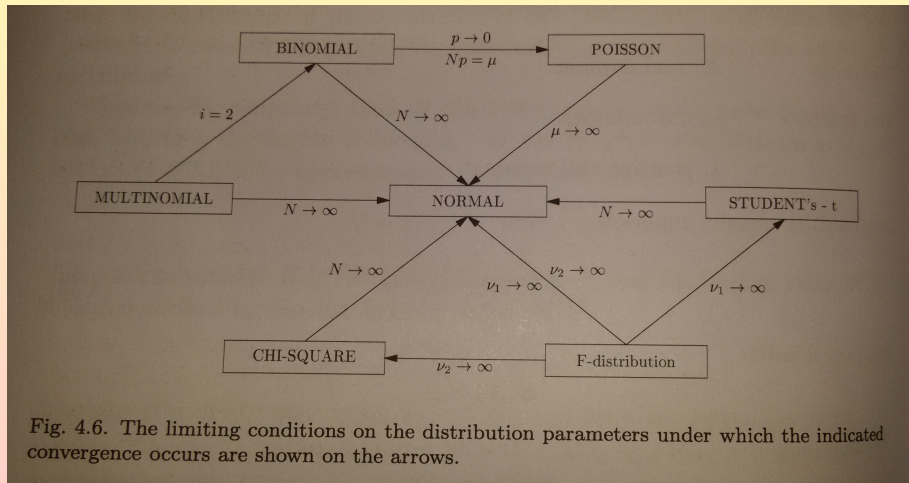
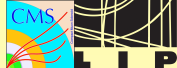


Fig. 4.6. The limiting conditions on the distribution parameters under which the indicated convergence occurs are shown on the arrows.

## Let's make it funnier: more variables!

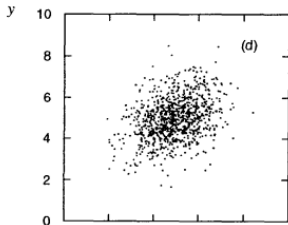
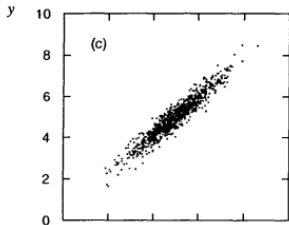
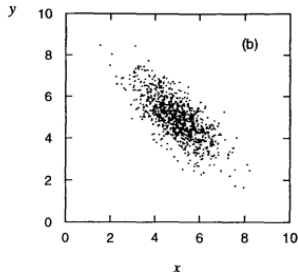
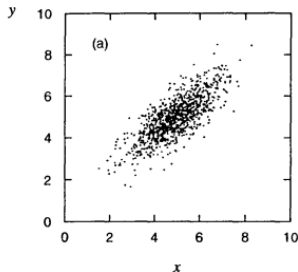


- Let our function  $g(\vec{X})$  be a function of more variables,  $\vec{X} = (X_1, X_2, \dots, X_n)$  (with p.d.f.  $f(\vec{X})$ )
- The expectation value for  $g(\vec{X})$  is:  
$$E(g(\vec{X})) = \int g(\vec{X})f(\vec{X})dX_1dX_2\dots dX_n = \mu_g$$
- The variance for  $g(\vec{X})$  is:  
$$V[g] = E[(g - \mu_g)^2] = \int (g(\vec{X}) - \mu_g)^2 f(\vec{X})dX_1dX_2\dots dX_n = \sigma_g^2$$
- But why limit ourselves to pitting each variable with itself?
- **Covariance:** of two variables  $X, Y$ :  
$$V_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X\mu_Y = \int XYf(X, Y)dXdY - \mu_X\mu_Y$$
  - It is also called “error matrix”, and sometimes denoted  $cov[X, Y]$
  - It is symmetric by construction:  $V_{XY} = V_{YX}$ , and  $V_{XX} = \sigma_X^2$
  - To have a dimensionless parameter: correlation coefficient  $\rho_{XY} = \frac{V_{XY}}{\sigma_X\sigma_Y}$

## Understanding covariance



- $V_{XY}$  is the expectation for the product of deviations of  $X$  and  $Y$  from their means
- If having  $X > \mu_X$  enhances  $P(Y > \mu_Y)$ , and having  $X < \mu_X$  enhances  $P(Y < \mu_Y)$ , then  $V_{XY} > 0$ : positive correlation!



