Perhaps our standards are too high…

- Maybe I am giving you too much food for thought on this one issue
- The realization comes after reading a preprint by a >100-strong collaboration, MINOS
- In a recent paper [MINOS 2011] they derive 99.7% upper limits for some parameters, using antineutrino interactions. How then to combine with previous upper limits derived with neutrino interactions ?
- They use the following formula: $$\frac{1}{\mu_{up}^2} = \frac{1}{\mu_{up,1}^2} + \frac{1}{\mu_{up,2}^2}$$

The horror, the horror.

- You should all be able to realize that this cannot be correct! It looks like a poor man's way to combine two significances for Gaussian distributions, but it does not even work in that simple case.

Neyman vs MINOS

- To show how wrong one can be with the formula used by MINOS, take the Gaussian measurement of a parameter. Take **σ=1** for the Gaussian: this means, for instance, that if the unknown parameter is **μ=3**, there is then a 68% chance that your measurement **x** will be in the **2<x<4** interval.

  Since, however, what you know is your measurement and you want to draw conclusions on the unknown **μ**, you need to "invert the hypothesis". Let's say you measure x=2 and you want to know what is the maximum value possible for **μ**, at 95% confidence level. This requires producing a "Neyman construction". You will find that your limit is **μ<3.64**.

- So what if you got twice the measurement x=2, in independent measurements ? Could you then combine the limits as MINOS does ?
  If you combine two x=2 measurements, each yielding $\mu_1=\mu_2<3.64$ at 95%CL, according to the MINOS preprint you might do $\mu_{comb}=1/\mathrm{sqrt}(1/\mu_{1,up}^2 + 1/\mu_2^2) <$ **2.57**.
- Nice: you seem to have made great progress in your inference about the unknown μ. But unfortunately, the correct procedure is to first combine the measurement in a single pdf for the mean: this is $x_{ave}=2$, with standard deviation $\sigma=1/\mathrm{sqrt}(2)$. The limit at 95% CL is then **μ<3.16**, so quite a bit looser!

  Also note that we are being cavalier here: if x_1 and x_2 become different, the inference one can draw worsens considerably in the correct case, while in the MINOS method at most reduces to the most stringent of the two limits.

  For instance, $x_1=2$, $x_2=4$ yields the combined limit $\mu<4.16$, while MINOS would get $\mu<3.06$. This is consistent with the mistake of ignoring that the two confidence limits belong to the same confidence set.

The Jeffreys-Lindley Paradox

- One specific problem (among many!) which finds Bayesians and Frequentists in stark disagreement on the results: charge bias of a tracker at LEP
- Imagine you want to investigate whether your tracker has a bias in reconstructing positive versus negative curvature. We work with a zero-charge initial state (e^+e^-). You take a unbiased set of events, and count how many positive and negative curvature tracks you have reconstructed in a set of n=1,000,000. You get n^+=498,800, n^-=501,200. You want to test the hypothesis that R=0.5 with *a size* **α=0.05**.
- Bayesians will need a prior to make a statistical inference: their typical choice would be to assign equal probability to the chance that R=0.5 and to it being different (R<>0.5): a "point mass" of p=0.5 at R=0.5, and a uniform distribution of the remaining p in [0,1]
- The calculation goes as follows: we are in high-statistics regime and away from 0 or 1, so Gaussian approximation holds for the Binomial. The probability to observe a number of positive tracks as small as the one observed can then be written, with $x=n^+/n$, as $N(x,\sigma)$ with $\sigma^2=x(1-x)/n$. The posterior probability that R=0.5 is then

$$P(R=\frac{1}{2}\,|\,x,n) \approx \frac{\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}}{\left[\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}+\frac{1}{2}\int_0^1 \frac{e^{-\frac{(x-R)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\,dR\right]} = 0.97816$$

from which a Bayesian concludes that there is **no evidence against R=0.5,** and actually the data strongly supports the null hypothesis (P>1-α)

Jeffreys-Lindley: frequentist solution

- Frequentists will not need a prior, and just ask themselves how often a result "as extreme" as the one observed arises by chance, if the underlying distribution is N(R,σ) with R=1/2 and σ^2=x(1-x)/n as before.
- One then has

$$P(x \leq 0.4988 | R = \frac{1}{2}) = \int_0^{0.4988} \frac{e^{-\frac{(t-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi\sigma}} dt = 0.008197$$

$$\Rightarrow P'(x | R = \frac{1}{2}) = 2 * P = 0.01639$$

(we multiply by two since we would be just as surprised to observe an excess of positives as a deficit).

From this, frequentists conclude that the tracker is biased, since there is a less-than 2% probability, P'<α, that a result as the one observed could arise by chance! A frequentist thus draws the **opposite conclusion** that a Bayesian draws from the same data .

Likelihood ratio tests

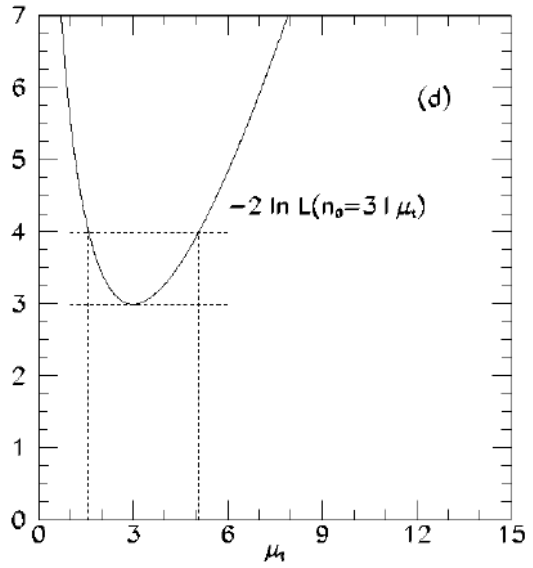
- Because of the invariance properties of the likelihood under reparametrization (L not a density!), a ratio of likelihood values can be used to find the most likely values of a parameter θ, given the data X
  - a reparametrization from θ to $f(\theta)$ will not modify our inference: if $[\theta_1, \theta_2]$ is the interval containing the most likely values of θ, $[f(\theta_1), f(\theta_2)]$ will contain the most likely values of $f(\theta)$!
  - log-likelihood differences also invariant
- One may find the interval by *selecting all the values of θ such that*

  $$-2\,[\ln L(\theta) - \ln L(\theta_{max})\,] <= Z^2$$

  The interval approaches asymptotically a central confidence interval with C.L. corresponding to ±Z Gaussian standard deviations. E.g. if we want 68% CL intervals, choose Z=1; for five-sigma, Z^2=25, etc.
- It is an approximation! Sometimes it undercovers (e.g. Poisson case)
- But a very good one in typical cases. The property depends on *Wilks' theorem* and is based on a few regularity conditions.
- LR tests are popular because it is what MINUIT MINOS gives Problems when θ approaches boundary of definition

Example: likelihood-ratio interval for Poisson process with n=3 observed: **$L(\mu) = \mu^3 e^{-\mu}/3!$** has a maximum at μ= 3. **$\Delta(2\ln L) = 1^2$** yields approximate ±1 Gaussian standard deviation interval : [1.58, 5.08]

For comparison: Bayesian central with flat prior yields [2.09,5.92]; NP central yields [1.37,5.92]



R. Cousins, Am. J. Phys. 63 398 (1995)

The likelihood principle

- As noted above, in both Bayesian methods and likelihood-ratio based methods, only the probability (density) for obtaining the data at hand is used: it is contained in the likelihood function. *Probabilities for obtaining other data are **not used***
- In contrast, in typical frequentist calculations (e.g., a p-value which is the probability of obtaining a value as extreme or *more extreme than that observed), one uses probabilities of data not seen.*

- This difference is captured by the *Likelihood Principle:*

  *If two experiments yield likelihood functions which are **proportional**, then Your inferences from the two experiments should be identical.*

- The likelihood Principle is built into Bayesian inference (except special cases). It is instead violated (sometimes badly) by p-values and confidence intervals.

- You cannot have both the likelihood principle fulfilled and guaranteed coverage.

- Although practical experience indicates that the Likelihood Principle may be too restrictive, it is useful to keep it in mind. Quoting Bob Cousins:

  "When frequentist results 'make no sense' or 'are unphysical', the underlying reason can be traced to a bad violation of the L.P."

Example of the Likelihood Principle

- Imagine you expect background events sampled from a Poisson mean b, assumed known precisely.

- For signal mean μ, the total number of events n is then sampled from Poisson mean μ+b. Thus,

  $P(n) = (\mu+b)^n e^{-(\mu+b)}/n!$

- Upon performing the experiment, you see no events at all, n=0. You then write the likelihood as

  $L(\mu) = (\mu+b)^0 e^{-(\mu+b)}/0! = \exp(-\mu)\exp(-b)$

- Note that changing b from 0 to any $b^*>0$, L(μ) only changes by the constant factor $\exp(-b^*)$. This gets renormalized away in any Bayesian calculation, and is *a fortiori* irrelevant for likelihood *ratios. So for zero events observed, likelihood-based inference about signal mean μ is independent of expected b.*

- You immediately see the difference with the Frequentist inference: in the confidence interval constructions, the fact that n=0 is less likely for b>0 than for b=0 results in *narrower confidence intervals for μ as b increases.*

Conditioning and ancillary statistics

- An "ancillary statistic" is a function of the data which carries information about the *precision* of the measurement of the parameter of interest, but no information about the parameter's value.
- Most typical case in HEP: branching fraction measurement. With N_A, N_B event counts in two channels one finds that

  $$\mathbf{P(N_A,N_B) = Poisson\,(N_A) \times Poisson\,(N_B) = Poisson\,(N_A+N_B) \times Binomial\,(N_A \mid N_A+N_B)}$$

  By using the expression on the right, one may ignore the ancillary statistics N_A+N_B, since all the information on the BR is in the conditional binomial factor → by restricting the sample space, the problem is simplified. This is relevant when one designs toy Monte Carlo experiments e.g. to evaluate uncertainties

- And it gets even more intriguing in the famous example by Cox (1958): flip a coin to decide whether to use a 10% scale (if you get tails) or a 1% scale (if you get head) to measure a weight. Which error do you quote for your measurement, upon getting a head ?
  - Of course the knowledge of your measuring device allows you to estimate that your precision is 1%
  - but a full NP construction (which seeks the highest power for a chosen α, unconditional on the outcomes) would require you to include the coin flipping in the procedure!

- *The quality of your inference depends on the breadth of the "whole space" you are considering*. The more you can restrict it, the better (i.e. the more relevant) your inference; but ancillary statistics are not easy to find
- The likelihood principle can be thought of as an extreme form of conditioning: you only consider the data you have !

