Statistical Methods Lecture 2

Claudio Campagnari
UCSB

U.S. DEPARTMENT OF ENERGY | Office of Science

UCSB

Welcome to part 2 ☺

I want to start today’s lecture by going back to yesterday’s slides on profile likelihood to emphasize a point that I sort of skipped

Profile likelihood

- Multi-dimensional likelihood $\mathcal{L}(\vec{a}) = \mathcal{L}(\overrightarrow{a_1}, \overrightarrow{a_2})$ where we want to focus on the subset of parameters $\overrightarrow{a_1}$
- Profile likelihood ratio $\lambda(\overrightarrow{a_1}) = \frac{\mathcal{L}(\overrightarrow{a_1}, \widehat{\widehat{\overrightarrow{a_2}}})}{\mathcal{L}(\widehat{\overrightarrow{a_1}}, \widehat{\overrightarrow{a_2}})}$
  - The denominator, with single hats, gives the maximum of the likelihood
  - The numerator is a function of $\overrightarrow{a_1}$ and at each value of $\overrightarrow{a_1}$ the likelihood is maximized with respect to $\overrightarrow{a_2}$ (hence the weird double hat notation).
- Wilk's theorem: $-2\ \log\lambda(\overrightarrow{a_1})$ is asymptotically distributed as χ^2 with N_1 degrees of freedom. (N_1=dimensionality of $\overrightarrow{a_1}$)
- A **profile** ΔNLL scan gives a function with the same properties as described earlier with respect to $\overrightarrow{a_1}$ only

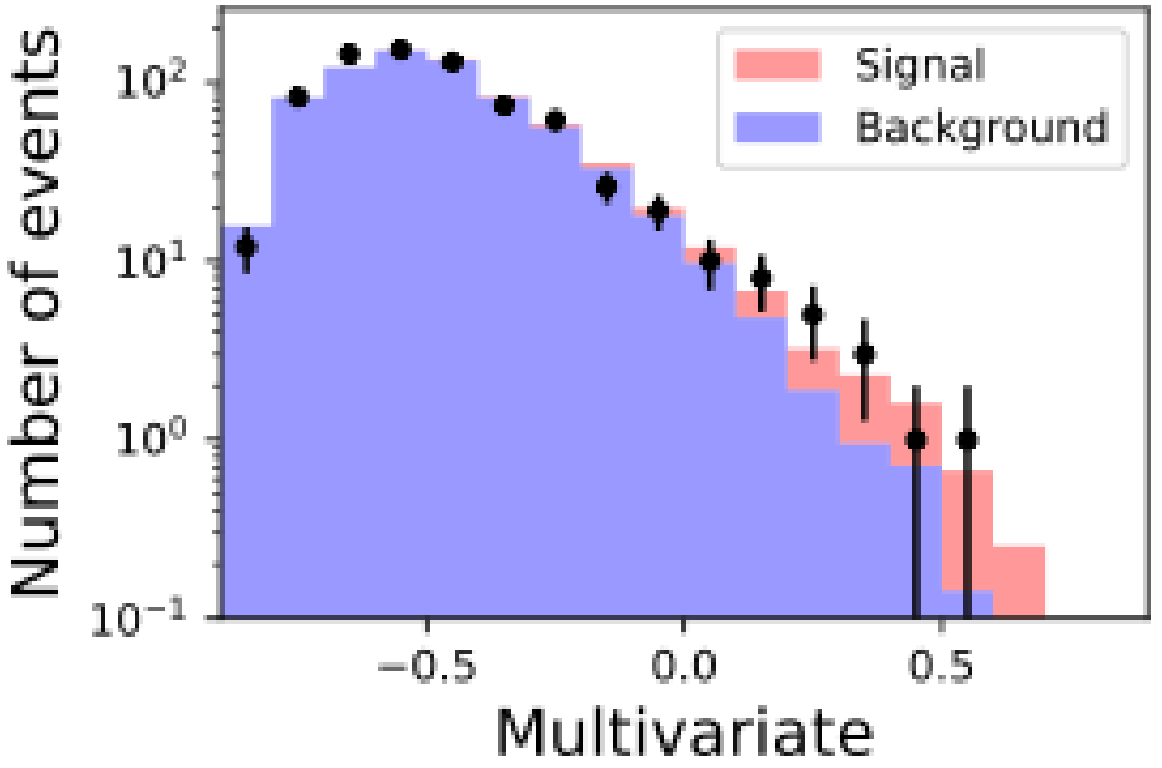
Profile likelihood (continued)

- The common use in frequentist statistics is to “profile away” the nuisances (ie: $\overrightarrow{a_2}$ in the previous page would be the set of nuisances)
- This means that the nuisances are fitted together with the quantity of interest.
  - The knowledge of the nuisances (systematics) is then improved in the analysis.
    - **Whether you like it or not**

