Phenomenology point of view of data analysis: statistics

O. L. G. Peres[1]

[1] Instituto de Fisica Gleb Wataghin
UNICAMP

26 September 2012

Summary

1. Introduction, Probability distribution, Poisson distribution,
2. The Gaussian Limit: The central limit theorem, Gaussian errors, Multi-dimensional gaussian erros, Error Matrix
3. Fitting and Hypothesis Testing

Summary

1. Introduction, Probability distribution, Poisson distribution,
2. The Gaussian Limit: The central limit theorem, Gaussian errors, Multi-dimensional gaussian erros, Error Matrix
3. Fitting and Hypothesis Testing

Summary

1. Introduction, Probability distribution, Poisson distribution,
2. The Gaussian Limit: The central limit theorem, Gaussian errors, Multi-dimensional gaussian erros, Error Matrix
3. Fitting and Hypothesis Testing

Introduction

Inferential statistics provides mathematical methods to infer the properties of a population from a randomly selected sample taken from it. A population is an arbitrary collection of elements, a sample just a subset of it [1]

Scientific measurements are subject to the same scheme. Let us look to few statistical problems.

1. A certain experiment detect neutrinos from reactors. It observe a distortion from expected from theory. It is possible to describe the distortion assuming oscillations?
2. A certain experiment is trying to look for angular distribution of events, and determine if is compatible with the expected from the theory or not.
3. 4. A distortion is observed in the spectrum of beta decay. Is it a background fluctuation or the signal for neutrino mass ?

[1] G. Bohm and G. zech, Introduction to Statistics and Data Analysis for Physicist, DOI 10.3204/DESY-BOOK/statistics (e-book) <http://www-library.desy.de/elbook.html>

Introduction

Scientific measurements are subject to the same scheme. Let us look to few statistical problems.

1. A certain experiment detect neutrinos from reactors. It observe a distortion from expected from theory. It is possible to describe the distortion assuming oscillations?
2. A certain experiment is trying to look for angular distribution of events, and determine if is compatible with the expected from the theory or not.
3. 4. A distortion is observed in the spectrum of beta decay. Is it a background fluctuation or the signal for neutrino mass ?

Introduction

Scientific measurements are subject to the same scheme. Let us look to few statistical problems.

1. A certain experiment detect neutrinos from reactors. It observe a distortion from expected from theory. It is possible to describe the distortion assuming oscillations?
2. A certain experiment is trying to look for angular distribution of events, and determine if is compatible with the expected from the theory or not.
3. 4. A distortion is observed in the spectrum of beta decay. Is it a background fluctuation or the signal for neutrino mass ?

Phenomenology and statistics

Experimental science/phenomenology concerned with two types of experimental measurement:

1. Measurement of a quantity : parameter estimation
2. Tests of a theory/model : hypothesis testing

Phenomenology and statistics

Experimental science/phenomenology concerned with two types of experimental measurement:

1. Measurement of a quantity : parameter estimation
2. Tests of a theory/model : hypothesis testing

For parameter estimation we usually have some data (a set of measurements) and from which we want to obtain

1. The best estimate of the true parameter; *the measured value*
2. The best estimate of how well we have measured the parameter; *the uncertainty*

Phenomenology and statistics

For parameter estimation we usually have some data (a set of measurements) and from which we want to obtain

1. The best estimate of the true parameter; *the measured value*
2. The best estimate of how well we have measured the parameter; *the uncertainty*

Phenomenology and statistics

For hypothesis testing we usually have some data (a set of measurements) and one or more theoretical models, and want

1. A measure of how consistent our data are with the model; *a probability*
2. Which model best describes our data; *a relative probability*

To address the above questions we need to use and understand statistical techniques

Phenomenology and statistics

For hypothesis testing we usually have some data (a set of measurements) and one or more theoretical models, and want

1. A measure of how consistent our data are with the model; *a probability*
2. Which model best describes our data; *a relative probability*

To address the above questions we need to use and understand statistical techniques

Probabilities: How to Define Probability?

In Statistics: *probability, a basic concept which may be taken as undefinable, expressing in some way a degree of belief, or as the limiting frequency in an infinite random series. Both approaches have their difficulties and the most convenient axiomatization of probability theory is a matter of personal taste. Fortunately both lead to much the same calculus of probability.*

Probabilities: How to Define Probability?

In the frequentist statistics[1], sometimes also called classical statistics, the probability of an event, the possible outcome of an experiment, is defined as the frequency with which it occurs in the limit of an infinite number of repetitions of the exper iment. If in throwing dice the result five occurs with frequency 1/6 in an infinite number of trials, the probability to obtain five is defined to be 1/6. Examples

1. Poisson distribution
$$P(n,\mu) = e^{-\mu}\frac{\mu^n}{n!}$$
2. Gaussian distribution
$$P(x,\mu) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

[1] Disclaimer: In this talk I will not mention Bayesian statistics

Mean and Variance

$$\text{Mean : } \mu =< x >= \int xP(x)dx$$

$$\text{Mean of Squares } < x^2 >= \int x^2 P(x)dx$$

$$\text{Variance Var(x)} = \sigma^2 \equiv< (x-\mu)^2 >= \int (x-\mu)^2 P(x)dx$$

μ and σ describe the mean and the *width* of probability density function (PDF).