Food for thought: relevant subsets

- Neyman's method for the Gaussian measurement with known sigma of parameter with unknown positive mean yields upper limits at 95% CL in the form $\mu_{UL}=x+1.64\sigma$
- This lends itself to a pointed criticism best highlighted by a hypothetical betting game
  - The procedure is guaranteed to cover the unknown true value in 95% of experiments by the math of Neyman's construction
  - Yet one can devise a betting strategy against it at 1/20 odds, using no more information than the observed x, and be guaranteed to win in the long run!
  - How ? *Just choose a real constant k: bet that the interval does not cover when x<k, pass otherwise*.
  - For k<-1.64 this wins EVERY bet! For larger k, advantage is smaller but is still >0.
- Surely then, the procedure is not making the best inference on the data ?
- Another example:

  Find μ using x_1, x_2 sampled from p(x|μ)=Uniform [μ-1/2, μ+1/2]:
  - A: {0.99,1.01} ; B: {0.51,1.49}
  - N-P procedures maximizing power in the unconditional space yield the same confidence interval for both data sets A and B; however, B clearly restricts the set of possible μ to [0.99,1.01] while A only restricts it to [0.51,1.49] !
  - **There exists in fact a ancillary statistics $|x_1-x_2|$** which carries no information on μ, yet can be used to divide the sample space in subsets where inference can be different.
  - See R. Cousins, Arxiv:1109.2023 for more discussion

Comparing methods to compute intervals

- **Bayesian credible intervals**:
  - need a prior (can be a good thing –allows a means to put in your personal prior belief)
  - random variable in construction is true value
  - usually obey the likelihood principle
  - can be basis for decision theory (provides $p(\theta|\text{data})$)
  - do not guarantee coverage

- **Frequentist confidence intervals**:
  - do not need a prior (can do inference reporting the result of your data keeping it objective)
  - random variables are extrema of intervals
  - do not obey the likelihood principle
  - guarantee coverage
  - use p(data not obtained) for inference about θ

- **Likelihood ratio intervals**:
  - do not need a prior
  - random variables are extrema of intervals
  - obey the likelihood principle

The three methods at work

- Let us take the classical example of a zero-background counting experiment, $N_{obs}=3$ case (as above): determine upper limit on signal. This boils down to three different recipes:

    1. **Bayesian upper limit** at 90% credibility: determine posterior $p(\mu|N)$;
       find μ_u such that posterior probability $P(\mu>\mu_u) = 0.1$.
    2. **Likelihood ratio** method for approximate 90% C.L. upper limit: find μ_u such that $L(\mu_u) / L(3)$ has prescribed value
    3. **Frequentist one-sided** 90% C.L. upper limit: find μ_u such that
       $P(n\leq 3 \,|\mu_u) = 0.1$.

- They give different answers ! That is because they ask different questions.

- Which method is best ? Not decidable – and certainly the answer cannot be given by HEP physicists !

- Several factors contribute to the practical choices made
  - Frequentist vs Bayesian preconceptions
  - Technical problems (eg. with the integration of the nuisance parameters in the Bayesian case → until MCMC tools became available, the problem was intractable in all but the easiest cases)
  - Peculiarities of the problem at hand. For instance, small statistics causes the Likelihood intervals, which rest on asymptotic properties of the form of L (Wilks' theorem) to have poor properties

Hypothesis testing: generalities

We are often concerned with proving or disproving a theory, or comparing and choosing between different hypotheses.

In general this is a different problem than that of estimating a parameter, but the two are tightly connected.

If nothing is known a priori about a parameter, naturally one uses the data to **estimate** it; if however a theoretical prediction exists on a particular value, the problem is more proficuously formulated as a **test of hypothesis**.

Within the idea of hypothesis testing one must also consider **goodness-of-fit tests**: in that case there is only one hypothesis to test (e.g. a particular value of a parameter as opposed to any other value), so some of the possible techniques are not applicable

A hypothesis is **simple** if it is completely specified; otherwise (e.g. if depending on the unknown value of a parameter) it is called **composite**.

Nuts and bolts of Hypothesis testing

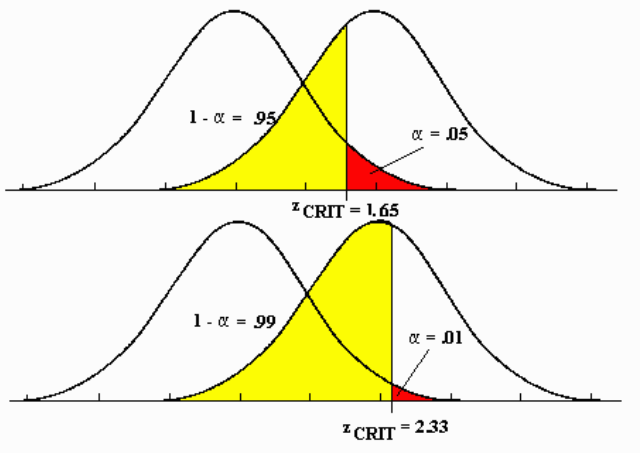
- H_0: null hypothesis
- H_1: alternate hypothesis
- Three main parameters in the game:
  - α: **type-I error rate**; probability that H_0 is true although you accept the alternative hypothesis
  - β: **type-II error rate**; probability that you fail to claim a discovery (accept H_0) when in fact H_1 is true
  - θ, parameter of interest (describes a continuous hypothesis, for which H_0 is a particular value). E.g. θ=0 might be a zero cross section for a new particle
- Common for H_0 to be nested in H_1

Can compare different methods by plotting α vs β vs the parameter of interest

- Usually there is a tradeoff between α and β; often a subjective decision, involving cost of the two different errors.
- Tests may be more powerful in specific regions of an interval (e.g. a Higgs mass)

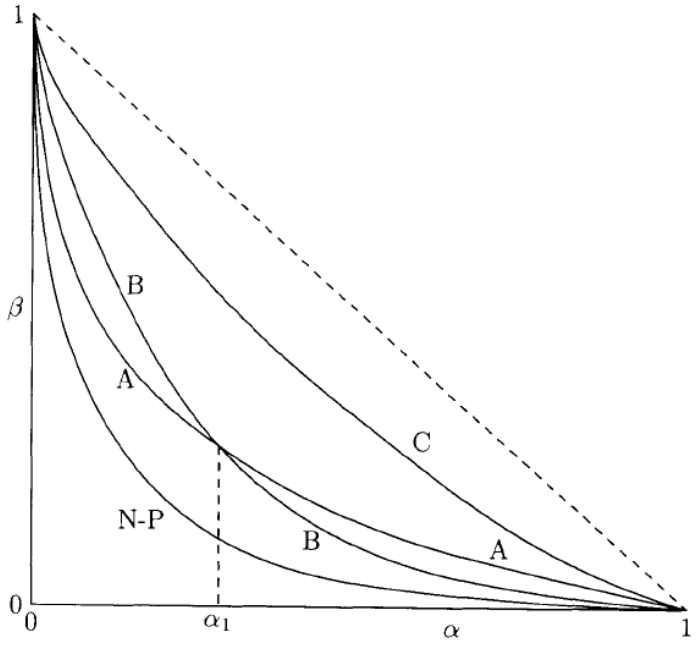
There is a 1-to-1 correspondence between hypothesis tests and interval construction

In classical hypothesis testing, test of σ=0 for the Higgs equates to asking whether σ=0 is in the confidence interval.



Above, a smaller α is paid with a larger type-II error rate (yellow area)
→ smaller power 1-β

Alpha vs Beta and power graphs



- Very general framework of classification
- Choice of α and β is conflicting: where to stay in the curve provided by your analysis method highly depends on habits in your field
- What makes a difference is the test statistics: note how the N-P likelihood-ratio test outperforms others in the figure [James 2006] – reason is N-P lemma
- As data size increases, power curve becomes closer to step function

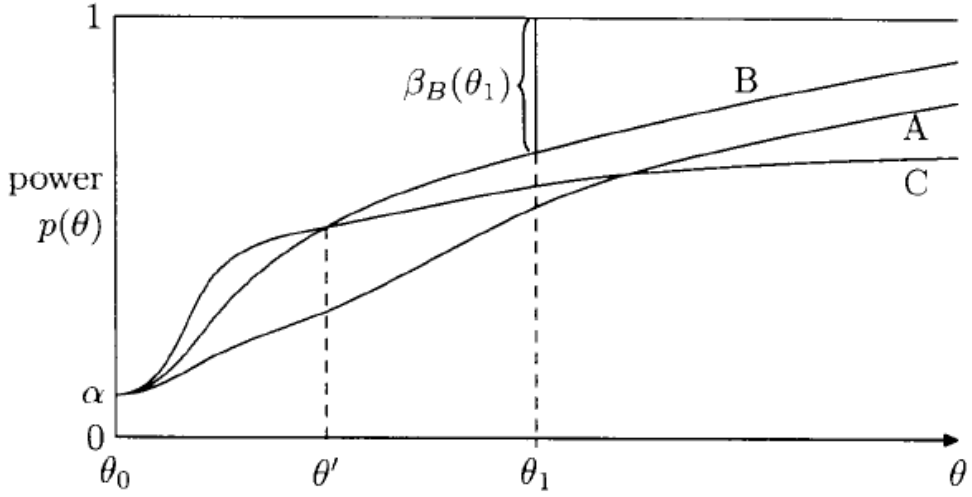
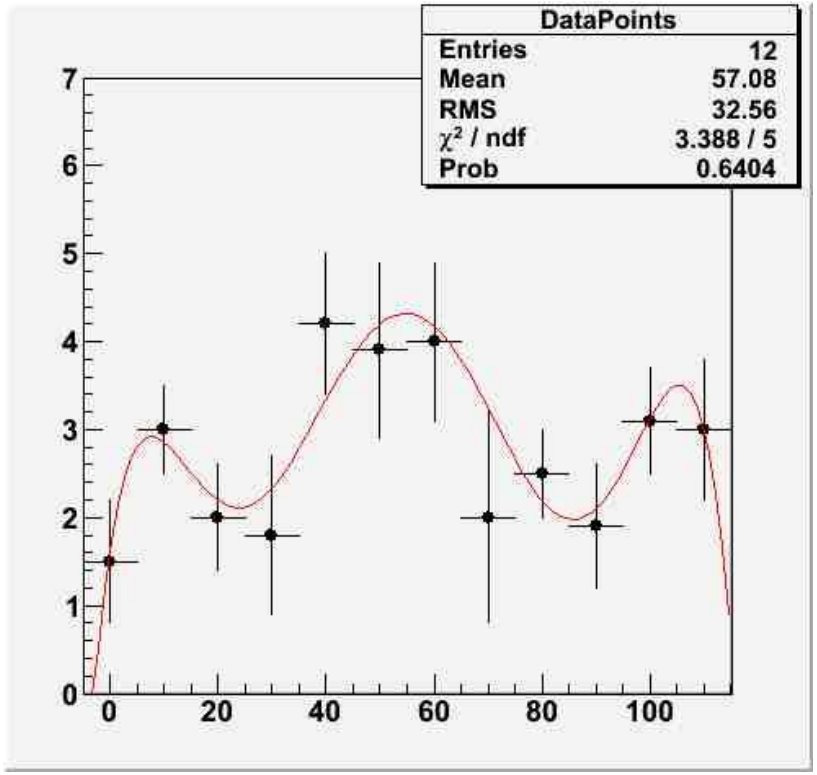


