SPRACE

Aspects of HEP – Data Analysis

Thiago R. F. P. Tomei

SPRACE-Unesp

Introduction to Data Analysis

In broad terms, we can divide the LHC data analyses in two camps:

- □ **Measurements**, in which one is trying to measure some known standard model quantity. Examples go from very simple quantities, like the Higgs boson mass M_H, to convoluted quantities like the fully inclusive top quark pair production at the LHC at 14 TeV, $\sigma(\mathrm{pp} \to \mathrm{t\bar{t}} +$ anything$)$.
- □ **Searches**, in which one tries to uncover evidence of discrepancies between the standard model predictions and the observed data. Examples include new searches for resonances, supersymmetry, dark matter...

But this distinction is a bit artificial! Consider:

- □ $\mathrm{H} \to \mu\mu$ at the 13 TeV LHC has a well-defined SM prediction: $\sigma(\mathrm{pp} \to \mathrm{H} \to \mu\mu) \simeq$ 12.08 fb. This has not been observed yet, so we call it a search.
- □ When searching for a new resonance, $\mathrm{pp} \to \mathrm{Z}' \to \mathrm{ee}$, we usually make (multiple) assumptions on the value of its mass, spin, etc. We then try to measure its production cross-section – usually coming up with a value statistically compatible with zero.

What do we Actually Measure?

Easy answer – **we count the number of events in a given configuration**. The observed number of events for a given physics process N_{obs} is:

$$\boxed{N_{\text{obs}} = \sigma \times \mathcal{L} \times A \times \epsilon}$$

where:

- □ σ is the production cross-section. This is a purely theoretical-driven quantity – it is the very σ we learned how to calculate on the first lecture.
- □ $\mathcal{L}$ is the accelerator luminosity. This is a measure of how many particles we are able to fit through a given space in a given time. We discussed it on the second lecture.
- □ A is the acceptance. It measures, for that given process, the ratio of detectable particles that actually go into the detector volume.
  - Technically, this is also theoretically-driven, but it is customary to factor it out like this.
- □ ϵ is the selection efficiency. It can be subdivided into two parts, $\epsilon = \epsilon_o \times \epsilon_a$:

- ϵ_o is the efficiency of reconstructing a given set of objects in the detector. We discussed it on the third lecture.
- ϵ_a is the efficiency of any further requirements done in the analysis (a.k.a. fun ☺).

The “given configuration” we discussed above is determined by the acceptance × efficiency product. The total number of events we observe in that configuration is:

$$N_{\mathrm{obs}} = \sum_i N^i_{\mathrm{obs}} + N_{\mathrm{fakes}}$$

where the sum runs over **all** of the physics process that produce events in that configuration. There may be spurious contributions N_{fakes} from any kinds of non-collision effects. Those include:

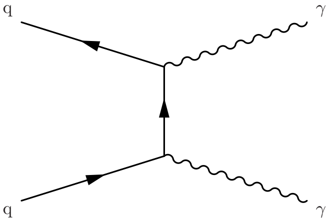
- Cosmic rays going through the detector (fake muons).
- Longitudinal particles from beam interactions with accelerator elements (beam halo).
- Hardware failures: “hot cells”, dead channels, high voltage spikes.
- Software failures: e.g., unusual configurations of hits leading to high numbers of fake tracks.
- The phase of the moon (?)
    - Nucl. Instrum. Methods Phys. Res., A 357 (1995) 249-252
- The seasonal variation of rainfall (???)
    - Proceedings of the 1999 Particle Accelerator Conference, New York, 1999
- The schedule of the French high-speed rail trains (?????)
    - Nucl. Instrum. Methods Phys. Res., A 417 (1998) 9-15

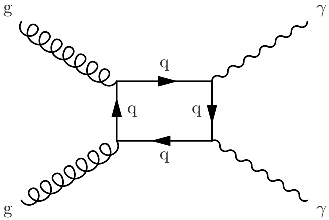
But let’s go back to collision processes...