Central Limit Theorem

It can be proved that for large μ that a Poisson distribution tends to a Gaussian.This is one example of a more general theorem, the Central Limit Theorem:

If n random variables,x_i, each distributed according to any PDF, are combined then the sum $y = \sum_i x_i$ will have a PDF which , for large n, tends to a Gaussian.

For now we are going to use Gaussian distribution

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

where the width (variance) is given by Var(x)=σ.

Gaussian distribution

It is natural to introduce $\chi^2(x)$

$$\chi^2 = \frac{(x-\mu)^2}{\sigma^2} \qquad P(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\chi^2/2}$$

Fraction of events

$$\begin{array}{rll} 68.3\% & |x-\mu| < 1\sigma & \chi^2 < 1 \\ 95.5\% & |x-\mu| < 2\sigma & \chi^2 < 4 \\ 99.7\% & |x-\mu| < 3\sigma & \chi^2 < 9 \\ 6\times 10^{-6} & (x-\mu) > 5\sigma & \chi^2 > 25 \end{array} \tag{1}$$

Gaussian distribution

Fraction of events

$$\begin{array}{rll} 68.3\% & |x-\mu| < 1\sigma & \chi^2 < 1 \\ 95.5\% & |x-\mu| < 2\sigma & \chi^2 < 4 \\ 99.7\% & |x-\mu| < 3\sigma & \chi^2 < 9 \\ 6 \times 10^{-6} & (x-\mu) > 5\sigma & \chi^2 > 25 \end{array} \tag{1}$$

The region defined by $|x-\mu| < 1\sigma$, or 1σ region is called a confidence interval for a given confidence level (C.L.)

2-D dimensional Gaussian distribution I

If we assume two independent measurements x and y (ignoring correlations)

$$P(x,y) = P(x)P(y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-(x-\overline{x})^2)/(2\sigma_x^2)} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-(y-\overline{y})^2/(2\sigma_y^2)}$$
$$= \frac{1}{2\pi\sigma_x\sigma_y} e^{-(1/2)\left[(x-\overline{x})^2)/(\sigma_x^2)+(y-\overline{y})^2/(\sigma_y^2)\right]}$$

that describe two measurements with $x \pm \sigma_x$ and $y \pm \sigma_y$. We can rewrite as a χ^2,

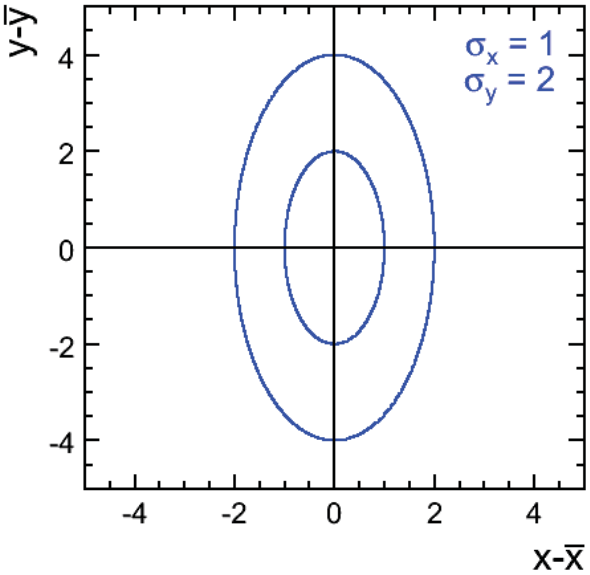
$$P(x,y) = \frac{1}{\sqrt{2\pi}\sigma_x\sigma_y} e^{-\chi^2(x,y)/2} \tag{2}$$

where

$$\chi^2(x,y) = \frac{(x-\overline{x})^2}{\sigma_x^2} + \frac{(y-\overline{y})^2}{\sigma_y^2} \tag{3}$$

2D- dimensional Gaussian distribution II

2D dimensional curve for Gaussian distribution



Inner (Outer) curve is for 1σ (2σ) :

We ask the question what is the allowed region for the parameter y or what is the allowed region for parameter x?

The answer is very easy:

68.3% events inside $\pm 1\sigma_x$ independently of y

68.3% events inside $\pm 1\sigma_y$ independently of x

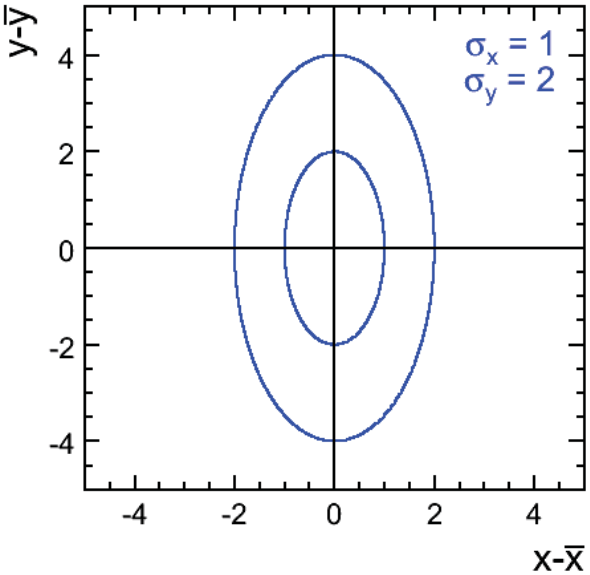
2D- dimensional Gaussian distribution II

Another question that we can answer is what is 68 % joint probability for the parameters x and y ?