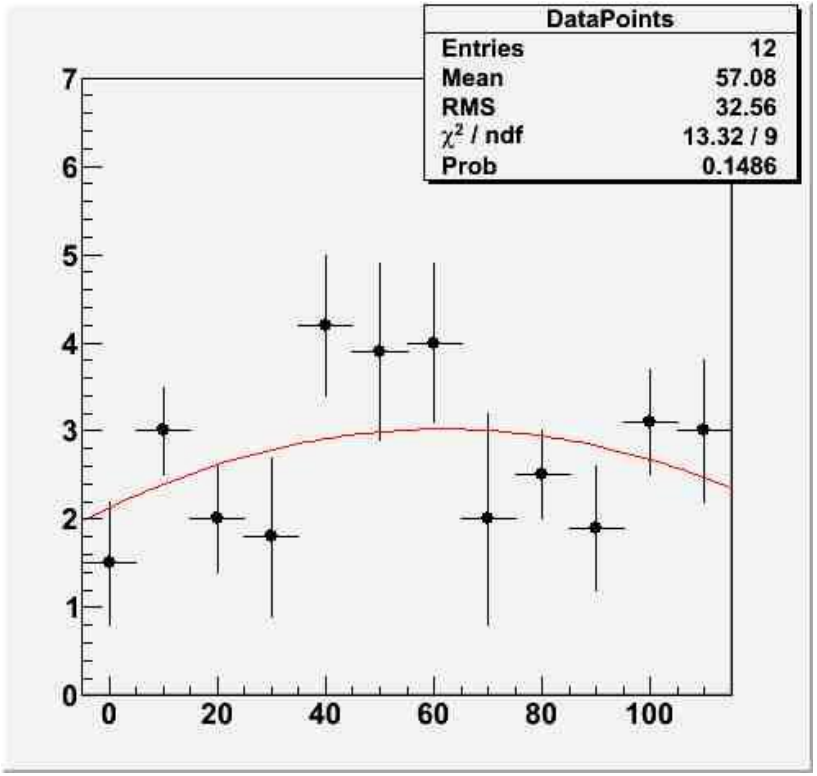
Fig. 10.3. Power functions of tests A, B, and C at significance level α. Of these three tests, B is the best for $\theta > \theta'$. For smaller values of θ, C is better.

The power of a test usually also depends on the parameter of interest: different methods may have better performance in different parameter space points UMP (uniformly most powerful): has the highest power for any θ

On overfitting

A complex problem in statistics is model selection. Upon getting some data, in the absence of a principled model of the pdf from which these were drawn, one needs to do some trial-and-error fitting. One danger is then to use models more complicated than would be needed by the data.
Which of the two functional forms do you think produced the data shown below ?





This leads us to a little but important side-topic: the F-test

Eye fitting strikes back: Fisher's F-test

- Suppose you have no clue of the real functional form followed by your data (**n** points)
  - or even suppose you know only its general form (e.g. polynomial, but do not know the degree)
- You may try a function $\mathbf{f_1(x;\{p_1\})}$ and find it produces a good fit (→goodness-of-fit); however, you are unsatisfied about some additional feature of the data that appear to be systematically missed by the model
- You may be tempted to try a more complex function –usually by adding one or more parameters to $\mathbf{f_1}$
  - **this ALWAYS improves the absolute χ^2**, as long as the new model "embeds" the old one (the latter means that given any choice of $\mathbf{\{p_1\}}$, there exists a set $\mathbf{\{p_2\}}$ such that $\mathbf{f_1(x;\{p_1\})==f_2(x;\{p_2\})}$
- How to decide whether $\mathbf{f_2}$ is more motivated than $\mathbf{f_1}$, or rather, that the added parameters are doing something of value to your model ?
- Don't use your eye! *Doing so may result in choosing more complicated functions than necessary to model your data, with the result that your statistical uncertainty (e.g. on an extrapolation or interpolation of the function) may abnormally shrink, at the expense of a modeling systematics which you have little hope to estimate correctly.*

→ Use the F-test: the function F

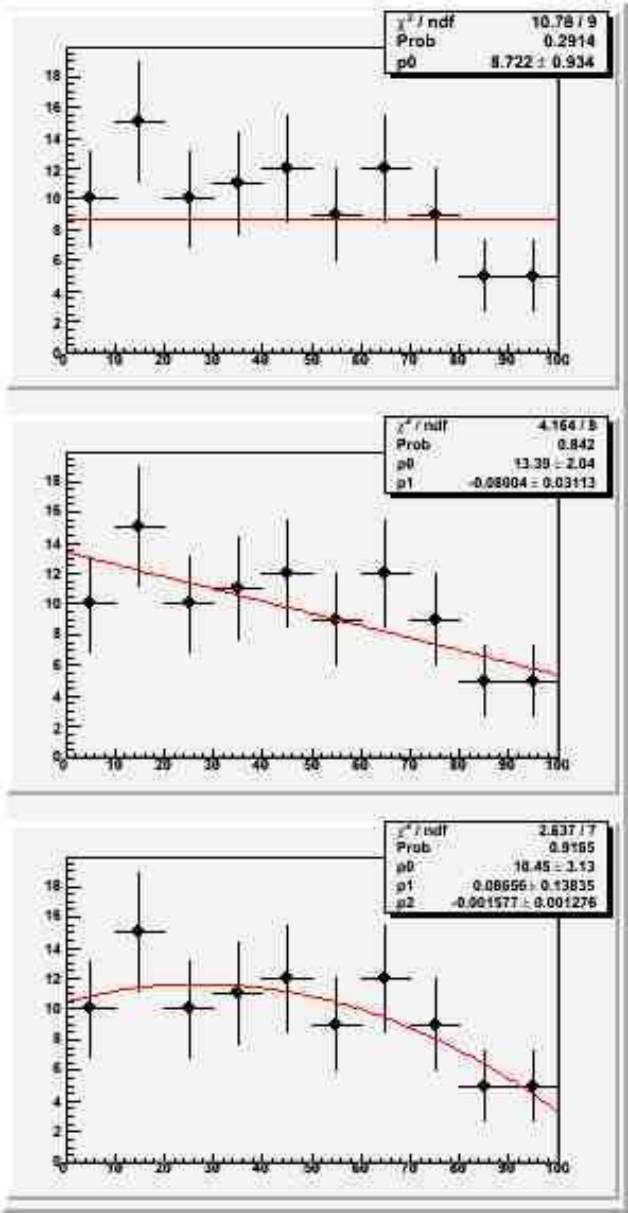
$$F = \frac{\dfrac{\sum_i (y_i - f_1(x_i))^2 - \sum_i (y_i - f_2(x_i))^2}{p_2 - p_1}}{\dfrac{\sum_i (y_i - f_2(x_i))^2}{n - p_2}}$$

has a Fisher distribution if the added parameter is **not** improving the model.

$$f(F;\nu_1,\nu_2) = \frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{F^{\frac{\nu_1}{2}-1}}{(\nu_1+\nu_2 F)^{\frac{\nu_1+\nu_2}{2}}}$$