Fitting a distribution to several components

- For example, Signal (S) and Background (B). But maybe more.
  - One dimensional here

- Two different questions
  1. You care about the numbers S and B in this sample (sample composition)
     - N=S+B events is fixed
  2. You care about the "primordial" values of S (and maybe B) in an equivalent sample of given luminosity or p.o.t.
     - N = S+B events is not fixed
     - S is used here to estimate some physics quantity, eg, a cross-section



First case, S+B=N fixed

Unbinned

Probability of a given event with value x_i

$$p_i = \frac{S}{S+B}\, p_S(x_i) + \frac{B}{S+B}\, p_B(x_i)$$

$$\mathcal{L} = \prod p_i$$

$$\mathrm{NLL} = -\sum \log\left(S\, p_S(x_i) + B\, p_B(x_i)\right) + \log N$$

This is a constant, so we can drop it

$$\boxed{\mathrm{NLL} = -\sum \log\left(S\, p_S(x_i) + B\, p_B(x_i)\right) = -\sum \log\left(S\, p_S(x_i) + (N-S)\, p_B(x_i)\right)}$$

NLL is only function of one quantity

First case, S+B=N fixed

Unbinned

Probability of a given event with value x_i

$$p_i = \frac{S}{S+B}\, p_S(x_i) + \frac{B}{S+B}\, p_B(x_i)$$

$$\mathcal{L} = \prod p_i$$

$$\mathrm{NLL} = -\sum \log\left(S\, p_S(x_i) + B\, p_B(x_i)\right) + \log N$$

$$\mathrm{NLL} = -\sum \log\left(S\, p_S(x_i) + B\, p_B(x_i)\right) = -\sum \log\left(S\, p_S(x_i) + (N-S)\, p_B(x_i)\right)$$

Binned

Probability of a given bin with entries d_i

$$p_i = e^{-\mu_i} \frac{\mu_i^{d_i}}{d_i!} \qquad \mu_i = S p_{iS} + B p_{iB}$$

$$\mathcal{L} = \prod p_i$$

$$\mathrm{NLL} = -\sum d_i \log\left(S\, p_{iS} + B\, p_{iB}\right) = -\sum d_i \log\left(S\, p_{iS} + (N-S)\, p_{iB}\right)$$

NLL is only function of one quantity

Second Case, S+B=N not fixed

Now *N* is **<u>also</u>** a random variable

$$p_N = e^{-(S+B)} \frac{(S+B)^N}{N!}$$

$$\mathcal{L} = p_N \prod p_i$$

Leads to

$$\mathrm{NLL} = S + B - \sum \log \left(S\, p_S(x_i) + B\, p_B(x_i) \right)$$

Unbinned

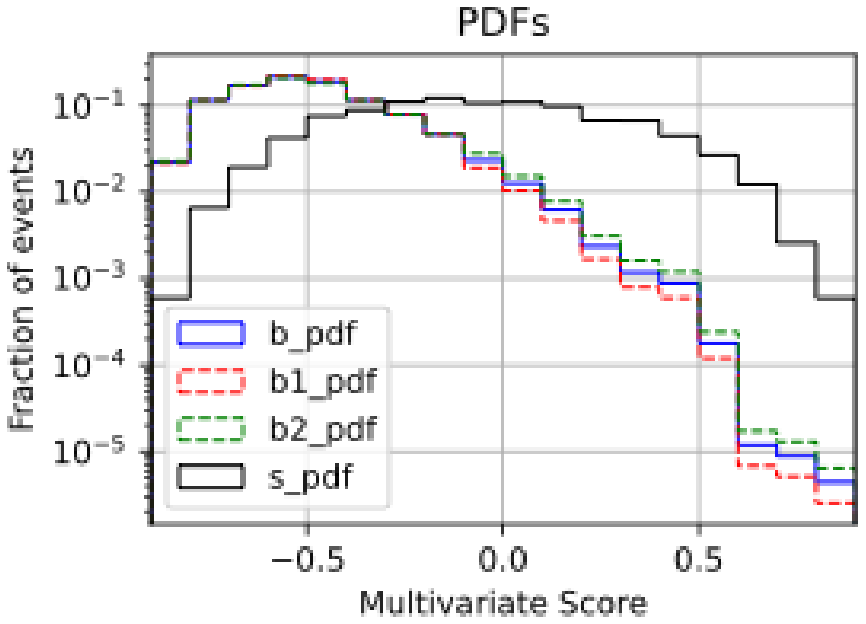
$$\mathrm{NLL} = S + B - \sum d_i \log \left(S\, p_{iS} + B\, p_{iB} \right)$$