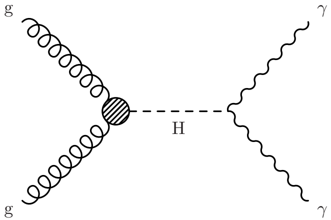
Signal and Background

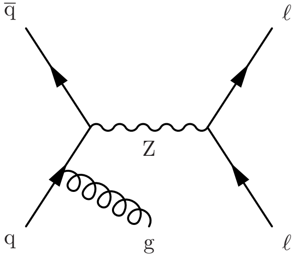
The fact that the sum $\sum_i N^i_{\mathrm{obs}}$ goes over all physics processes means that we cannot readily separate the process in which we are interested – the **signal**. All of the other processes constitute the **background** for that measurement. We can separate backgrounds in two sets:

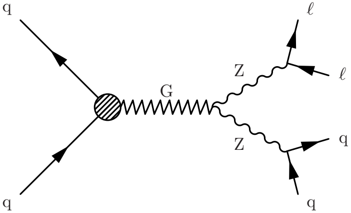
- □ Irreducible backgrounds are those that share the exact same final state as the signal. For instance, the nonresonant diphoton production $\mathrm{pp} \to \gamma\gamma$ is an irreducible background to Higgs boson production in that channel, $\mathrm{pp} \to \mathrm{H} \to \gamma\gamma$. The only option is to model them as well as possible.
- □ Reducible backgrounds are those where the final state differs from that of the signal, but due to various reasons end up being selected by our analysis. An example would be inclusive $\mathrm{Z} \to \ell\ell$ production being a background for a $\mathrm{ZZ} \to \ell\ell\mathrm{qq}$ search: the former could appear as a "dilepton $+$ jet" final state, and that jet could be mistaken to be the $\mathrm{Z} \to \mathrm{qq}$ leg of the latter.


q
γ
q
γ


g
γ
q
q
q
q
g
γ


g
γ
H
g
γ


$\overline{\mathrm{q}}$
ℓ
Z
q
g
ℓ


q
ℓ
Z
ℓ
G
Z
q
q
q

Separating Signal and Background

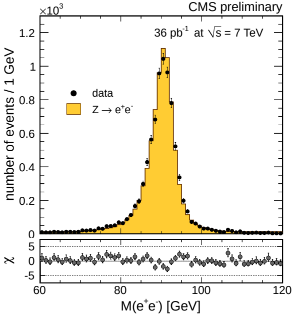
Our first task is to search for observables that are differently distributed for signal and background. Some searches are easy, for $\mathrm{Z} \to \ell\ell$:

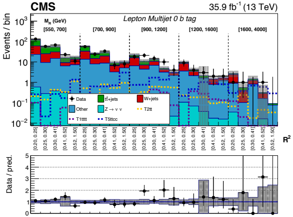
- □ 2 high-p_{T} leptons.
- □ Same flavour, opposite charges.
- □ Invariant mass $M_{\ell\ell}$ in 70–110 GeV range.

Other searches are harder. The Razor variables used for SUSY searches:

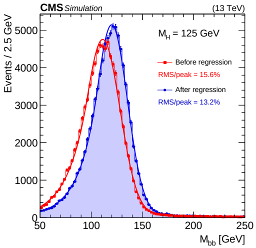
$$M_{\mathrm{R}} \equiv \sqrt{\left(|\vec{p}^{\,j_1}| + |\vec{p}^{\,j_2}|\right)^2 - \left(p_z^{j_1} + p_z^{j_2}\right)^2} \text{ and } R^2 \equiv \left(\frac{M_{\mathrm{T}}^{\mathrm{R}}}{M_{\mathrm{R}}}\right)^2$$