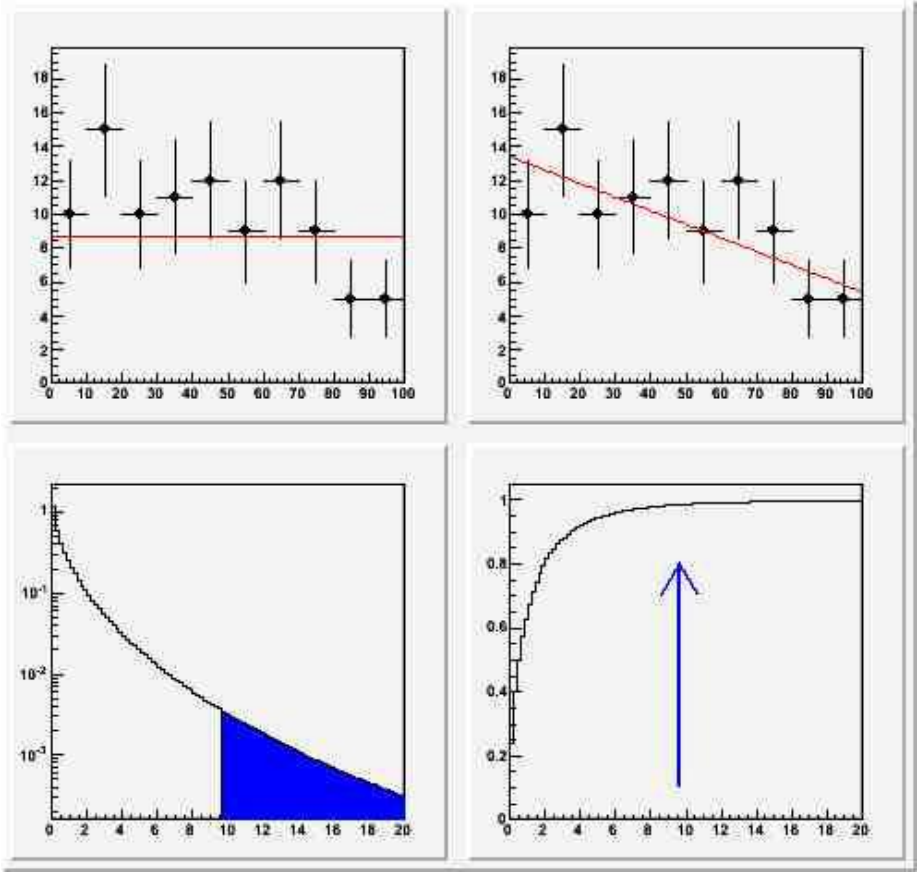
Example of F-test

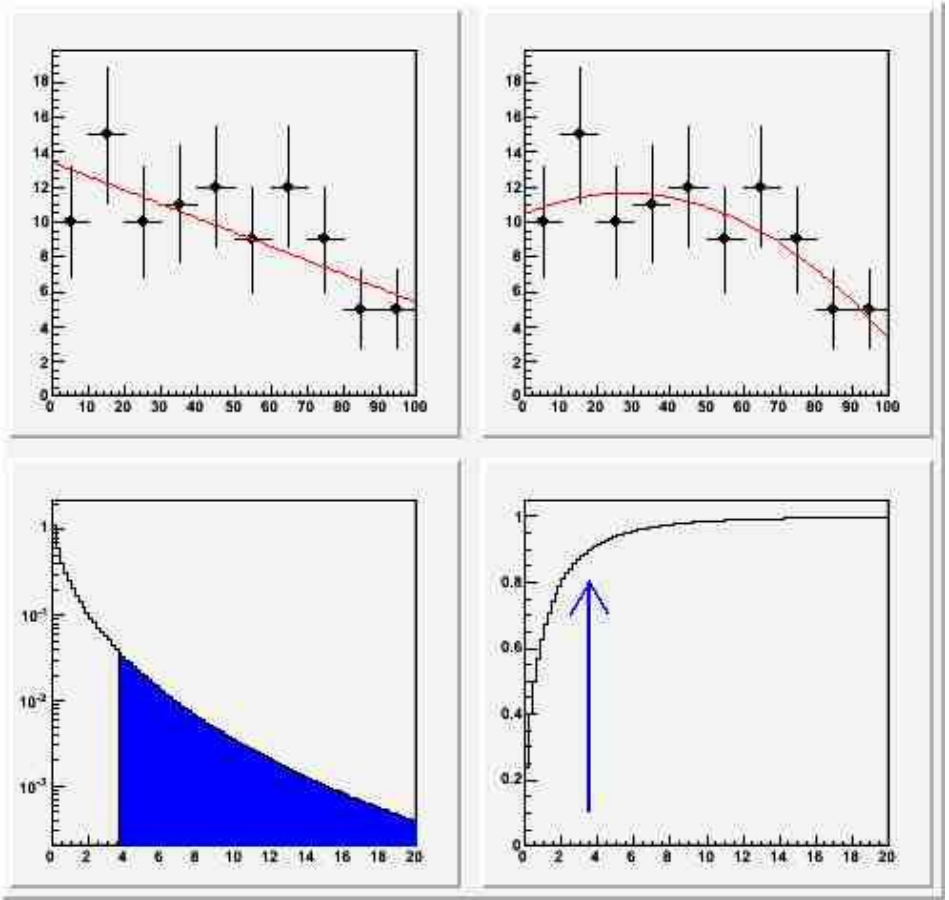
- Imagine you have the data shown on the right, and need to pick a functional form to model the underlying p.d.f.
- At first sight, any of the three choices shown produces a meaningful fit. P-values of the respective χ^2 are all reasonable (0.29, 0.84, 0.92)
- The F-test allows us to pick the right choice, by determining whether the additional parameter in going from a constant to a line, or from a line to a quadratic, is really needed.
- We need to pre-define **a size** of our test: we will reject the "null hypothesis" that the additional parameter is useless if **p<0.05**. We define p as the *probability that we observe a F value at least as extreme as the one in the data*, if it is drawn from a Fisher distribution with the corresponding degrees of freedom
- Note that we are implicitly also selecting a "region of interest" (large values of F)! More on this later.



How many of you would pick the constant model ? The linear ? The quadratic ?

The test between constant and line yields **p=0.0146**: there is evidence against the null hypothesis (that the additional parameter is useless), so we reject the constant pdf and take the linear fit





The test between linear and quadratic fit yields **p=0.1020**: there is no evidence against the null hypothesis (that the additional parameter is useless). We therefore keep the linear model.

The Neyman-Pearson Lemma

- For simple hypothesis testing there is a recipe to find the most powerful test. It is based on the likelihood ratio.
- Take data X={X_1…X_N} and two hypotheses depending on the values of a discrete parameter: H_0={θ=$θ_0$} vs H_1{θ=$θ_1$}. If we write the expressions of size α and power 1-β we have

$$\int_{w_\alpha} f_N(X \mid \theta_0) dX = \alpha$$

$$1-\beta = \int_{w_\alpha} f_N(X \mid \theta_1) dX$$

The problem is then to find the critical region w_α such that 1-β is maximized, given α. We rewrite the expression for power as

$$1-\beta = \int_{w_\alpha} \frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} f_N(X \mid \theta_0) dX$$

which is an expectation value:

$$= E_{w_\alpha}\left[\frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} \mid \theta = \theta_0 \right]$$

This is maximized if we accept in w_α all the values for which

$$l_N(X, \theta_0, \theta_1) = \frac{f_N(X \mid \theta_1)}{f_N(X \mid \theta_0)} \geq c_\alpha$$

So one chooses H_0 if $l_N(X, \theta_0, \theta_1) > c_\alpha$

and H_1 if instead $l_N(X, \theta_0, \theta_1) \leq c_\alpha$

In order for this to work, the likelihood ratio must be defined in all space; hypotheses must be simple. The test above is called Neyman-Pearson test, and a test with such properties is the **most powerful**.

Treatment of Systematic Uncertainties

- Statisticians call these *nuisance parameters*
- Any measurement in HEP is affected by them: the turning of an observation into a measurement requires assumptions about parameters and other quantities whose exact value is not perfectly known → their uncertainty affects the main measurement
  - Going from a event count to a cross section requires knowing N_b, L, ε_{sel}, ε_{trig} …
  - measurements which are subsidiary to the main result

- Inclusion of effect of nuisances in interval estimation and hypothesis testing introduces complications. Each of the methods has recipes, but not universal nor always applicable
  - **Bayesian treatment**: one constructs the multi-dimensional prior pdf $p(\theta)\Pi_i p(\lambda_i)$ including all the parameters λ_i, multiplies by $p(X_0|\theta,\lambda)$, and integrates all of the nuisances out, remaining with $p(\theta|X_0)$
  - **Classical frequentist treatment**: scan the space of nuisance parameters; for each point do Neyman construction, obtaining multi-dimensional confidence region; project on parameter of interest
  - **Likelihood ratio**: for each value of the parameter of interest θ^*, one finds the value of nuisances that globally maximizes the likelihood, and the corresponding $L(\theta^*)$. The set of such likelihoods is called the **profile likelihood.**

- Each "method" has problems (B: multi-D priors; C: overcoverage and intractability; L: undercoverage) – will not discuss them here, but note that this is a topic at the forefront of research, for which no general recipe is valid.
- Often used are "hybrid" methods for integrating nuisance parameters out: for instance, treat nuisance parameters in a Bayesian way while treating the parameter of interest in a frequentist way, or "profile away" the nuisance parameters and then use any method. Also possible is using Bayesian techniques and then evaluate their coverage properties.

Notes on Goodness-of-fit tests

- If H_0 is specified but the alternative H_1 is not, then only the Type I error rate α can be calculated, since the Type II error rate β depends on having specified a particular H_1.

  In this case the test is called a test for *goodness-of-fit (to H_0).*

- The question "Which g.o.f. test is best?" is ill-posed, since the power depends on the alternative hypothesis, which is not given.

- In spite of the popularity of tests which give a statistics one may directly connect with the size α (in particular χ^2 and Kolomogorov tests), their ability to discriminate against variations with respect to H_0 may be poor, i.e. they may have small power $(1-\beta)$ against relevant alternative hypotheses
  - χ^2 throws away information (sign, ordering)
  - Kolmogorov –Smirnov test only sensitive to biases, not to shape variations, and has terrible performance on tails

- It is in general hard to define what is random and what is not. Imagine you get three p-values: would you like to see them evenly spaced in [0,1] ? Would it induce you to doubt of the null if they all came out within 0.01 of 0.5 ? What if they are all close to 0.624 ? Or all close to zero ?

More on GoF

- Note the duality with confidence intervals: one might test the hypothesis $\theta=\theta_{test}$ using θ^* as test statistic. If we define the region $\theta^*>=\theta^*_{obs}$ as having equal or less agreement with the hypothesis than the result obtained, then the p-value of the test is α.
  - but for the c.i. the probability α is specified first, and the value θ_{test} is the random variable (depends on data); in a G.o.F. test for θ_{test}, we specify θ_{test} and the p-value is the result.

- In HEP, despite their limitations, Goodness-of-Fit tests are useful for a number of applications:
  - consistency checks
  - defining a control region
  - model testing
- The job of the experimenter is to find a suitable test statistic, *and* a region of interest of the latter. An example will clarify matters.

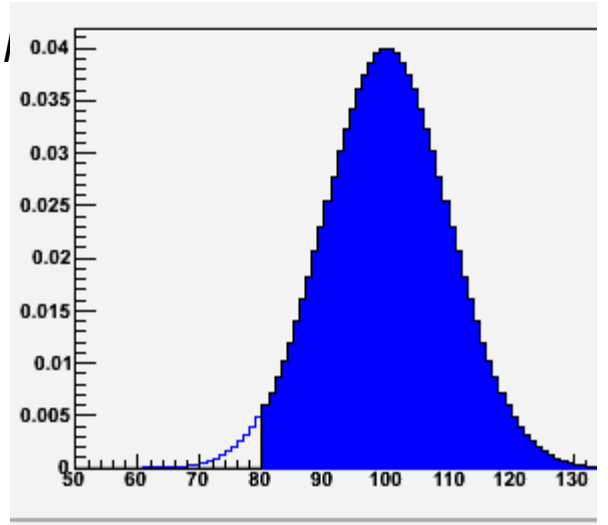
Choosing the region of interest

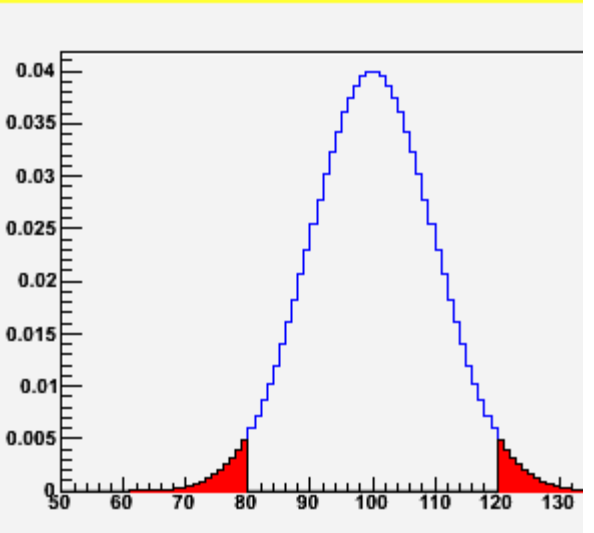
- Feynman's example:

  "*Upon walking here this morning, the strangest thing ever happened to me. A car passed by, and I could read the plate:* ***JKZ 0533****. How weird is that ??! The probability that I saw such a combination of letters and numbers (assuming they are all used in this country) is one in 10000*26^3, or one in eightyeight millions!*"

  Correct… The paradox arises from not having defined beforehand the region of interest!

- A more common one: you have a counting experiment where background is predicted to be 100 events. You observe 80 events. How rare is that ?