Binned

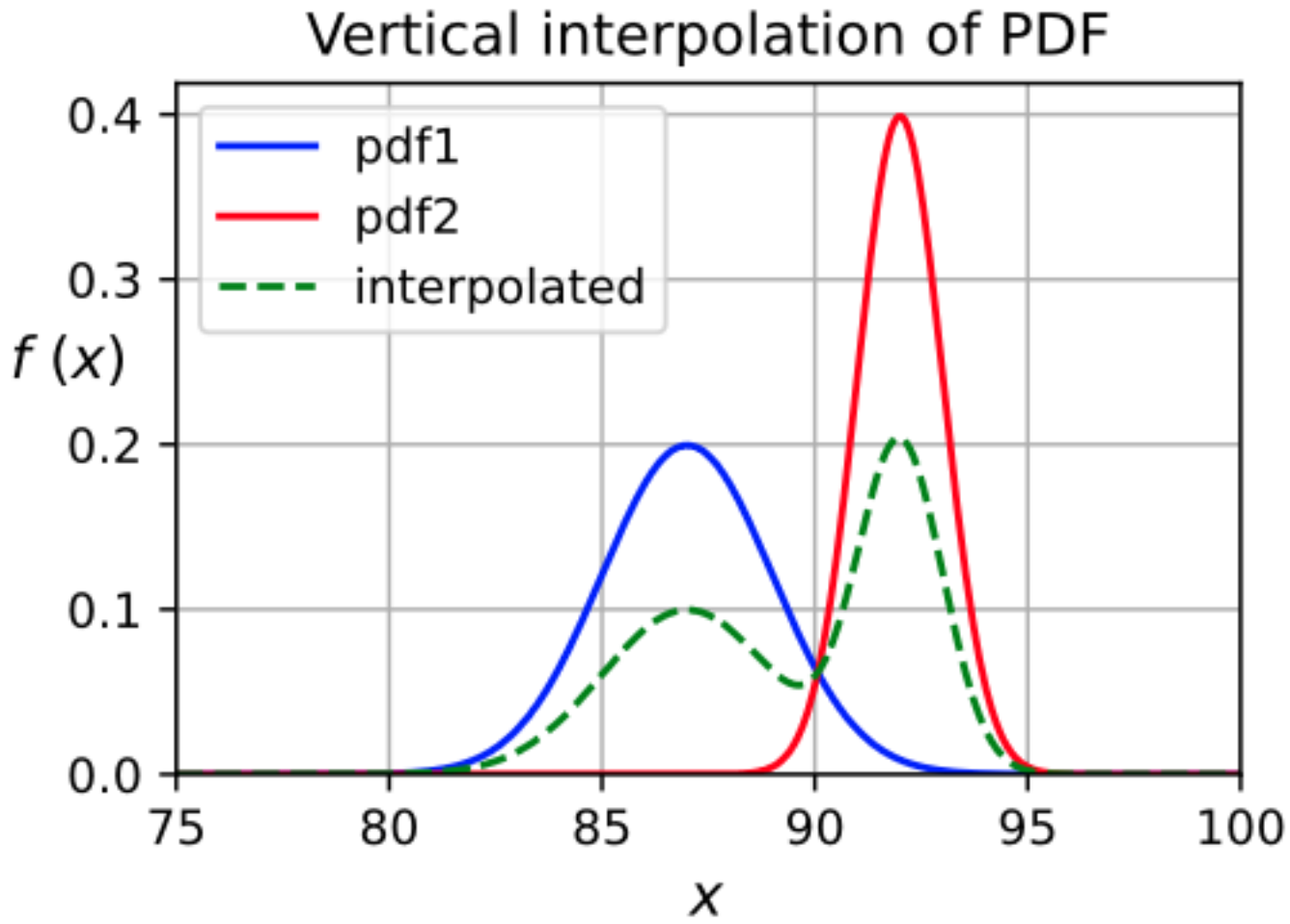
Extended log-likelihood

Fitting a distribution: pdf shapes not perfectly known

1. If the pdfs can be expressed analytically solution straight-forward
   - The parameters of the analytic functions are nuisance parameters that come with their own pdfs
   - In a Bayesian analysis: marginalize them away
   - In a Frequentist analysis: profile them away
2. Sometime the pdf's are not analytic, they might come (as histograms?) from some study with a $\pm 1\sigma$ variation.
   - "Morphing": Introduce a nuisance parameters α with Normal pdf which interpolates/extrapolates such that
     - $\alpha = 0$ gives central value of pdf
     - $\alpha = \pm 1$ gives $\pm 1\sigma$ variation of pdf
     - **"Vertical Morphing"**