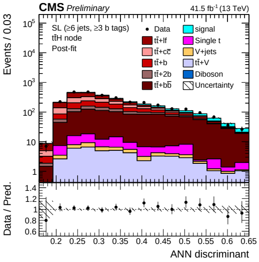
$$\text{with } M_{\mathrm{T}}^{\mathrm{R}} \equiv \sqrt{\frac{p_{\mathrm{T}}^{\mathrm{miss}}\left(p_{\mathrm{T}}^{j_1} + p_{\mathrm{T}}^{j_2}\right) - \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} \cdot \left(\vec{p}_{\mathrm{T}}^{\,j_1} + \vec{p}_{\mathrm{T}}^{\,j_2}\right)}{2}}$$





Machine Learning in Analysis



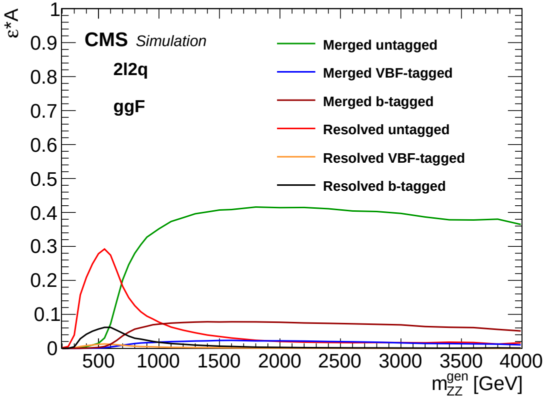


Some times, we don't need to search for variables... the computer can do it for us! Machine Learning (ML) techniques can be applied to two kinds of problems:

- □ **Classification problems:** the output variable takes class labels. Useful for analysis, e.g. classify the event as signal or background.
  - Usually we unpack the classification and work with the output variable directly.
- □ **Regression problems:** the output variable takes continuous values. Useful for reconstruction, e.g. energy of a b-jet based on its kinematics and flavour content.

Cut-Based Analyses

- □ Design a set of cuts by as unbiased a procedure as possible.
  - Blind study: try to avoid observer bias and confirmation bias. Don't look at the data until you have frozen the analysis.
- □ Choose cuts that optimise the final overall accuracy of the result.
  - Difficult tradeoff between statistical and systematic uncertainties (see later).
- □ Always study the marginal effect of each of your cuts by tables and plots.
  - Cuts with no marginal effect (that is, they remove no events after all other cuts) are quite useless.

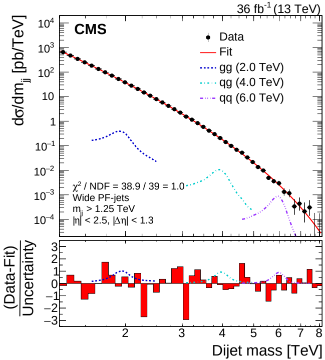


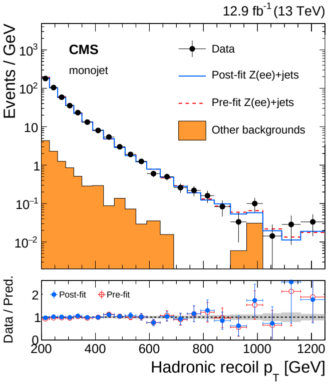
Background Estimation

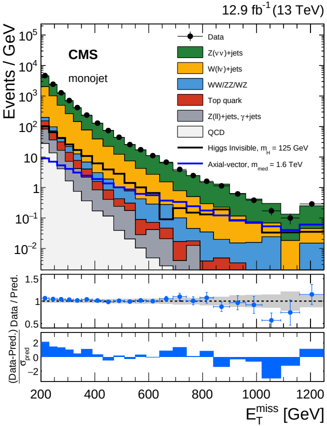
After you optimise your analysis, some background always* remains. An array of options are available to estimate it:

- Fully trust the simulation: not recommended, except maybe for backgrounds so small that even an error by a large factor would make no difference.
- Trust a "data corrected" simulation: usually done by defining **control regions** (a.k.a "sidebands"), in which you expect similar behaviour of your background but a near-absence of your signal.
- Model your background "in situ": the same, but your control regions act simultaneously as measurement regions for some other modality of your search.