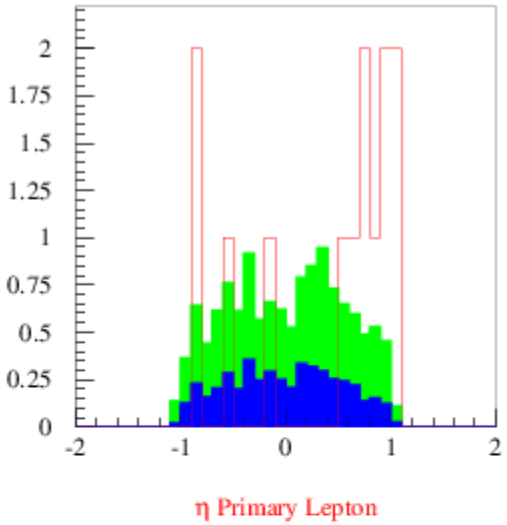




  - **Ill-posed question** ! Depends, to say the least, on whether you are interested only in excesses or in absolute departures!
  - In the first case the region of interest is N>=x, which, for x=80, corresponds to a fractional area p = 0.977.
  - In the second case, the region of interest is |N-100|>=|x-100| which for x=80 has an integral p = 0.0455.
  - And one might imagine other ways to answer – a no-brainer being $p=e^{-100}\ 100^{80}/80!$

The Kolmogorov Test: an example

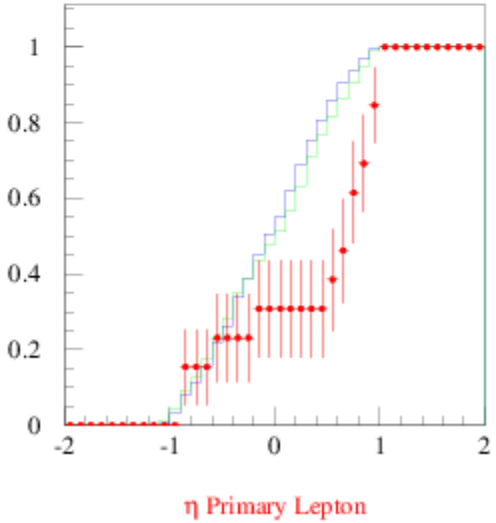
- CDF, circa 2000: 13 weird events identified in a subset of sample used to extract top quark cross section
  - contain a "superjet": a jet with a b-quark tag also containing a soft-lepton tag
  - expected 4.4 +-0.6 events from background sources
  - P(>=13|4.4+-0.6)=0.001
  - Kinematic characteristics found in stark disagreement with expectation from SM sources
- Have no alternative model to compare → try a Goodness-of-Fit test



- Kolmogorov-Smirnov test: compare cumulative distributions of data and model f(x); find largest difference

$$d_{KS} = \underset{x\in[a,b]}{Max}\left[\left|\int_a^x data(t)dt - \int_a^x f(t)dt\right|\right]$$

Value of d_{KS} can then be used to extract a p-value, given data size.



Intermezzo: combination of p-values

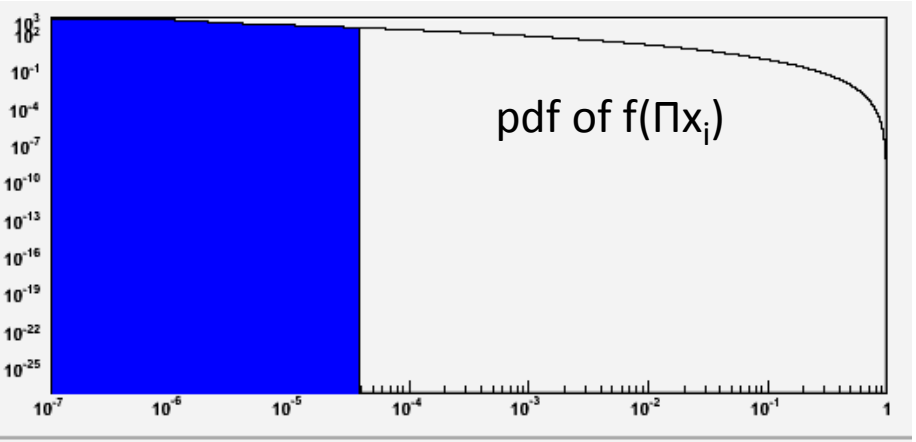
- Suppose you have several p-values, derived from different, independent tests. You may ask yourself several questions with them.
  - What is the probability that the smallest of them is as small as the one I got ?
  - What is the probability that the largest one is as small as the largest I observed ?
  - What is the probability that the product is as small as the one I can compute with these N values ?
- Please note! Your inference on the data at hand **strongly** depends on what test you perform, for a given set of data. In other words, you cannot choose which test to run only upon seeing the data…

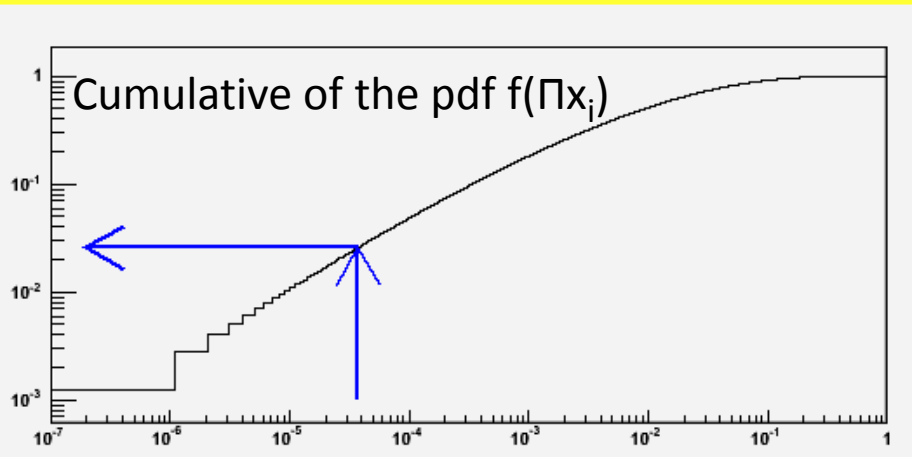
- Suppose anyway you believe that each p-value tells something about the null hypothesis you are testing, so you do not want to discard any of them. Then the reasonable (not the optimal!) thing to do is to use the product of the N values. The formula providing the cumulative distribution of the density of $x=\Pi x_i$ can be derived by induction (see [B.Roe 1992], p.129) and is

$$F_N(x) = x\sum_{j=0}^{N-1}\frac{1}{j!}|\log^j(x)|$$

This accounts for the speed with which the product of N numbers in [0,1] tends to zero as N grows.

Some examples





To start let us take five *really uniformly* distributed p-values, x_1=0.1, x_2=0.3, x_3=0.5, x_4=0.7, x_5=0.9. Their product is 0.00945, and with the formula just seen we get P(0.00945)=0.5017. As expected.

And what if instead x_1=0.00001, x_2=0.3, x_3=0.5, x_4=0.7, x_5=0.9 ? The result is P($9.45*10^{-7}$) =0.00123, which is rather large: one might think that the chance of getting one in five numbers as small as 10^{-5} must occur only a few times in 10^{-5}. But we are testing the product, not the smallest of the five numbers !

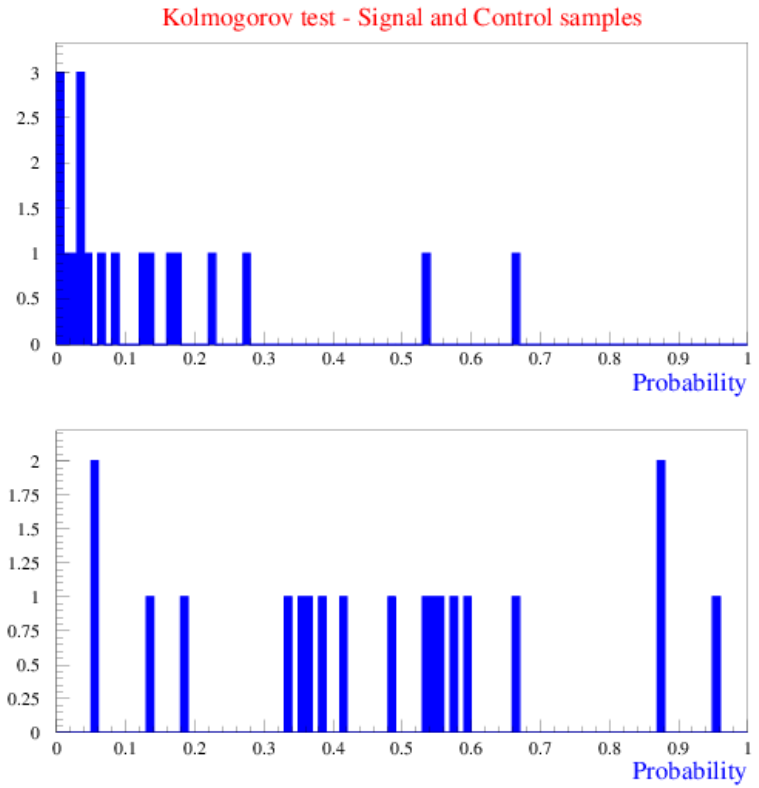
And if now we let x_1=0.05, x_2=0.10, x_3=0.15, x_4=0.20, x_5=0.25, the test for the product yields P($3.75*10^{-5}$)=0.0258 (see picture on the right).

Also not a compelling rejection of the null…

Compare with what you would get if you had asked "*what is the chance that five numbers are all smaller than 0.25 ?*", whose answer is $(0.25)^5$=0.00098. This demonstrates that the a-posteriori choice of the test is to be avoided !

Global P from set of p-values

- Authors of CDF “superjet” analysis tested a “complete set” of kinematical quantities; then computed global P of set of KS p-values using formula of combining p-values (assumed sampled from a Uniform distribution):



→ >6-sigma result!

… But in absence of an alternate model (really hard to cook given the weird kinematic properties of the set) one cannot thus “disprove” the Standard Model…

The real nature of events remained mysterious; at heated meetings, famous physicist argued that it was wrong to draw statistical inferences based on extreme values of some of the kinematical quantities

But the KS test is especially unsuited to spot those! In fact, one can move events in the tails back to center of distribution without p(KS) changing at all !!

GoF tests with Max Likelihood

- The maximum likelihood is a powerful method to estimate parameters, but no measure of GoF is given, because the value of L at maximum is not known, even under the hypothesis that the data are indeed sampled from the pdf model used in the fit
- The distribution of L_{max} can be studied with toy MC → one derives a p-value that a value as small as the one observed in the data arises, under the given assumptions
- Alternatively, one can bin the data, obtaining estimated mean values of entries per bin from the ML fit:

$$\hat{\nu}_i = n_{tot} \int_{x_i^{min}}^{x_i^{max}} f(x;\hat{\theta}) dx$$

Then one can derive a χ^2_L statistic using the ratio of likelihoods $\lambda = \frac{L(n|\nu)}{L(n|n)}$

and computing $\chi^2 = -2\log\lambda$

since in this case the latter follows a χ^2 distribution.

The quantity $\lambda(\nu)=L(n|\nu)/L(n|n)$ differs from the likelihood function by a normalization factor, and can thus be used for both parameter estimation and Goodness of fit.

Evaluating significance

- In HEP a common problem is the evaluation of a significance in a counting experiment. Significance is usually measured in "number of sigma's" → implicit Gaussian approx.

- We have already seen examples of this. It is common to cast the problem in terms of a Goodness-of-Fit test of a null hypothesis H_0

- Expect *b* events from background, test for a signal contributing *s* events by a Poisson experiment: then

$$f(n|b+s) = (b+s)^n e^{-(b+s)}/n!$$

- Upon observing N_{obs}, can assign a probability to the observation as

$$P(n \geq N_{obs}) = 1 - \sum_{n=0}^{N_{obs}-1} \frac{b^n e^{-b}}{n!}$$

- **Please note**: this is not the probability of H_0 being true !! It is the probability that, H_0 being true, we observe N_{obs} events or more

- Take b=1.1, N_{obs}=10: then p=2.6E-7 → a 5σ discovery. Similar for b=0.05, N_{obs}=4.