## Mutual information: take it to the next level



- Covariance acts taking into account only the the first order in the expansion

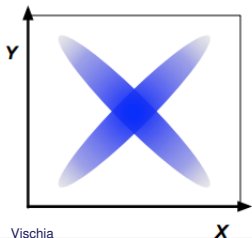
### Mutual Information



A more general notion of 'correlation' comes from **Mutual Information**:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p_1(x) p_2(y)} \right),$$
$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

- ▶ it is symmetric:  $I(X; Y) = I(Y; X)$
- ▶ if and only if  $X, Y$  totally independent:  $I(X; Y) = 0$
- ▶ possible for  $X, Y$  to be uncorrelated, but not independent



Mutual Information doesn't seem to be used much within HEP, but it seems quite useful

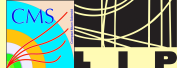
- Assume that we have  $n$  random variables  $\vec{X} = (X_1 \dots X_n)$  with p.d.f.  $f(\vec{X})$
- Assume that we do not know completely  $f(\vec{X})$ !!! We only know the mean values  $\mu_1 \dots \mu_n$  and the covariance matrix  $V_{ij}$
- Consider a function  $g(\vec{X})$ : if we want its p.d.f., in principle we could operate a change of variables using the Jacobian of  $g(\vec{X})$ 
  - But we do not know fully the p.d.f., so we cannot write analytically the Jacobian



- We can still expand  $g(\vec{X})$  to first order around the mean values  $\mu_i$ :  
$$g(\vec{X}) \simeq g(\vec{\mu}) + \sum_{i=1}^N \left[ \frac{\partial g}{\partial X_i} \right]_{X=\mu} (X_i - \mu_i)$$
- Expectation value of  $g$  at the first order:  $E[g(\vec{X})] \simeq g(\vec{\mu})$  (because  $E[X_i - \mu_i] = 0$ )
- Variance:  $\sigma_g^2 \simeq \sum_{i,j=1}^N \left[ \frac{\partial g}{\partial X_i} \frac{\partial g}{\partial X_j} \right]_{X=\mu} V_{[ij]}$
- Covariance (for set of functions  $g_1 \dots g_m$ ):  
$$\text{cov}[g_k, g_l] \simeq \sum_{i,j=1}^N \left[ \frac{\partial g_k}{\partial X_i} \frac{\partial g_l}{\partial X_j} \right]_{X=\mu} V_{[ij]}$$
- The variances are propagated from the  $X_i$  to the functions  $g_k$ , via the jacobian of the tranformation

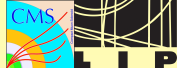
- From the general formula one can derive expressions for any possible transformation of the variables.
- A couple simple cases:
- $Y = X_1 + X_2$ 
  - $\sigma_Y^2 = \sigma_1^2 + \sigma_2^2 + 2V_{12}$
  - If variables are uncorrelated, the last term disappears ☺
- $Y = X_1 X_2$ 
  - $\frac{\sigma_Y^2}{Y^2} = \frac{\sigma_1^2}{X_1^2} + \frac{\sigma_2^2}{X_2^2} + 2 \frac{V_{12}}{X_1 X_2}$
  - If variables are uncorrelated, the last term disappears ☺

## What about combining the results of many experiments?



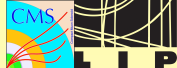
- Let's have several experiments measuring the same physical quantity, giving a set of answers  $a_i$  with errors  $\sigma_i$
- Best estimate of  $a$ :  $a = \frac{\sum a_i/\sigma_i^2}{\sum 1/\sigma_i^2}$
- Best estimate of the accuracy  $\sigma$ :  $\frac{1}{\sigma^2} = \sum 1/\sigma_i^2$
- The best estimate arises when each experiment is weighted by  $\frac{1}{\sigma_i^2}$ , which in some sense gives a measure of the information content of that particular experiment!

## A little exercise to fix ideas



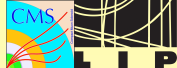
- We are trying to determine the number of *married people* in an underdeveloped country (where marriage is only between a man and a woman)
- We have performed two experiments, yielding the following results:
  - Number of married men:  $10.0 \pm 0.5$  million
  - Number of married women:  $8 \pm 3$  million
- How many married people are in the country?