* Even if it doesn't, you still have to find a way to put an uncertainty on that zero!


36 fb^{-1} (13 TeV)
CMS
Data
Fit
gg (2.0 TeV)
qg (4.0 TeV)
qq (6.0 TeV)
χ^2 / NDF = 38.9 / 39 = 1.0
Wide PF-jets
m_{jj} > 1.25 TeV
$|\eta|$ < 2.5, $|\Delta\eta|$ < 1.3
$d\sigma/dm_{jj}$ [pb/TeV]
10^4
10^3
10^2
10
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
(Data-Fit) / Uncertainty
3
2
1
0
−1
−2
−3
2
3
4
5
6
7
8
Dijet mass [TeV]


12.9 fb^{-1} (13 TeV)
CMS
monojet
Data
Post-fit Z(ee)+jets
Pre-fit Z(ee)+jets
Other backgrounds
Events / GeV
10^3
10^2
10
1
10^{-1}
10^{-2}
Data / Pred.
Post-fit
Pre-fit
2
1
0
200
400
600
800
1000
1200
Hadronic recoil p_T [GeV]


12.9 fb^{-1} (13 TeV)
CMS
monojet
Data
Z(νν)+jets
W(lν)+jets
WW/ZZ/WZ
Top quark
Z(ll)+jets, γ+jets
QCD
Higgs Invisible, m_H = 125 GeV
Axial-vector, m_{med} = 1.6 TeV
Events / GeV
10^5
10^4
10^3
10^2
10
1
10^{-1}
10^{-2}
Data / Pred.
1.5
1
0.5
(Data-Pred.) / σ_{pred}
2
0
−2
200
400
600
800
1000
1200
E_T^{miss} [GeV]

Basic Concepts in Probability

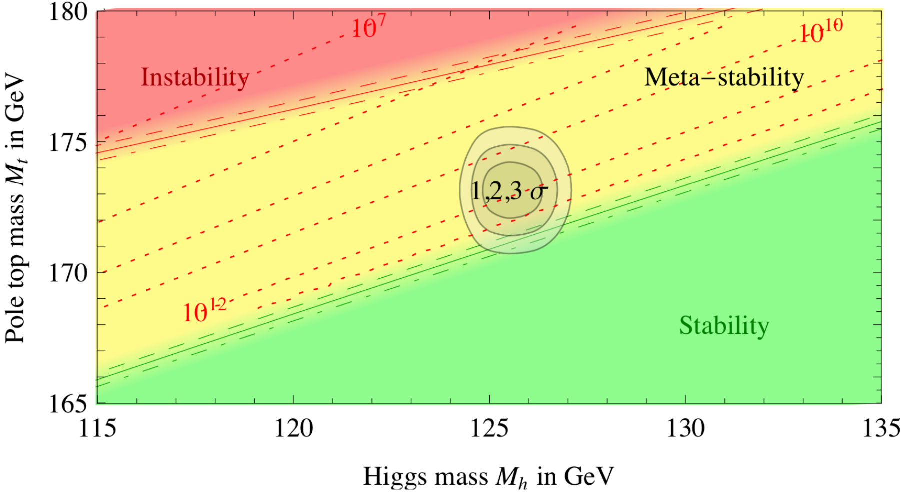
- □ **Mathematical probability:** given the set of all possible exclusive, elementary events X_i, probability of occurrence of X_i is $P(X_i)$ and follows the Kolmogorov axioms:
  - $P(X_i) \geq 0 \quad$ for all i
  - $P(X_i$ or $X_j) = P(X_i) + P(X_j)$
  - $\sum_{\Omega} P(X_i) = 1$
- □ **Frequentist probability:** if your observe N events, and n of them are of type N, the probability that any single event will be of type X is the "empirical" limit of the frequency ratio:

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

  - Approximate the probability by making N large.
  - Experiments have to be **repeatable** – but repeatable means that all the **relevant** conditions are the same. Good science should produce **reproducible results**.