Consider the contours of $\chi^2 = \frac{(x- < x >)^2}{\sigma_x^2} + \frac{(y- < y >)^2}{\sigma_y^2}$.

Then $\chi^2 = 1$ correspond to contour of PDF that fails to have $e^{-1/2}$ of the peak (remember prob=$\sim e^{-\chi^2/2}$)

We should take into account that now we have a multidimensional surface:



Types of errors in a measurement

- Statistics errors:Typically $\sigma^{stat} = \sqrt{N}$

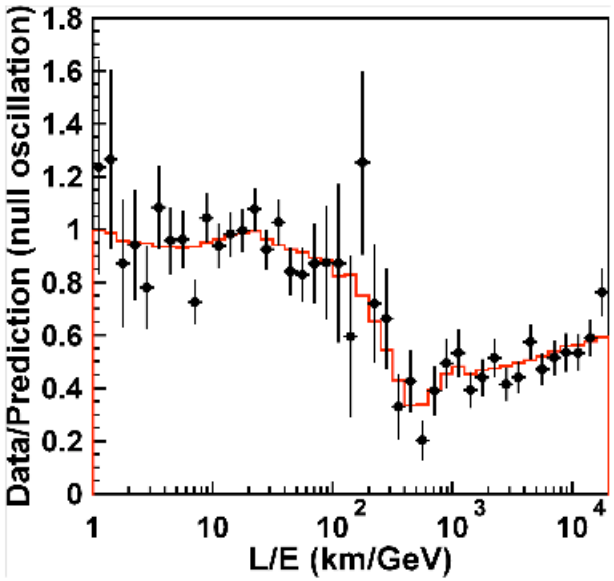
how many electrons were detected at fixed time
tossing a coin

- Systematic errors:

energy calibration
imperfect theory prediction

- Blunders errors

bugs in the analysis/errors in Monte Carlo code/forgot to include a particular background



Statistical errors shown only.

Suppose that x and y have two source of errors[1]: statistical :(s_x and s_y with no correlations and systematics (c_x and c_y with full correlation.
This mean that $x = x_0 \pm s_x \pm c_x$ and $y = y_0 \pm s_y \pm c_y$.
Then the error matrices sum up

$$\sigma^2 = \begin{pmatrix} s_x^2 & 0 \\ 0 & s_y^2 \end{pmatrix} + \begin{pmatrix} c_x^2 & c_x c_y \\ c_x c_y & c_y^2 \end{pmatrix} \equiv \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

where we define $\sigma_x^2 = s_x^2 + c_x^2$, $\sigma_y^2 = s_y^2 + c_y^2$, $\rho = \dfrac{c_x c_y}{s_x s_y}$ and the total error is correlated.

[1] E. Lisi, Neutrino physics tutorials

As an example of the general case, we have statistical and systematic errors, in the atmospheric neutrino, all points have independent statistical errors, $(\sigma^2_{stat})_{ij} = \delta_{ij}\sigma^{exp}_i\sigma^{exp}_j$ but due the theoretical neutrino flux have the same source, then the predictions are correlated. We can write $(\sigma^2_{syste})_{ij} = \rho^{theo}_{ij}\sigma^{theo}_i\sigma^{theo}_j$, where ρ_{ij} should be found or given by some experimentalist. The total error of each point is described by
$\sigma^2_{ij} = (\sigma^2_{stat})_{ij} + (\sigma^2_{syste})_{ij} = \delta_{ij}\sigma^{exp}_i\sigma^{exp}_j + \rho^{theo}_{ij}\sigma^{theo}_i\sigma^{theo}_j$: total errors are correlated.