- **Also, please note:** if you use a small number of events to measure a cross section, you will have large error bars (whatever your method of evaluating a confidence interval for the true mean!). For instance if b=0, N=5, Likelihood-ratio intervals give 3.08 < s < 7.58, i.e. $s=5_{-1.92}^{+2.58}$. Does that mean we are less than 3-sigma away from zero ? NO !

Bump hunting: Wilks' theorem

- A typical problem in HEP: test for the presence of a Gaussian signal on top of a smooth background, using a fit to B(M) (H_0: null hypothesis) and a fit to B(M)+S(M) (H_1: alternative hypothesis)
- This time we have both H_0 and H_1. One can thus easily derive the local significance of a peak from the likelihood values resulting from fits to the two hypotheses. The standard recipe uses **Wilks' theorem**:
  - get L_0, L_1
  - evaluate -2ΔLogL
  - Obtain p-value from probability that $\chi^2(N_{dof})>-2\Delta LogL$
  - Convert into number of sigma for Gaussian distribution using the inverse of the error function
  - Four lines of code !
- Convergence of -2ΔlnL to χ^2 distribution is fast. But certain regularity conditions need to hold! In particular, models need to be nested, and we need to be away from a boundary in the parameter of interest.
  - In principle, allowing the mass of the unknown signal to vary in the fit violates the conditions of Wilks' theorem, since for zero signal normalization H_0 corresponds to any $H_1(M)$ (mass is undefined under H_0: it is a *nuisance parameter present only in the alternative hypothesis*);
  - But it can be proven that approximately Wilks' theorem still applies (see [Gross 2010])
  - Typically one runs toys to check the distribution of p-values
  - but this is not always practical
- Upon obtaining the local significance of a bump, one needs to account for the multiplicity of places where the signal might have arisen by chance.
  - Is rule of thumb valid ? $TF = (M_{max}-M_{min})/\sigma_M$

More on the Look-Elsewhere Effect

- The problem of accounting for the multiplicity of places where a signal could have arisen by chance is apparently easy to solve:
  - Rule of thumb ?
  - Run toys by simulating a mass distribution according to H_0 alone, with $N=N_{obs}$ (remember: **thou shalt condition!**), deriving the distribution of $-2\Delta\ln L$

- Running toys is sometimes impractical (see Higgs combination); it is also illusory to believe one is actually accounting fully for the trials factor
  - In typical analyses one has looked at a number of distributions for departures from H_0
  - Even if the observable is just one (say a M_{jj}) one often is guilty of having checked many possible cut combinations
  - If a signal appears in a spectrum, it is often natural to try and find the corner of phase space where it is most significant; then "a posteriori" one is often led into justifying the choice of selection cuts
  - A HEP experiment runs O(100) analyses on a given dataset and O(1000) distributions are checked for departures. A departure may occur in any one of 20 places in a histogram → trials factor is O(20k)
  - This means that one should expect a 4-sigma bump to naturally arise by chance in any given HEP experiment ! (→ Well borne out by past experience…) Beware of quick conclusions!

- In reality the trials factor depends also on the significance of the local fluctuation (which can be evaluated by fixing the mass, such that $\Delta N_{dof}=1$). Gross and Vitells [Vitells 2010] demonstrate that a better "rule of thumb" is provided by the formula

$$TF = k\frac{M_{\max}-M_{\min}}{\sigma_M}Z_{fix}$$

where k is typically 1/3 and can be estimated by counting the average number of local minima $\langle N\rangle = k\,(M_{max}-M_{min})/\sigma_M$

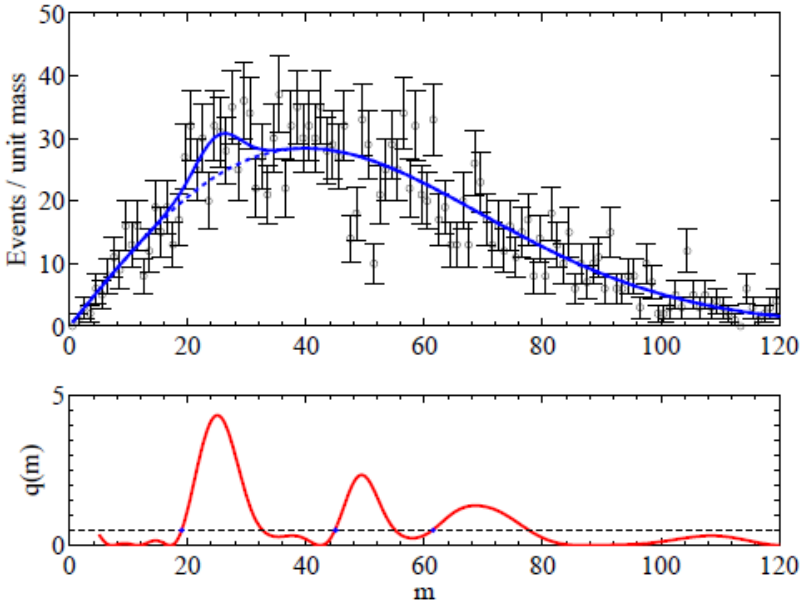
Local minima and upcrossings

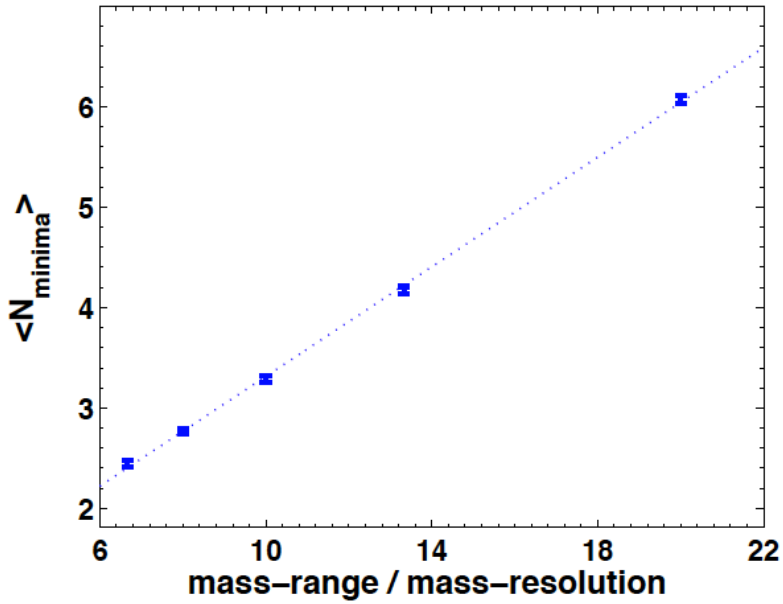
When dealing with complex cases (Higgs combination), a study [Vitells 2010] comes to help. One counts the number of "upcrossings" of the distribution of p-values, or the value of the test statistics itself, as a function of mass. Its wiggling tells how many independent places one has been searching in ! [CMS 2011]

- The number of local minima in the fit to a distribution is closely connected to the freedom of the fit to pick signal-like fluctuations in the investigated range

The number of times that the test statistics (below, the likelihood ratio between H_1 and H_0) crosses some reference point is a measure of the trials factor. One estimates the global p-value with the number N_0 of upcrossings from a minimal value of the q_0 test statistics (for which $p=p_0$) by the formula

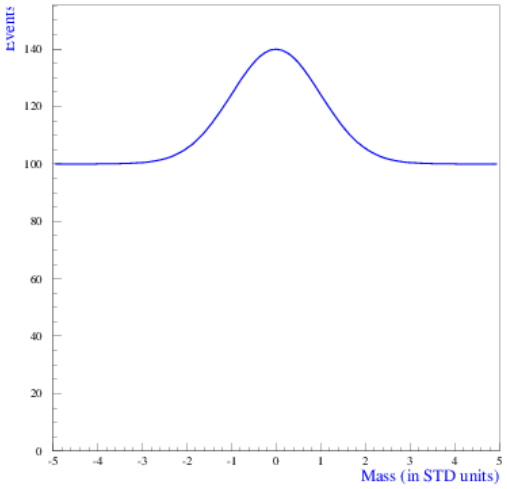
$$p_0^{\text{global}} \sim p_0^{\text{min}} + N_0\, e^{-\frac{1}{2} Z_{\text{max}}^2}$$

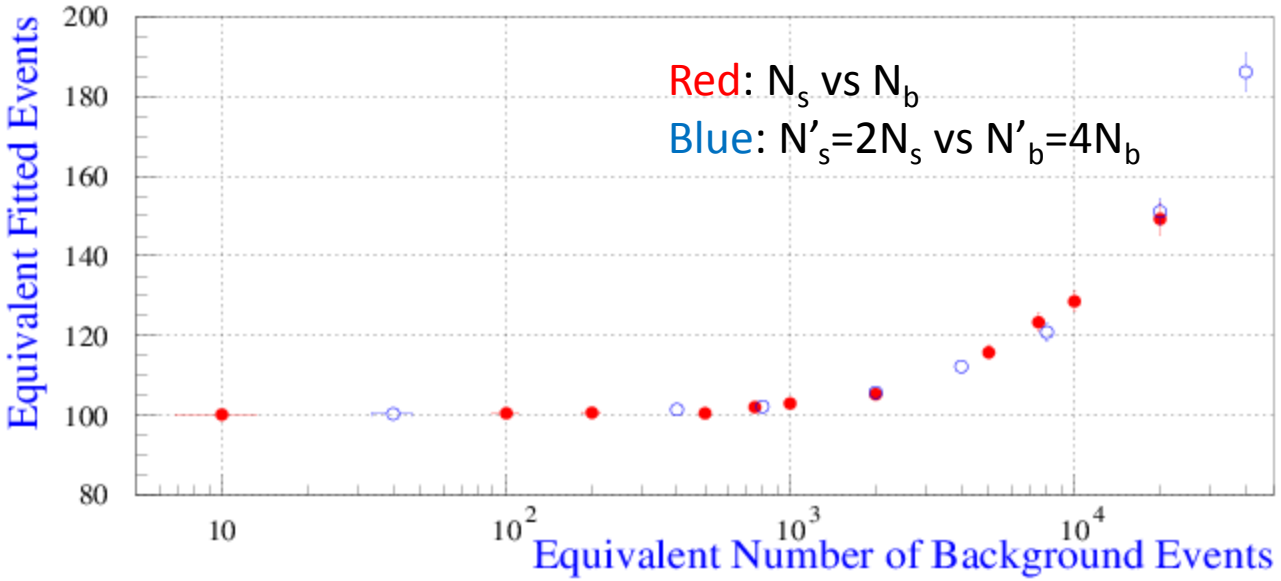




Second-order LEE

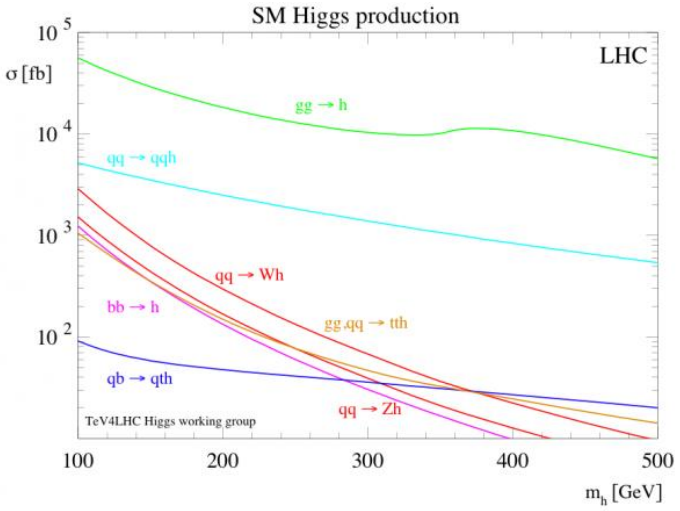
- Besides the above discussed approximate methods to compute the trials factor, there are practical ways to overcome the LEE bias
- The typical, sound recipe of the navigated HEP researcher to prevent the problem of LEE in estimating significance: upon observing a signal, wait for a new set of data, freezing cuts and the signal mass.
- But care is still required! In the fit to the second half of your data you cannot allow the mass to float around, not even only "just a bit", in the region where you spotted the signal
- In fact, there is a subtle, second-order LEE at work. The fitter will "pick up" the noise around the signal, biasing the signal normalization and the corresponding significance to be larger. This is connected with the linear growth of the trials factor with Z already discussed.
- Effect dubbed "Greedy bump bias" in [Dorigo 2000].