□ **Bayesian probability:** it is the degree of belief in X. Operational definition is based on the **coherent bet:** [Finetti1974]

- "The idea is to determine how strongly a person believes that X will occur by determining how much he would be willing to bet on it, assuming that he would be willing to bet on it, assuming that he wins a fixed amount if X does later occur and nothing if it fails to occur. Then $P(X)$ is defined as the largest amount he would be willing to bet, divided by the amount he stands to win." [James2006].
- This follows the Kolmogorov axioms.
- However, it is a property of both the observer and the observed system – it will in general change if the observer obtains more knowledge. It is a **subjective** probability!
- On the other hand, it helps addressing some questions that we want to try to answer:
  - ○ "What is the probability that the universe is (cosmologically) flat?"
  - ○ "What is the probability that the Higgs vacuum is stable?"
- There is a lot of work in studying objective Bayesian statistics (H Jeffreys, E. T. Jaynes, S. James, J. Berger,…. The science is far from settled!


180
175
170
165
115
120
125
130
135
Pole top mass M_t in GeV
Higgs mass M_h in GeV
Instability
Meta−stability
Stability
10^7
10^{10}
10^{12}
1,2,3 σ

Bayes Theorem

For discrete events:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \overset{\text{Bayesian}}{\Rightarrow} P(\theta_i|\boldsymbol{X}^0) = \frac{P(\boldsymbol{X}^0|\theta_i) \cdot P(\theta_i)}{P(\boldsymbol{X}^0)}$$

For continuous random variables:

$$q(Y|X) = \frac{p(X|Y)\, h(y)}{g(X)} \overset{\text{Bayesian}}{\Rightarrow} p(\theta|\boldsymbol{X}^0) = \frac{p(\boldsymbol{X}^0|\theta)\, p(\theta)}{\int p(\boldsymbol{X}^0|\theta)\, p(\theta) d\theta}$$

- □ $p(\theta|\boldsymbol{X}^0)$ is a p.d.f, the posterior probability density for θ.
- □ $p(\boldsymbol{X}^0|\theta)$ is the likelihood function $L(\theta)$. It is not a p.d.f
- □ $p(\theta)$ is the prior probability density for θ. **Here lies the major problem!**

Example (straight from Wikipedia):

- There are two subspecies of beetle – the “common one” C and the “rare one” R.
- An entomologist spots what might be a rare subspecies of beetle, due to the pattern X on its back.
- In the rare subspecies, 98% have the pattern, or $P(X|R) = 98\%$. In the common subspecies, 5% have the pattern, or $P(X|C)$.
- The rare subspecies accounts for only 0.1% of the population.
- How likely is the beetle having the pattern to be rare, or what is $P(R|X)$?

Example (straight from Wikipedia):

- □ There are two subspecies of beetle – the "common one" C and the "rare one" R.
- □ An entomologist spots what might be a rare subspecies of beetle, due to the pattern X on its back.
- □ In the rare subspecies, 98% have the pattern, or $P(X|R) = 98\%$. In the common subspecies, 5% have the pattern, or $P(X|C)$.
- □ The rare subspecies accounts for only 0.1% of the population.
- □ How likely is the beetle having the pattern to be rare, or what is $P(R|X)$?

$$P(R|X) = \frac{P(X|R)P(R)}{P(R)} = \frac{0.98 \times 0.001}{0.98 \times 0.001 + 0.05 \times 0.999} \simeq 1.9\% \text{ (!!!)}$$

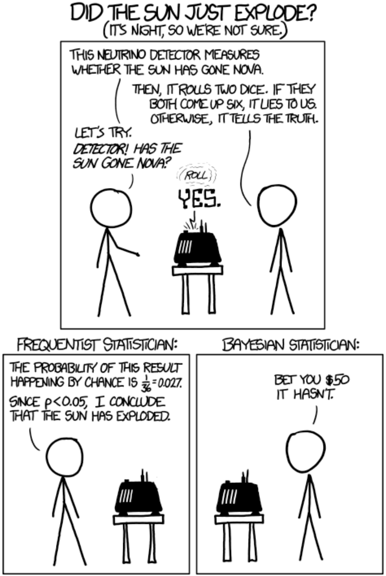
The Prior ProblemS – stolen from M. Pierini (CERN)