The general 2D dimensional distribution

In general we can have a correlation, ρ between the errors of variable x and y,

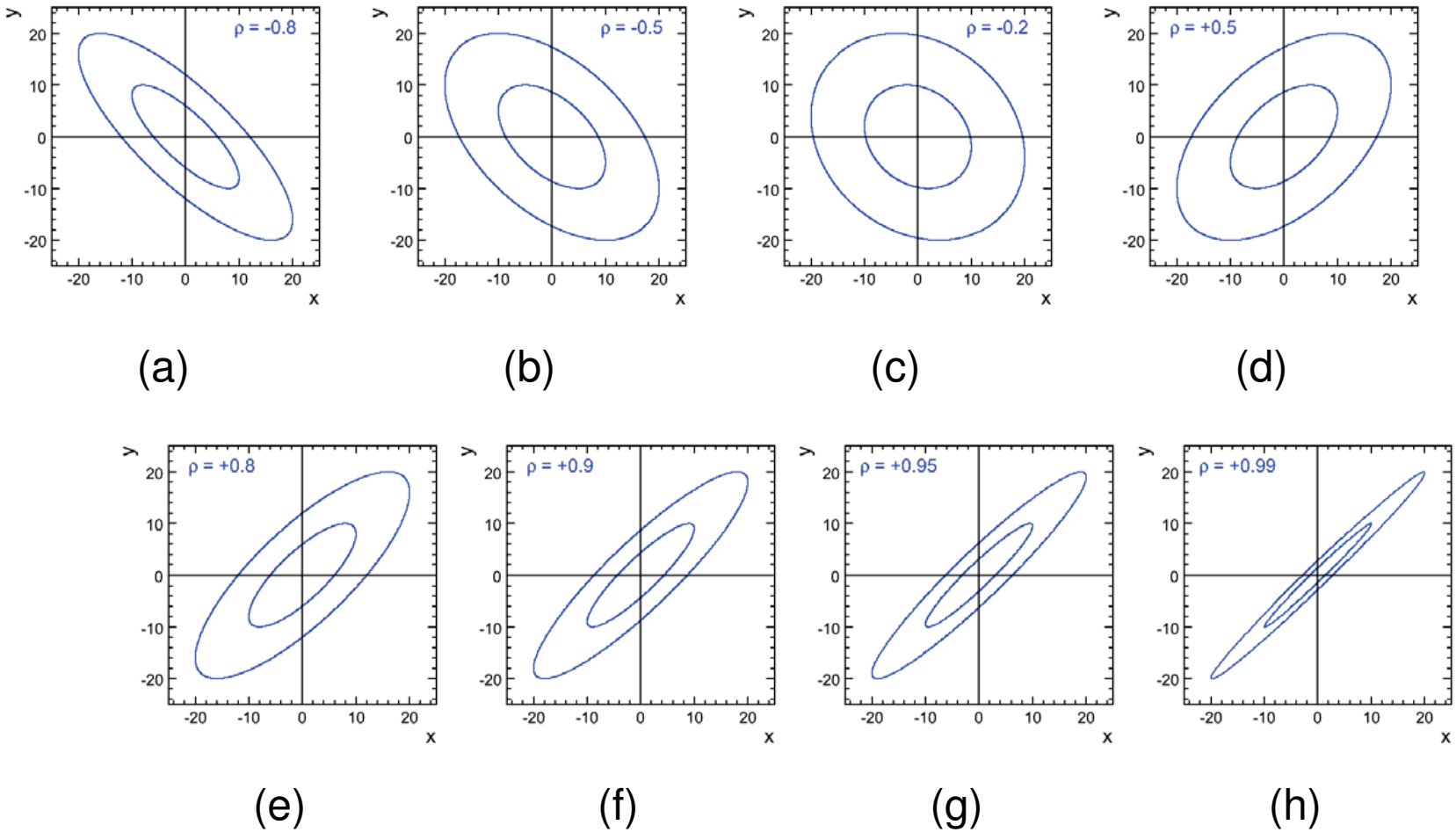


Figure : From negative correlation to positive correlation,

The general 2D dimensional distribution

$$P(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_y} e^{-\chi^2/2}$$

$$\chi^2 = -\frac{1}{(1-\rho^2)}\left[\frac{(x-\overline{x})^2}{\sigma_x^2} + \frac{(y-\overline{y})^2}{\sigma_y^2} - \frac{2\rho(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y}\right]$$

If we define the error Matrix,

$$\mathbf{M} = \begin{pmatrix} < x^2 > & < xy > \\ < xy > & < y^2 > \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

then we can define the probability as , where (we need to compute the N-dimensional inverse matrix $\mathbf{M^{-1}}$).

$$P(x,y) = \frac{1}{2\pi|\mathbf{M}|} e^{-(\mathbf{x^T M^{-1} x})/2}$$

For general multi-dimensional

$$P(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{n/2}|\mathbf{M}|^{1/2}} e^{-(\mathbf{x^T M^{-1} x})/2}$$

The χ^2 analysis

Given some data (event counts, distributions) and a particular theoretical model

are the data consistent with the model:

- hypothesis testing
- goodness of fit

in the context of the model, what are our best estimates of its parameters:

- fitting

In both cases, need a measure of consistency of data with our model. Start with a discussion of χ^2

The χ^2 analysis

- Suppose we measure a parameter, $x \pm \sigma$, which a theorist says should have the value μ.

Within this simple model, we can write down the prior probability of obtaining the value $x \pm \sigma$, given the prediction

$$P(\text{data,prediction}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- To express the consistency of the data,

ask the question if the model is correct what is the probability of obtaining result at least far as far from the prediction as the observed value.This is simply the fraction of the area under the Gaussian with $|x - \mu| > |x_{obs} - \mu|$

- The degree of consistence is

$$P(\chi^2 > \chi^2_{obs}) \qquad \text{where} \chi^2 = \left(\frac{x-\mu}{\sigma}\right)^2$$

The χ^2 analysis

For n dimensional probability

$$P(\chi^2, n) \propto \left(\chi^2\right)^{(n-2)/2} e^{-\chi^2/2}$$

And for any number of variables we have

$$P(\chi^2 > \chi^2_{obs}) = \int_{(\chi^2)_{obs}}^{\infty} P(\chi^2, n) d\chi^2$$

Notice the dependence on n, as we see before.

The χ^2 analysis

Recipe for fitting data with a model

Build $\chi^2 = \mathbf{X}^T(\sigma^2)^{-1}\mathbf{X}$,

$$\mathbf{X} = \begin{pmatrix} x_1^{theo}(\vec{p}) - x_1^{exp} \\ x_2^{theo}(\vec{p}) - x_2^{exp} \\ \cdots\cdots \\ x_N^{theo}(\vec{p}) - x_N^{exp} \end{pmatrix}$$

where $\vec{p}$ is a parameter space of the model (dimension $N_p \neq N$). For example to fit KamLand data, we can use a two-generation oscillation mechanism, that have two parameters and KamLand have 20 points. Then N=20, and N_p=2.

The χ^2 analysis

Find

$$\chi^2_{min} = \min_{\vec{p}} \left(\chi^2(\vec{p})\right)$$

at $\vec{p} = \vec{p_0}$, where $\vec{p_0}$ is the point that minimize this function.