## A little exercise to fix ideas



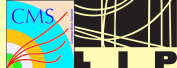
- We are trying to determine the number of *married people* in an underdeveloped country (where marriage is only between a man and a woman)
- We have performed two experiments, yielding the following results:
  - Number of married men:  $10.0 \pm 0.5$  million
  - Number of married women:  $8 \pm 3$  million
- How many married people are in the country?
- The number of married people is the sum of the married men and married women
  - Summing and propagating errors for the sum:  $18 \pm 3$  million. 17% accuracy

## A little exercise to fix ideas



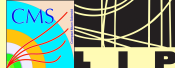
- We are trying to determine the number of *married people* in an underdeveloped country (where marriage is only between a man and a woman)
- We have performed two experiments, yielding the following results:
  - Number of married men:  $10.0 \pm 0.5$  million
  - Number of married women:  $8 \pm 3$  million
- How many married people are in the country?
- The number of married people is the sum of the married men and married women
  - Summing and propagating errors for the sum:  $18 \pm 3$  million. 17% accuracy
- Because of the laws of the country, the number of married men must be equal to the number of married women, meaning that each of those numbers is the estimate of the same physical quantity, “the half of the number of married people”
  - Combining the two experiments together:  $20 \pm 1$  million. 5% accuracy!
- Adding extra information improves the accuracy of the answer! It is all a matter of information!

## A dirty trick: the Central Limit Theorem - 1

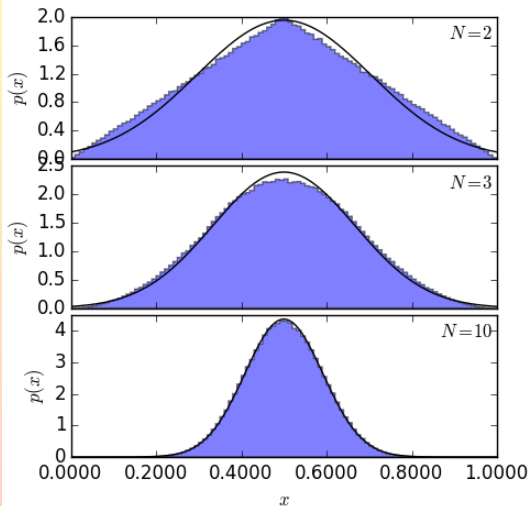


- Take a set  $\vec{X} = (X_1 \dots X_N)$  of independent variables following an *arbitrary distribution* with mean  $\mu$  and variance  $\sigma^2$ .
- Take their arithmetic mean:  $\hat{X} = \frac{1}{N} \sum_{i=1}^N X_i$
- In the limit  $N \rightarrow \infty$ , the mean follows a Gaussian distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{N}$
- How large do you think  $N$  needs to be to have a “good” approximation?

## A dirty trick: the Central Limit Theorem - 1

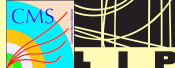


- Take a set  $\vec{X} = (X_1 \dots X_N)$  of independent variables following an *arbitrary distribution* with mean  $\mu$  and variance  $\sigma^2$ .
- Take their arithmetic mean:  $\hat{X} = \frac{1}{N} \sum_{i=1}^N X_i$
- In the limit  $N \rightarrow \infty$ , the mean follows a Gaussian distribution with mean  $\mu$  and variance  $\frac{\sigma^2}{N}$
- How large do you think  $N$  needs to be to have a “good” approximation?