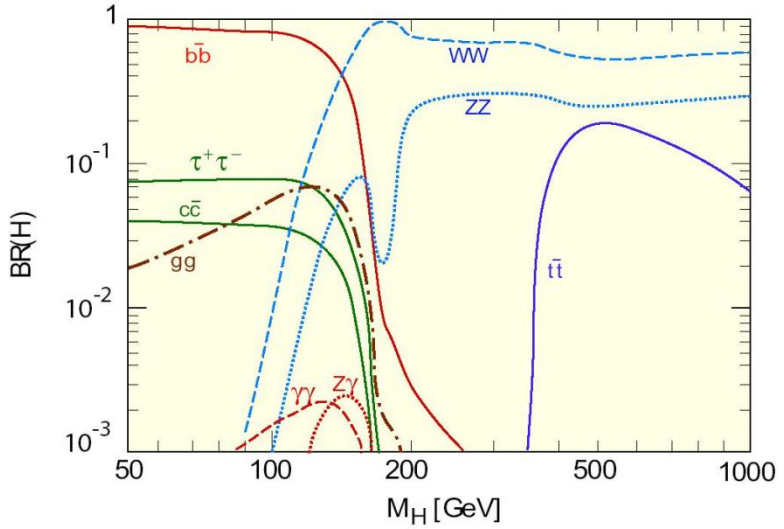




Higgs Searches at LHC

- The Higgs boson has been sought by ATLAS and CMS in all the main production processes and in a number of different final states, resulting from the varied decay modes:
  - qq→Hqq
  - gg→H
  - qq(′)→VH
  - H→ZZ
  - H→WW
  - H→gg
  - H→tt
  - H→bb





- The importance of the goal brought together some of the best minds of CMS and ATLAS, to define and refine the procedures to combine the above many different search channels, most of which have marginal sensitivity by themselves
- The method used to set upper limits on the Higgs boson cross section is called CL_s and the test statistics is a profile log-likelihood ratio. Dozens of nuisance parameters, with either 0% or 100% correlations, are considered
- Results are produced as a combined upper limit on the "strength modifier" $\mu=\sigma/\sigma_{SM}$, as well as a "best fit value" for μ, and a combined p-value of the null hypothesis. All of these are produced as a function of the unknown Higgs boson mass.
- The technology is strictly experts-only stuff, and it would take a couple of hours to go through all the main issues. We can just give a peek at the construction of the CL_s statistics, to understand the main architecture

Nuts and Bolts of Higgs Combination

The recipe must be explained in steps. The first one is of course the one of writing down extensively the likelihood function!

1) One writes a global likelihood function, whose parameter of interest is the strength modifier μ. If s and b denote signal and background, and θ is a vector of systematic uncertainties, one can generically write for a single channel:

$$\mathcal{L}(\text{data}\,|\,\mu,\theta) = \text{Poisson}\,(\,\text{data}\,|\,\mu\cdot s(\theta)+b(\theta)\,)\cdot p(\tilde{\theta}|\theta)$$

Note that θ has a "prior" coming from a hypothetical auxiliary measurement.

In L one may combine many different search channels where a counting experiment is performed as the product of their Poisson factors:

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!}\; e^{-\mu s_i - b_i}$$

or from a unbinned likelihood over k events, factors such as:

$$k^{-1}\prod_i (\mu S f_s(x_i) + B f_b(x_i))\;\cdot\; e^{-(\mu S+B)}$$

2) One then constructs a profile likelihood test statistics q_μ as $\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\hat{\theta}}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}$

Note that the denominator has L computed with the values of $\hat{\mu}$ and $\hat{\theta}$ that globally maximize it, while the numerator has $\theta=\hat{\theta}_\mu$ computed as the conditional maximum likelihood estimate, given μ.

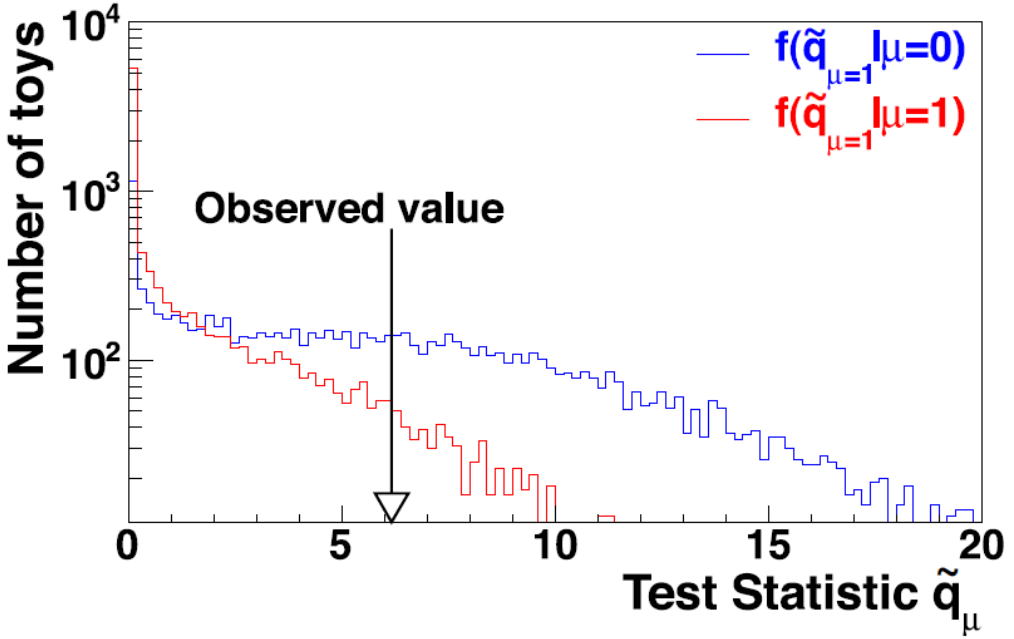
A constraint is posed on the MLE $\hat{\mu}$ to be confined in $0 \leq \hat{\mu} \leq \mu$: this avoids negative solutions and ensures that best-fit values *above* the signal hypothesis μ are not counted as evidence against it.

The above definition of a test statistics for CL_s in Higgs analyses differs from earlier instantiations

- LEP: no profiling of nuisances
- Tevatron: μ=0 in L at denominator

3) ML values of θ_μ for H_1 and θ_0 for H_0 are then computed, given the data

4) Pseudo-data is then generated for the two hypotheses, using the above ML estimates of the nuisance parameters. With the data, one constructs the pdf of the test statistics given a signal of strength μ (H_1) and μ=0 (H_0).



5) With the pseudo-data one can then compute the integrals defining p-values for the two hypotheses. For the signal plus background hypothesis H_1 one has

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{signal+background}) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{\text{obs}})\, d\tilde{q}_\mu$$

and for the null, background-only H_0 one has

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{background-only}) = \int_{q_0^{\text{obs}}}^{\infty} f(\tilde{q}_\mu|0, \hat{\theta}_0^{\text{obs}})\, d\tilde{q}_\mu$$

6) Finally one can compute the value called CL_s as

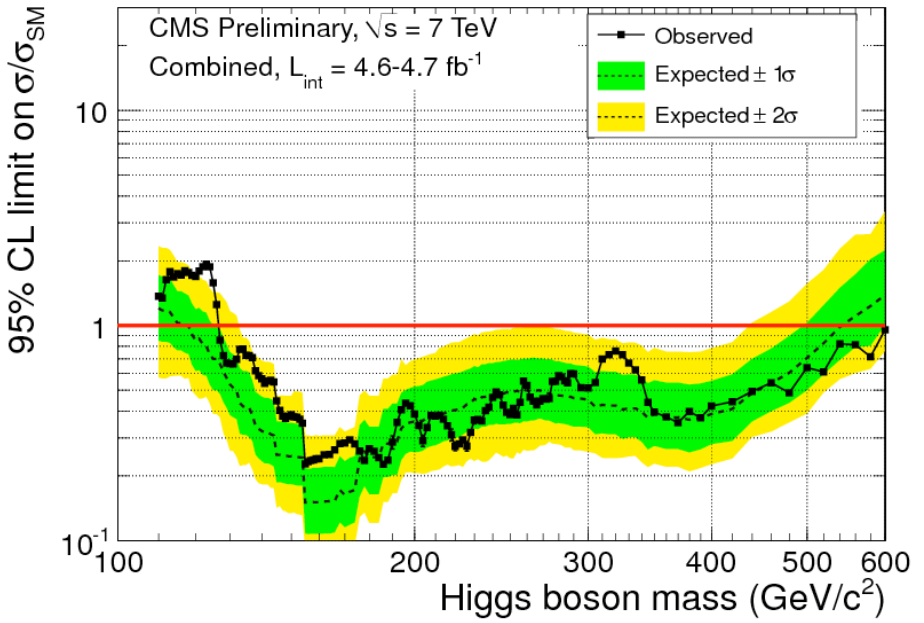
$$CL_s = p_\mu/(1-p_b)$$

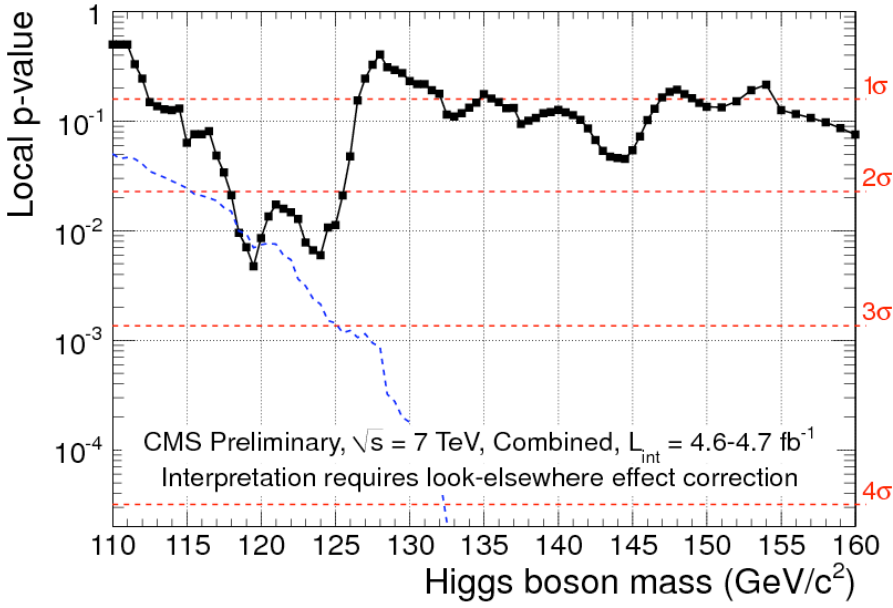
CL_s is thus a "modified" p-value, in the sense that it describes how likely it is that the value of test statistics is observed under the alternative hypothesis by also accounting for how likely the null is: the drawing incorrect inferences based on extreme values of p_μ is "damped", and cases when one has no real discriminating power, approaching the limit $f(q|\mu)=f(q|0)$, are prevented from allowing to exclude the alternate hypothesis.