Check if

$\chi^2_{min} \sim N - N_p \pm \sqrt{2 * (N - N_p)}$,

where N-N_p is the degrees of freedom for test of hypothesis.

If the previous condition it is not satisfied the model is either wrong (χ^2_{min} too high) ou *too good and suspect* (χ^2_{min} too low).

If χ^2_{min} is reasonable, try to estimate parameters around best fit $\vec{p_0}$ (parameter estimation)

The χ^2 analysis

Parameter estimation

Suppose you want $\pm 1\sigma$ ranges for each parameter $p_1, p_2, p_3, ...$ independetenly of the others (*marginalizing the others*)

Then

- Build $\Delta\chi^2 = \chi^2(\vec{p}) - \chi^2_{min}$, N_p dimensional manifold
- Project $\Delta\chi^2$ onto axis p_i, get $p_i^0 - \sigma_p^- < p^0 < p_i^0 + \sigma_p^+$.

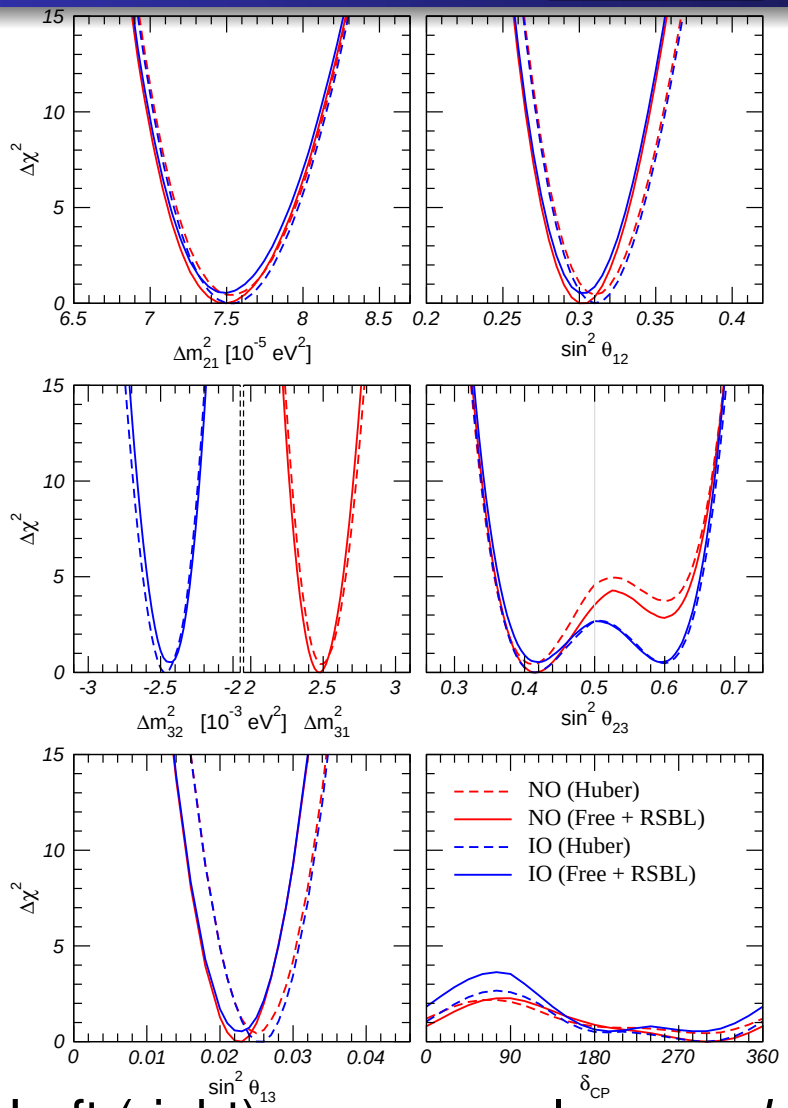
In practice the projection operation mean to impose that

$$\Delta\chi^2_i = \min_{p_j \neq p_i} (\chi^2(\vec{p}) - \chi^2_{min}) = 1$$

where we marginalize all the other variables

The χ^2 analysis

NuFIT 1.0 (2012)