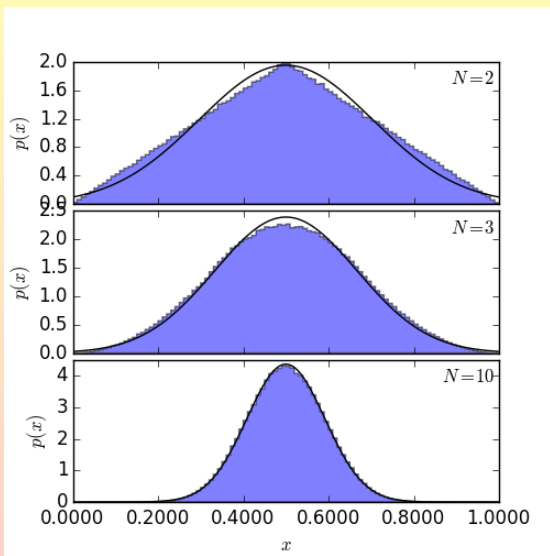




## A dirty trick: the Central Limit Theorem - 2



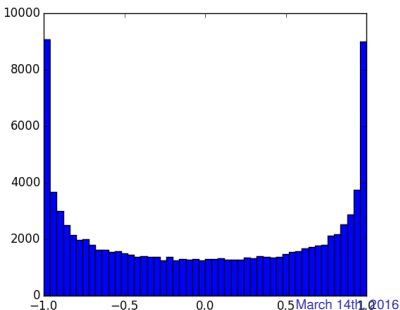
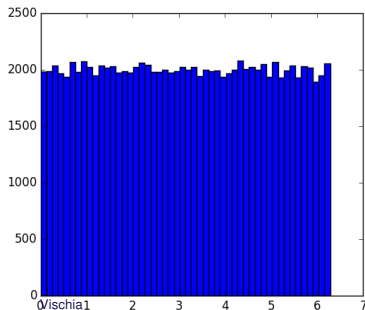
- Remarkable facts
- $N$  does not need to be really large
- The underlying p.d.f. of the variables **does not matter!** (this example: uniform)
- This is a very powerful technique to switch from any p.d.f. to a gaussian



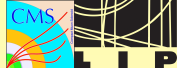
- Consider the p.d.f. for the independent variables  $\mathbf{X}$ , given the parameters  $\theta$
- $P(\mathbf{X}|\theta) = P(X_1, \dots, X_N|\theta) = \prod_{i=1}^N f(X_i|\theta)$
- Now, replace the variable  $\mathbf{X}$  with the *observed data*  $\mathbf{X}^O$
- $P$  is no longer a p.d.f., and is a function of  $\theta$  only.
- Likelihood function:  $L(\theta) = P(\mathbf{X}^O|\theta)$

## Likelihood function is awesome - 1

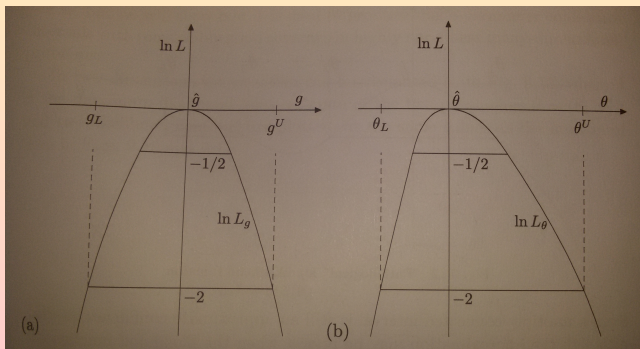
- Take a variable  $X$  with an uniform p.d.f.
- Transform it by  $Y = \cos(X)$
- The new p.d.f. is different from the beginning:  $P(Y(X)|\theta) = \frac{P(X|\theta)}{|dY/dX|}$
- But the jacobian transformation act on probability *density*, guaranteeing that  $P(Y(X_1) < Y < Y(X_2)) = P(X_1 < X < X_2)$
- *Probabilities* are invariant under change of variable
- The mode of the probability density is not invariant, so any “maximum probability density” criterion is broken
- Likelihood *ratio* is invariant under change of variable (Jacobian in denominator and numerator cancel out)



## Likelihood function is awesome - 2



- Let's now change the parameter  $\theta$  to  $u(\theta)$
- The likelihood function is invariant!!!  $L(\theta) = L(u(\theta))$
- The likelihood function being invariant under reparametrization of  $\theta$  reinforces the fact that the likelihood function is NOT a p.d.f. in  $\theta$
- Criteria of maximization of the likelihood survive perfectly to reparametrization
  - Actually, computationally it is easier to minimize  $-\log(L)$
- It is practically always possible to work with parabolic likelihoods

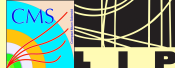


- Suppose we want to decide between two hypotheses
  - $H^0$  (for example: only Standard Model exists, no Higgs)
  - $H^1$  (for example: there is a Higgs)
- **We want to test the null hypothesis  $H^0$  against the alternative hypothesis  $H^1$** 
  - **We are not testing  $H^1$**
- Let  $\mathbf{X}$  be a function of the observations (called “*test statistic*”)
- Let  $W$  be the space of all possible values of  $\mathbf{X}$