7) We can then **exclude H_1 when $CL_s < \alpha$**, the (defined in advance !) *size* of the test. In the case of Higgs searches, all mass hypotheses $H_1(M)$ for which $CL_s<0.05$ are said to be excluded (one would rather call them "disfavoured"…)

Results of Higgs Search: CMS

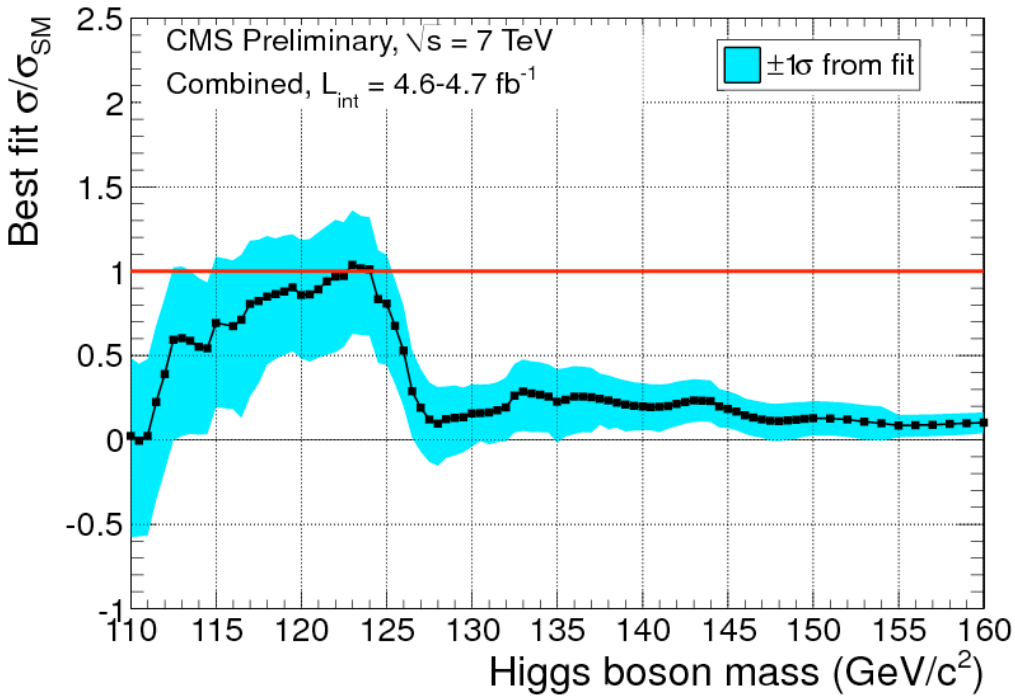
- Let us take the December 2011 CMS result as an example for looking at a few graphs.
- The observed limit on μ is compared with the expected one. The latter is derived from pseudo-data by performing the same procedure as on real data, deriving the shape of the 95% CL limit with CL_s for each mass point, and calculating the percentiles (2.3%, 15.9%, 50%, 84.1%, 97.7%) corresponding to median and 1- and 2-sigma bands
- To investigate the excess of events in the 118-125 GeV region, one may plot the p-value of the data given H_0. A comparison with the expected p-value given H_0 if the data contain a SM Higgs (with μ=1) is overlaid (blue dashes) only as a visual aid, and does not constitute a real test of that hypothesis



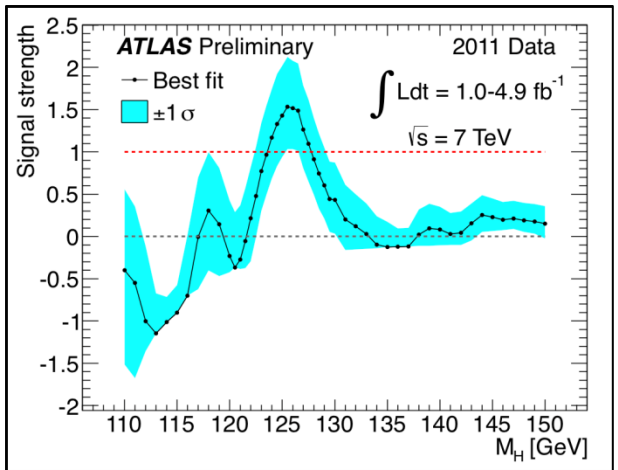


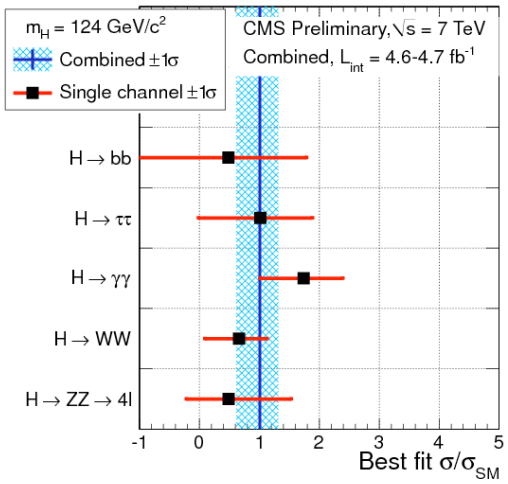
Best-fit σ/σ_{SM}

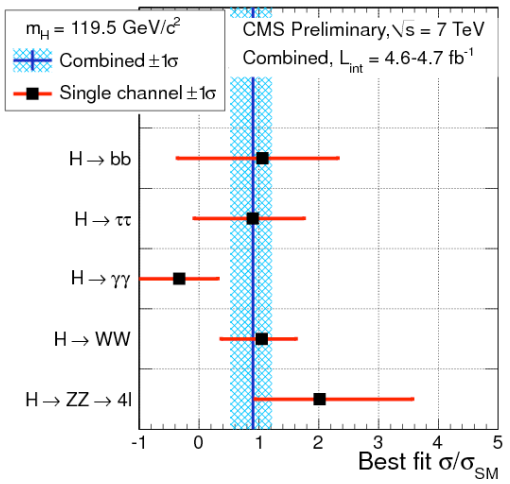
A better visual test of H_1 is provided by computing the best-fit value of μ from the likelihood function. This provides a more quantitative estimate of the compatibility of the data with the signal hypothesis.



- Best-fit μ values for the individual channels may be also compared for any given mass hypothesis. There is overall good compatibility between the CMS data and either M_H=119 and M_H=124 GeV; the latter appears more probable, given the ATLAS results (below, left).







Conclusions

- **Statistics is NOT trivial**. Not even in the simplest applications!
- A understanding of the different methods to derive results (eg. for upper limits) is crucial to make sense of the often conflicting results one obtains even in simple problems
  - The key in HEP is to try and derive results with different methods –if they do not agree, we get wary of the results, plus we learn something

- Making the right choices for what method to use is an expert-only decision, so…

  You should become an **expert in Statistics**, if you want to be a good particle physicist (or even if you want to make money in the financial market)
- The slide of this course are nothing but an appetizer. To really learn the techniques, you must put them to work

- Be careful about what statements you make based on your data! You should now know how to avoid:
  - Probability inversion statements: "The probability that the SM is correct given that I see such a departure is less than x%"
  - Wrong inference on true parameter values: "The top mass has a probability of 68.3% of being in the 171-174 GeV range"
  - Apologetic sentences in your papers: "Since we observe no significant departure from the background, we proceed to set upper limits"
  - Improper uses of the Likelihood: "the upper limit can be obtained as the 95% quantile of the likelihood function"
  - MINOS-like custom-made procedures: "The 95% CL limit can be combined with an earlier result by the formula …."

References

[James 2006] F. James, *Statistical Methods in Experimental Physics* (IInd ed.), World Scientific (2006)

[Cowan 1998] G. Cowan, *Statistical Data Analysis*, Clarendon Press (1998)

[Cousins 2009] R. Cousins, HCPSS lectures (2009)

[D'Agostini 1999] G. D'Agostini, *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN Yellow Report 99/03 (1999)

[Stuart 1999] A. Stuart, K. Ord, S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A, 6th edition (1999)

[Cox 2006] D. Cox, *Principles of Statistical Inference*, Cambridge UP (2006)

[Roe 1992] B. P. Roe, Probability and Statistics in Experimental Physics, Springer-Verlag (1992)

[Tucker 2009] R. Cousins and J. Tucker, 0905.3831 (2009)

[Cousins 2011] R. Cousins, Arxiv:1109.2023 (2011)

[Cousins 1995] R. Cousins, "Why Isn't Every Physicist a Bayesian ?", Am. J. Phys. 63, n.5, pp. 398-410 (1995)

[Gross 2010] E. Gross, "Look Elsewhere Effect", Banff (2010) (see p.19)

[Vitells 2010] E. Gross and O. Vitells, "Trials factors for the look elsewhere effects in High-Energy Physics", Eur.Phys.J.C70:525-530 (2010)

[Dorigo 2000] T. Dorigo and M. Schmitt,"On the significance of the dimuon mass bump and the greedy bump bias", CDF-5239 (2000)

[ATLAS 2011] ATLAS and CMS Collaborations, ATLAS-CONF-2011-157 (2011); CMS PAS HIG-11-023 (2011)

[CMS 2011] ATLAS Collaboration, CMS Collaboration, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in summer 2011", ATL-PHYS-PUB-2011-818, CMS NOTE-2011/005 (2011).

Also cited (but not on statistics):

[McCusker 1969] C.McCusker, I.Cairns, PRL 23, 658 (1969)

[MINOS 2011] P. Adamson et al., Arxiv:1201.2631 (2011)