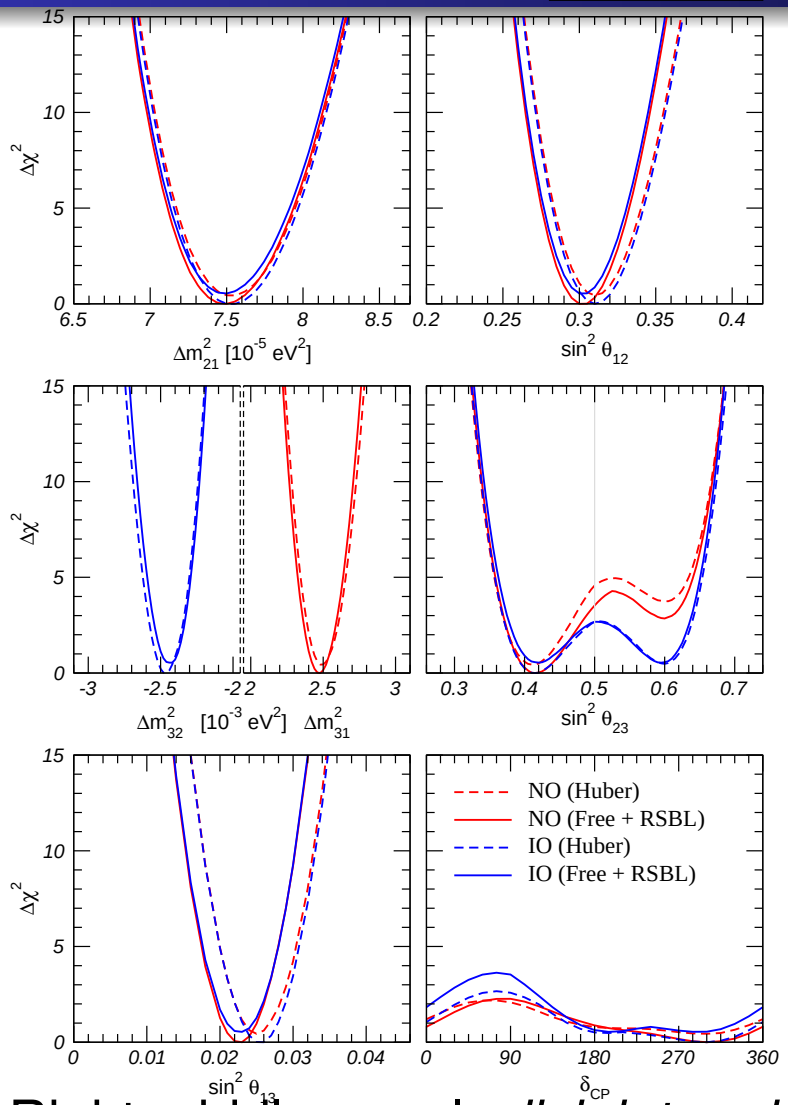


Left (right) upper panel: *normal* χ^2 behavior, projection on axis give the region for Δm^2_{21} $(\sin^2\theta_{12})$ [1]

[1] M. C. Gonzalez-Garcia et al, 1209.3023

The χ^2 analysis

NuFIT 1.0 (2012)



Right middle panel: *disjoint regions* for χ^2 , projection on axis give two region for $\sin^2\theta_{23}$ [1]

[1] M. C. Gonzalez-Garcia et al, 1209.3023

The χ^2 analysis

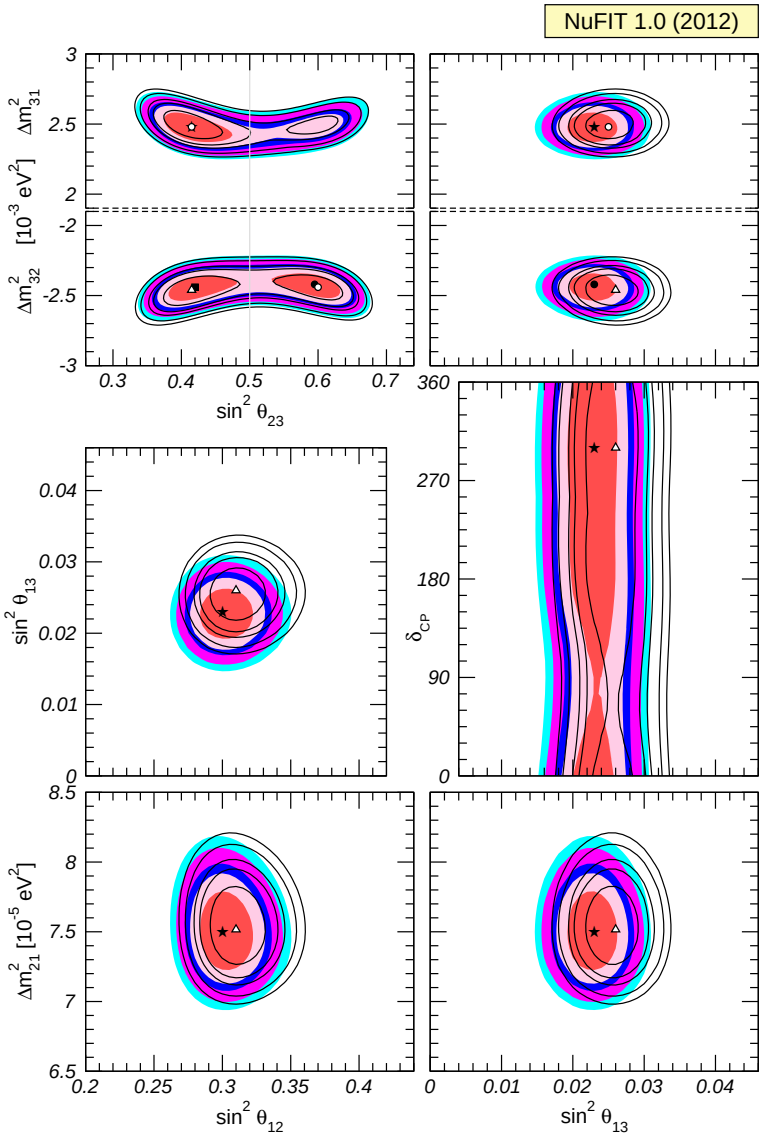
- We can justified this procedure ($\Delta\chi^2$ projections) as far the allowed manifold is a simply connected volume. For disconnected regions, there is not a definitive consensus!! Multiple domains of $\vec{p}$ are keep, waiting for some future experiments broke the degeneracy of solutions.

The joint probability of $\vec{p}$ sometimes is also interesting to quote, for example the allowed region on $\sin^2 2\theta$ and Δm^2 parameter space.