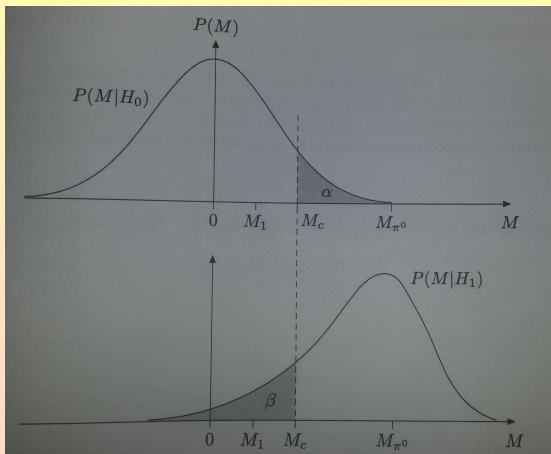
- Divide  $W$  into
  - A critical region  $w$ : observations  $X$  falling into  $w$  are regarded as suggesting that  $H^0$  is NOT true
  - A region of acceptance  $W - w$
- The size of the critical region is adjusted to obtain a desired *level of significance*  $\alpha$ 
  - Also called *size of the test*
  - $P(X \in w | H^0) = \alpha$
  - $\alpha$  is the probability of rejecting  $H^0$  when  $H^0$  is actually true

- The usefulness of the test depends on how well it discriminates against the alternative hypothesis
- The measure of usefulness is the *power of the test*
  - $P(X \in w|H^1) = 1 - \beta$
  - Power ( $1 - \beta$ ) is the probability of X falling into the critical region if  $H^1$  is true
  - $P(X \in W - w|H^1) = \beta$
  - $\beta$  is the probability that X will fall into the acceptance region if  $H^1$  is true
- NOTE: some authors use  $\beta$  where we use  $1 - \beta$ . Pay attention, and live with it.

# Let's decide something! - 4



- $H^0$ :  $pp \rightarrow pp$  elastic scattering
- $H^1$ :  $pp \rightarrow pp\pi^0$
- Compute the missing mass  $M$  (as total rest energy of unseen particles)
- Under  $H^0$ ,  $M = 0$
- Under  $H^1$ ,  $M = 135 \text{ MeV}$



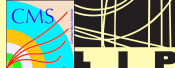
	Choose $H^0$	Choose $H^1$
$H^0$ is true	$1 - \alpha$	$\alpha$ (Type I error)
$H^1$ is true	$\beta$ (Type II error)	$1 - \beta$



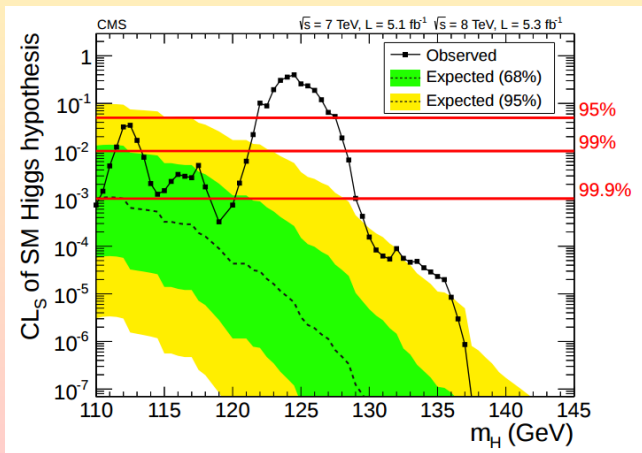
- Take the likelihood function  $L(\theta) = \prod_{i=1}^N f(X_i|\theta)$
- In general, the parameter space  $\theta$  of the parameters  $\theta$  can be partitioned such as
  - $H^0: \theta \in \nu$
  - $H^1: \theta \in \theta - \nu$
- A good test statistic for hypothesis testing is then the *maximum likelihood ratio*
- $\lambda = \frac{\max_{\theta \in \nu} L(\theta)}{\max_{\theta \in \theta} L(\theta)}$
- It usually produces the most powerful test.

- Systematic uncertainties can be (are usually) parametrized into the likelihood function
- Separate the parameters into
  - the *signal strength*  $\mu$  (often representing  $\sigma/\sigma_{SM}$ )
  - the parameters representing uncertainties, *nuisance parameters*  $\theta$
- $H^0$ :  $\mu = 0$  (Standard Model only, no Higgs)
- $H^1$ :  $\mu = 1$  (Standard Model + Standard Model Higgs)
- Find the maximum likelihood estimates (MLEs)  $\hat{\mu}$ ,  $\hat{\theta}$
- Find the conditional MLE  $\hat{\theta}(\mu)$ , i.e. the value of  $\theta$  maximizing the likelihood function for each fixed value of  $\mu$
- Write the test statistics as  $\lambda(\mu) = \frac{L(\mu, \hat{\theta}(\mu))}{L(\hat{\mu}, \hat{\theta})}$
- This beast depends on the signal strength  $\mu$ , but NOT on the nuisance parameters
- Nuisance parameters have been “profiled”, i.e. their MLE has been taken as a function of each value of  $\mu$