□ The need of priors in Bayesian statistics is a problem for some physicist

> The origin of the problem lies in the very first Bayesian assumption, namely that unknown model parameters are to be understood as mathematical objects distributed according to PDFs, which are assumed to be known: the priors. Obviously, the choice of the priors cannot be irrelevant; hence, the Bayesian treatment is doomed to lead to results which depend on the decisions made, necessarily on unscientific basis, by the authors of a given analysis, for the choice of these extraordinary PDFs.
>
> J. Charles et al. hep-ph/0607246

□ The lack of priors in nonBayesian statistics is a problem for some statistician

> The frequentist approach to hypothesis testing does not permit researchers to place probabilities of being correct on the competing hypotheses. This is because of the limitations on mathematical probabilities used by frequentists. For the frequentists, probabilities can only be defined for random variables, and hypotheses are not variables (they are not observables)... This limitation for frequentists is a real drawback because the applied researcher would really like to be able to place a degree of belief on the hypothesis. He or she would like to see how the weight of evidence modifies his/her degree of belief (probability) on the hypothesis being true.
>
> J. Press, *Subjective and Objective Bayesian Statistics*


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

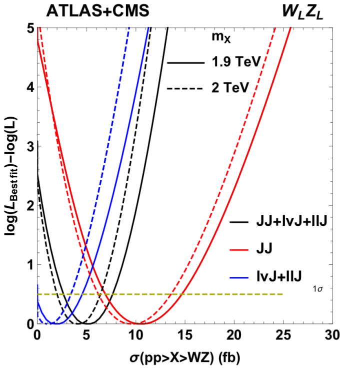
Practical Answers to Statistical Questions

- Point estimation – find a single value $\hat{\theta}$ that is "as close as possible" to the true parameter θ we want to measure. We usually use the maximum likelihood estimator

$$\frac{\partial \ln L}{\partial \theta_i} = 0$$

  which is optimal in the asymptotic limit of large N.
  - But in general it is better to report the likelihood function itself, at least near its maxima.
- Interval estimation – find the range $\theta_a \leq \theta \leq \theta_b$ that contains the true value θ_0 with probability β.
  - 1D: trivial, use the Neyman construction with the Feldman-Cousins "unified approach";
  - ND: use profile likelihood (MINOS). As a bonus, it allows for the removal of nuisance parameters $\boldsymbol{\mu}$ by maximising the full likelihood, at each value of the parameter of interest θ.

Example of Profile Likelihood



The profiled combined likelihoods for ATLAS and CMS Run 1 diboson resonance searches. The best-fit cross-section for the $\mathrm{W}' \to \mathrm{W}_L \mathrm{Z}_L$ with a W' mass of 1.9 TeV was $\sigma = 5.3^{+2.3}_{-2.0}\,\mathrm{fb}$.

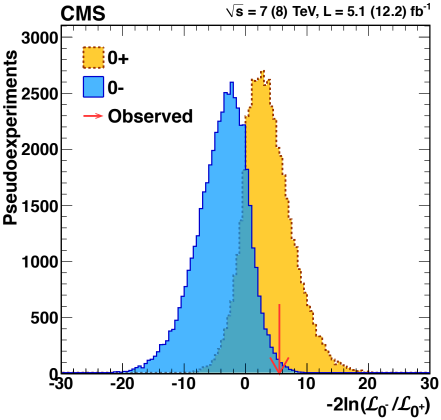
Hypothesis Testing – from Bob Cousins

Consider two hypotheses:

- □ H_0: is the null hypothesis. For instance, "the Standard Model is a true description of nature at the scales probed by the LHC".
- □ H_1: is the alternative hypothesis. For instance, "$\langle$INSERT YOUR MODEL HERE$\rangle$ is a true description of nature at the scales probed by the LHC".