In this case the volume defined by $\Delta\chi^2$=constant, the constant change with the dimension of parameter space N_p.

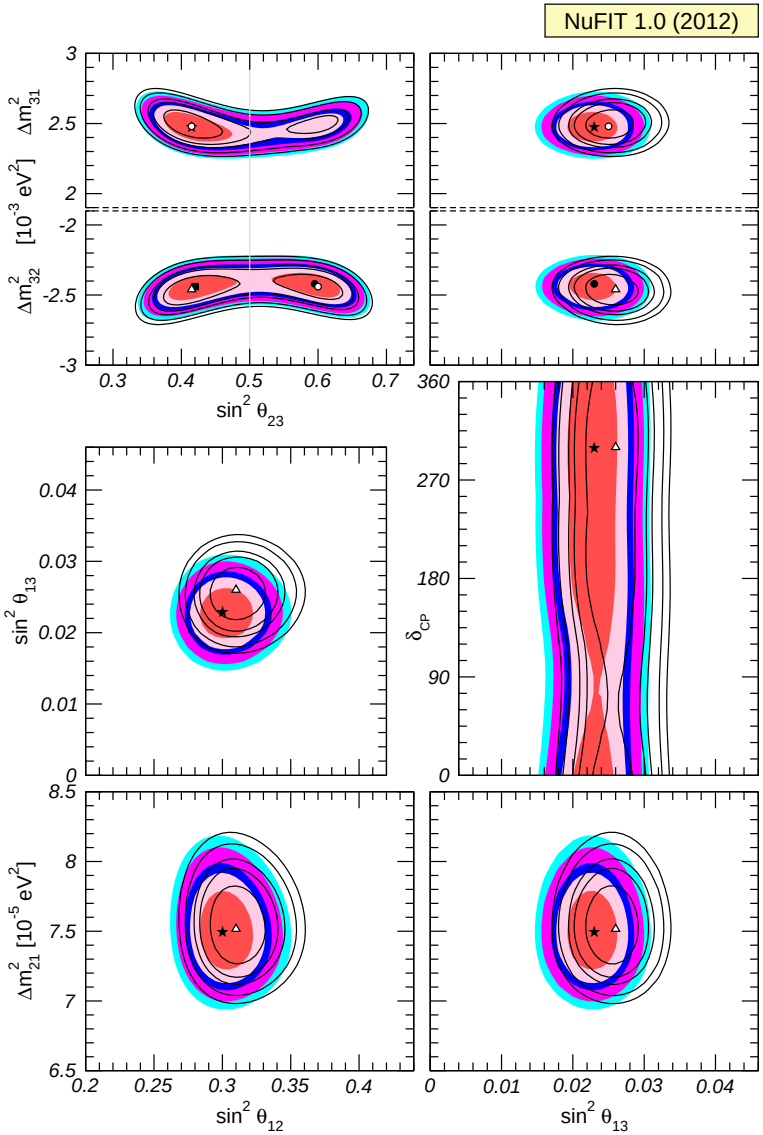
$C.L.(\%)$	$N_p = 1$	$N_p = 2$	$N_p = 3$
68.27	1.00	2.30	3.53
90	2.71	4.61	6.25
95	3.84	5.99	7.82
99	6.63	9.21	11.34
99.73	9.00	11.83	14.16

The χ^2 analysis



$C.L.(\%)$	$N_p = 2$
68.27	3.53
90	4.61
95	5.99
99	9.21

The χ^2 analysis



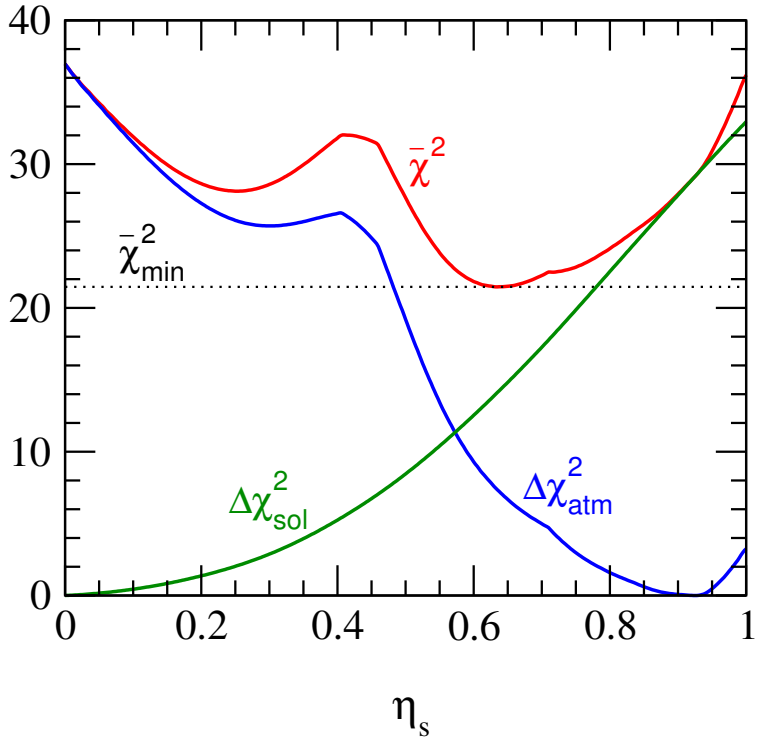
Regions for 1σ, 90%, 2σ, 99% and 3σ [1]

[1] M. C. Gonzalez-Garcia et al, 1209.3023

The χ^2 analysis

All cases showed so far are the good cases, all experiments combined show the same parameter region. But not always the case, sometimes we are faced with contradictory experiments!! What we should do?