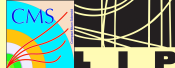
## A practical example: Higgs discovery - 1



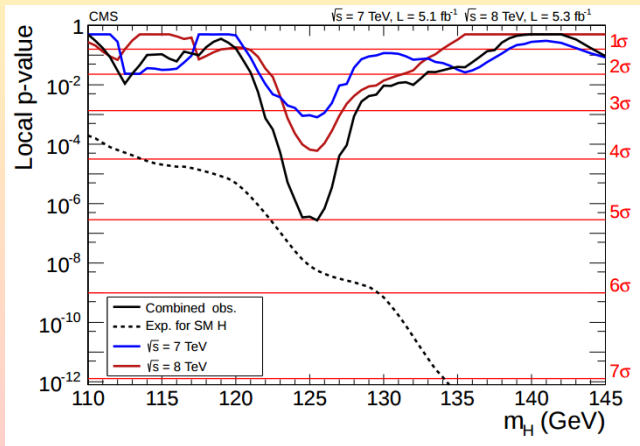
- Apply a full procedures to compute correctly all the needed quantities ( $CL_s$  method)
- End up with plotting the signal strength for different Higgs mass hypotheses
- What can we say about this plot?



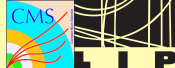
## A practical example: Higgs discovery - 2



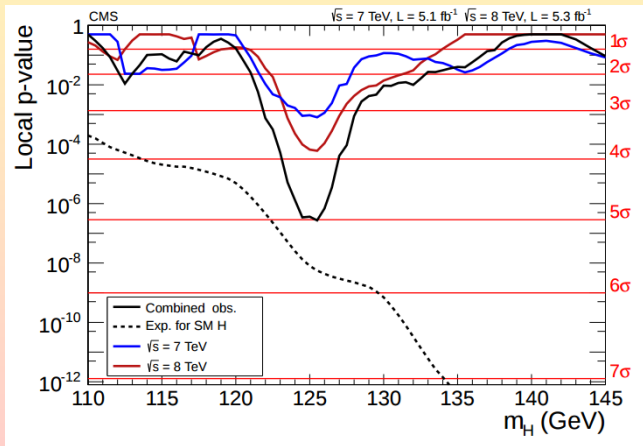
- Bonus: what does the significance of the discovery (“the sigmas”) represent?



## A practical example: Higgs discovery - 2

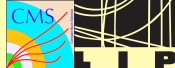


- Bonus: what does the significance of the discovery (“the sigmas”) represent?
- how likely is that  $H^0$  (SM-only) is true and the observed excess rises as consequence of a random fluctuation of the background?



- The original plan included much more material
  - Mainly practical case studies and hints for goodness of fit tests
- It would have been too much for a single lesson
  - If you are interested, I can add the material to the slides, for your offline fun
  - I can put online the code used to generate some of the plot examples in this presentation, if you are interested
- Feedback is welcome! These are my first lectures ever, and compressing in one hour “statistics” is almost impossible
  - Both for the giver and the receiver 😊
- In the next slide you can find some references

## Non-exhaustive list of references



- Frederick James: Statistical Methods in Experimental Physics - 2nd Edition, World Scientific
- Glen Cowan: Statistical Data Analysis - Oxford Science Publications
- Louis Lyons: Statistics for Nuclear And Particle Physicists - Cambridge University Press
- Louis Lyons: A Practical Guide to Data Analysis for Physical Science Students - Cambridge University Press
- Annis?, Stuard, Ord, Arnold: Kendall's Advanced Theory Of Statistics I and II
- R.J.Barlow: A Guide to the Use of Statistical Methods in the Physical Sciences - Wiley
- Kyle Cranmer: Lessons at HCP Summer School 2015
- Kyle Cranmer: Practical Statistics for the LHC - <http://arxiv.org/abs/1503.07622>
- Harrison Prosper: Practical Statistics for LHC Physicists - CERN Academic Training Lectures, 2015 <https://indico.cern.ch/category/72/>

THANKS FOR THE  
ATTENTION!



# Backup