$L(X, \theta)$ is different for H_0 and H_1. How do we test the two hypotheses against each other?

- For the null hypothesis H0, order possible observations x from least extreme to most extreme, using an ordering principle (which can depend on H1 as well). Choose a cutoff α (smallish number).
- □ "Reject" H_0 if the observed x_0 is in the most extreme fraction α of observations x (generated under H_0). By construction:
  - $\alpha =$ probability (with x generated according to H_0) of rejecting H_0 when it is true;
  - $\beta =$ probability (with x generated according to H_1) of not rejecting H_0 when it is false.



Expected distribution of $2\ln(L(0^-)/L(0^+))$ under the pure pseudoscalar and pure scalar hypotheses (histograms) for the Higgs boson. The arrow indicates the value determined from the CMS observed data with the discovery dataset (7 TeV, 5.1 fb^{-1} and 8 TeV, 12.2 fb^{-1}).

A decision on whether or not to declare discovery (falsifying H_0) requires 2 more inputs, both of which can affect the choice of α:

□ **Prior** belief in H_0 vs H_1.

□ Cost of Type I error (false discovery claim) vs cost of Type II error (missed discovery).

A one-size-fits-all criterion of α corresponding to 5σ is without foundation!

...and still, I (Thiago) am positive it will still be used for the rest of the days of the LHC.

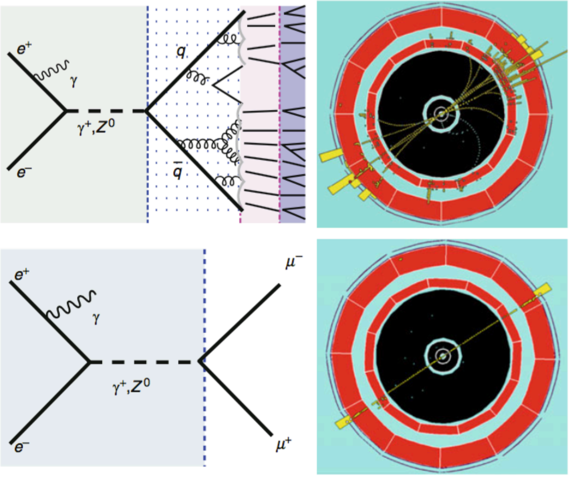
Example: Ratio of Hadronic & Leptonic Cross Sections in e^+e^- Reactions

□ For energy range where five quark flavours contribute and below the Z resonance (for lowest order in perturbation theory)

$$R_\gamma = \frac{\sigma(e^+e^- \to \text{hadrons})}{\sigma(e^+e^- \to \mu^+\mu^-)} = \frac{\sigma_{\text{had}}}{\sigma_{\text{lep}}}$$

$$= N_c \sum_q e_q^2 = N_c \frac{11}{9}$$

□ Goal: determine or constrain the number of colour states (N_c)



□ Count the number of events with hadronic (N_{had}) and leptonic (N_{lep}) final states

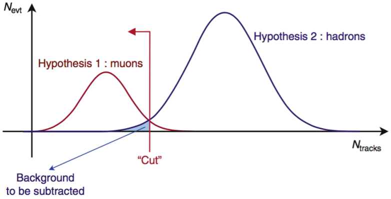
$$R_\gamma = \frac{\sigma_{\text{had}}}{\sigma_{\text{lep}}} = \frac{N_{\text{had}}/\mathcal{L}}{N_{\text{lep}}/\mathcal{L}} = \frac{N_{\text{had}}}{N_{\text{lep}}}$$

□ Define event selection and estimate backgrounds, possibly with input from simulation to define discrimination variables.