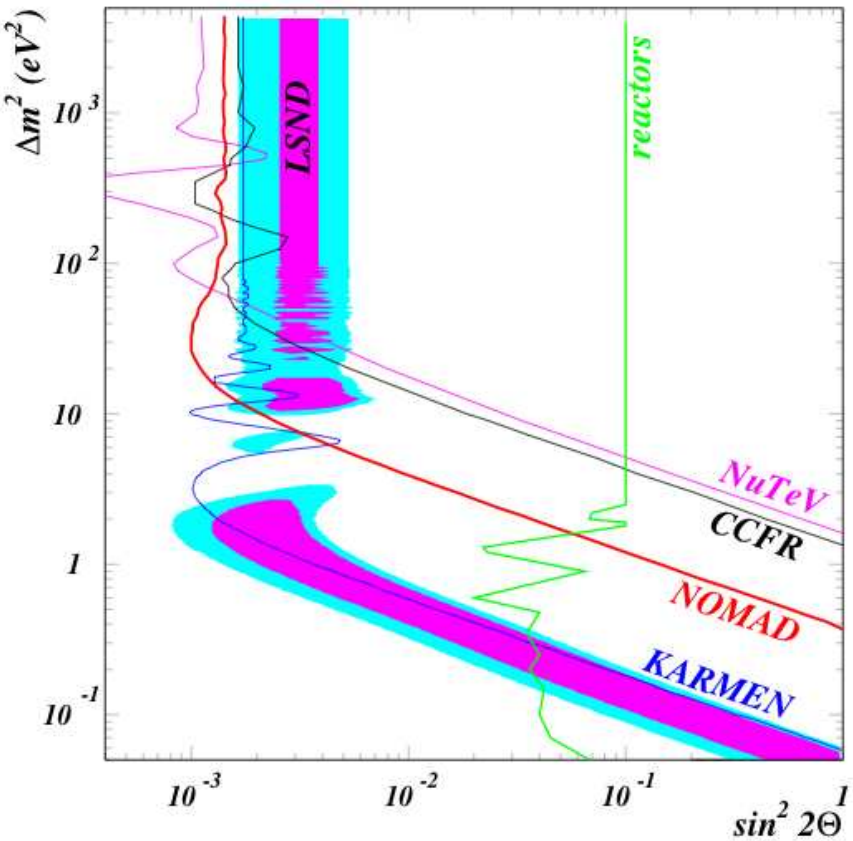
A χ^2 analysis[1] of solar and atmospheric experiments with one extra sterile neutrino, the so called 2+2 model can be parametrized as η_s the fraction of sterile neutrino in the oscillation.



[1] T. Schwetz and M. Maltoni, hep-ph/0304176v2

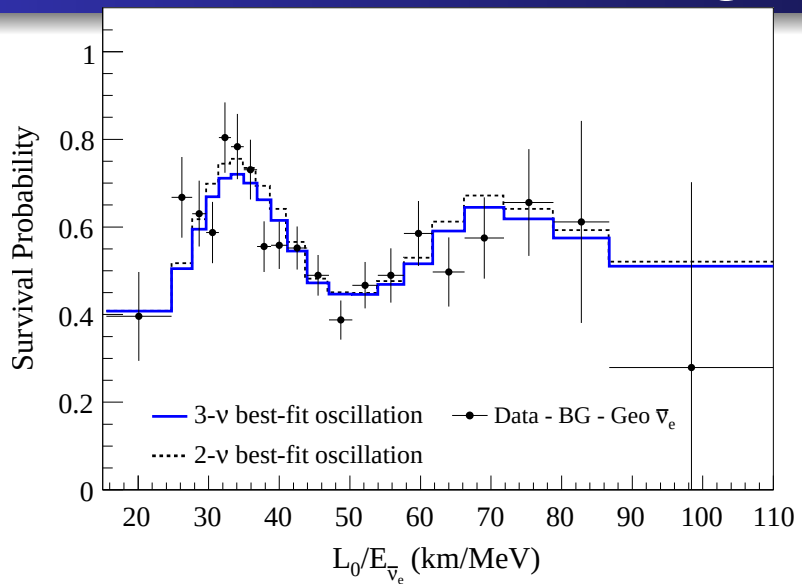
The χ^2 analysis

A intermediate case is that one experiment see a positive signal for neutrino oscillation and another dont see any signal of neutrino oscillation. BEWARE, we should combine the experiments and test if



we have a good solution.

Exercise: KamLand region



- Step 1: Define your χ^2, what are your input and theoretical prediction?
- Step 2: Fit curve with

$$P(\nu_\alpha \to \nu_\alpha) = 1 - \sin^2(2\theta)\sin^2\left(1.27\frac{\Delta m^2}{(\mathrm{eV})^2}\frac{\mathrm{L/Km}}{\mathrm{E/GeV}}\right)$$

- Assume Average distance 180 km. See Figure 2.3 in `http://kamland.lbl.gov/Dissertations/DetwilerJason-DoctorThesis.pdf`
- Find the allowed region for $\sin^2(2\theta)$ and Δm^2 parameters..
- There is a difference between your result and the numbers from literature?