- In this case, the number of charged particles (N_{tracks}) in each event helps to separate hadronic from leptonic events. Usually leptonic events have few tracks, whilst hadronic events have many more.
- But there is some overlap, so the selection has an efficiency ($\epsilon \leq 1$) to select a given type of event.

$$N^{meas}_{\text{had/lep}} = \epsilon^{meas}_{\text{had/lep}} \cdot N^{true}_{\text{had/lep}}$$

- $N^{meas}_{\text{had/lep}}$ should be corrected. In practice, what we do is estimate the efficiencies (using either simulation or data).



□ Analysis

- In general efficiency is not only due to the cut, but also includes: $\epsilon = A \cdot \epsilon_{\text{tri}} \cdot \epsilon_{\text{rec}} \cdot \epsilon_{\text{cut}}$
  - A: detector acceptance
  - ϵ_{tri}, ϵ_{rec}, ϵ_{cut}: trigger, reconstruction, and cut efficiencies
- Background subtraction

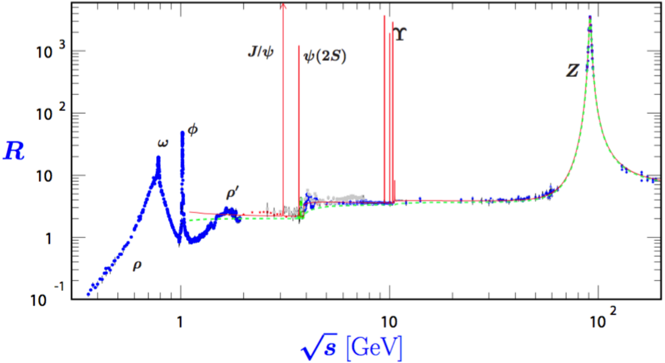
$$N_{\text{lep}}^{true} = (N_{\text{lep}}^{meas} - N_{\text{bckg}})/\epsilon_{\text{lep}}$$

- Statistical uncertainties
  - Can be estimated considering the statistical distributions followed in each measurement.
  - For counting experiment a Poisson distribution can be a proper choice.
  - For efficiency measurement (pass or fail), an uncertainty following a binomial would be preferable.
- Systematic uncertainties
  - In general, it will depend on each analysis (no standard procedure).
  - E.g. in the present case we could assign an uncertainty associated to the mismodelling of detector response by the simulation, which was used for efficiency correction.

Ratio of Hadronic & Leptonic Cross Sections in e^+e^- Reactions: Result

□ Measurements of R from different lepton colliders

- The relation for R_γ in previous slides applies to region $\sqrt{s} > 10\ \mathrm{GeV}$ (where quarks u,d,s,c, and b contribute) and far from the Z boson peak.
- In this case we have $R_\gamma = N_c \frac{11}{9} \approx \frac{11}{3} \Rightarrow N_c \approx 3$.



Tools of the Trade

ROOT is a modular scientific software toolkit. It provides all the functionalities needed to deal with big data processing, statistical analysis, visualisation and storage. It is currently used by all the LHC experiments.

SciPy is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, some of the core packages are NumPy, SciPy library, Matplotlib, IPython, SymPy, pandas.

Conclusions

High-energy physics is the field that studies the smallest building blocks of matter.

It is equally powered by contributions from theorists, experimentalists, computer scientists, engineers...The harmonious cooperation of those different groups is vital to the success of the field.

From the theoretical side, the field has had continued, resounding success with the standard model of particles and fields. Extensions to the standard model continue be proposed, exploring new ideas and addressing additional data produced by other fields, like astronomy and cosmology.

From the experimental side, the field has moved to global collaborations that design, build and operate extremely large and complex detectors. The data taken with those detectors dwarfs all other scientific datasets to date, and allows to measure the properties of the particles and fields to unprecedented precision.

High-energy physics is a long term endeavour, with experiment time scales measured in decades. The field is already preparing for the challenges aheads, with new experiments being proposed all around the world. Finally, the LHC is scheduled to run at least until 2035.

SPRACE

Thanks

And...

SPRACE

We need YOU!!! Please join us at `https://sprace.org.br`