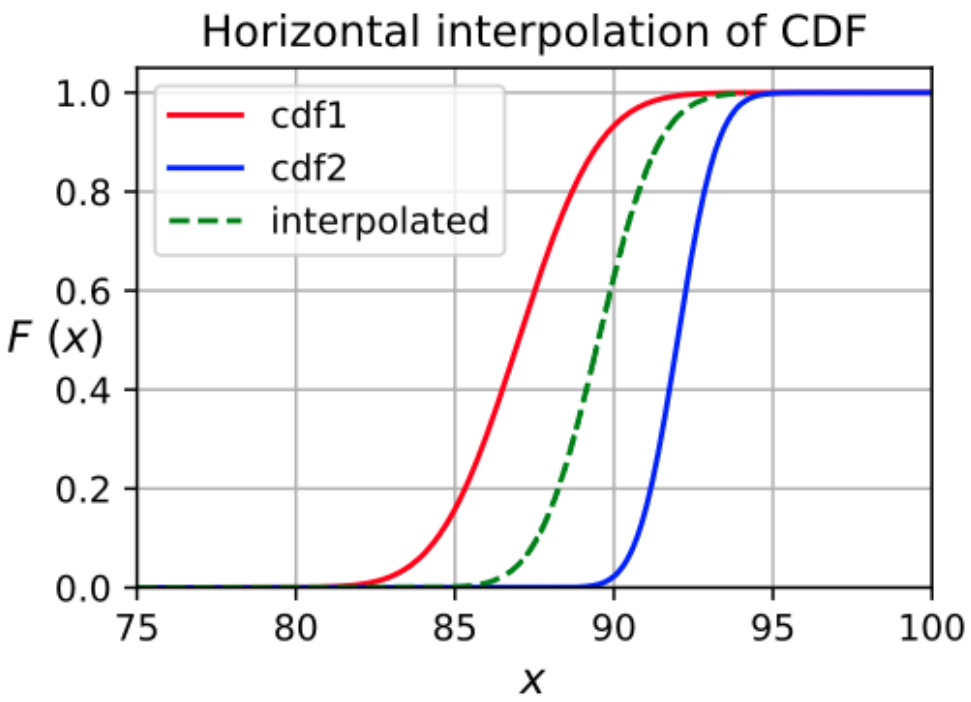


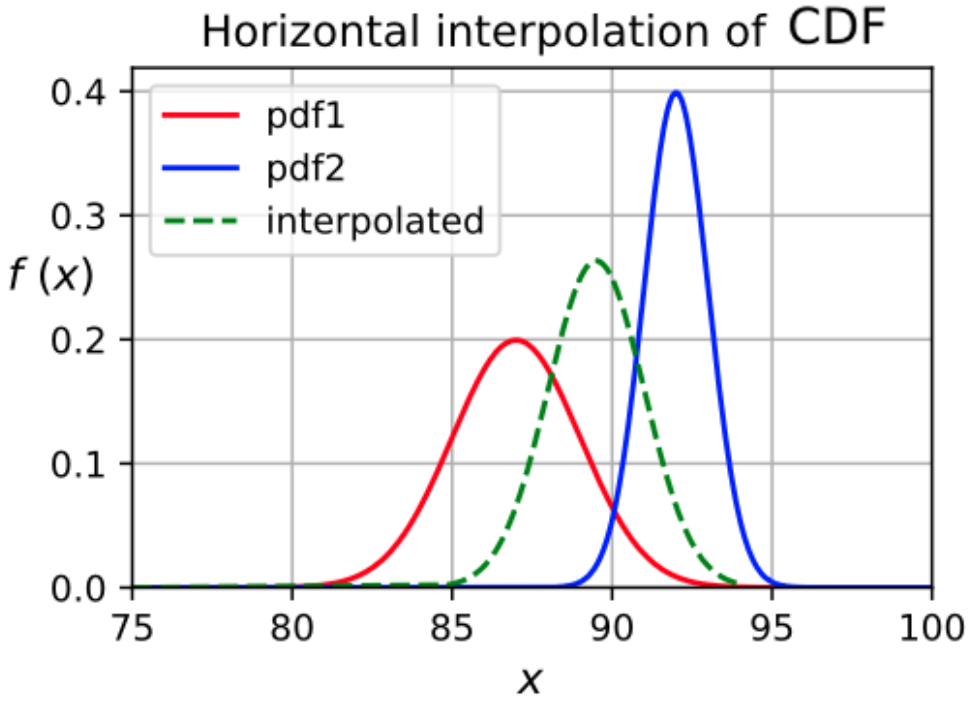
Careful about vertical morphing!



This is most likely NOT what you intended when you tried to interpolate between pdf1 and pdf2!!!

Sometime ”horizontal” morphing is better





- More computationally expensive.
- Not suited for multidimensions.
- More complicated methods, “moments based”, exist.
  - https://arxiv.org/pdf/1410.7388

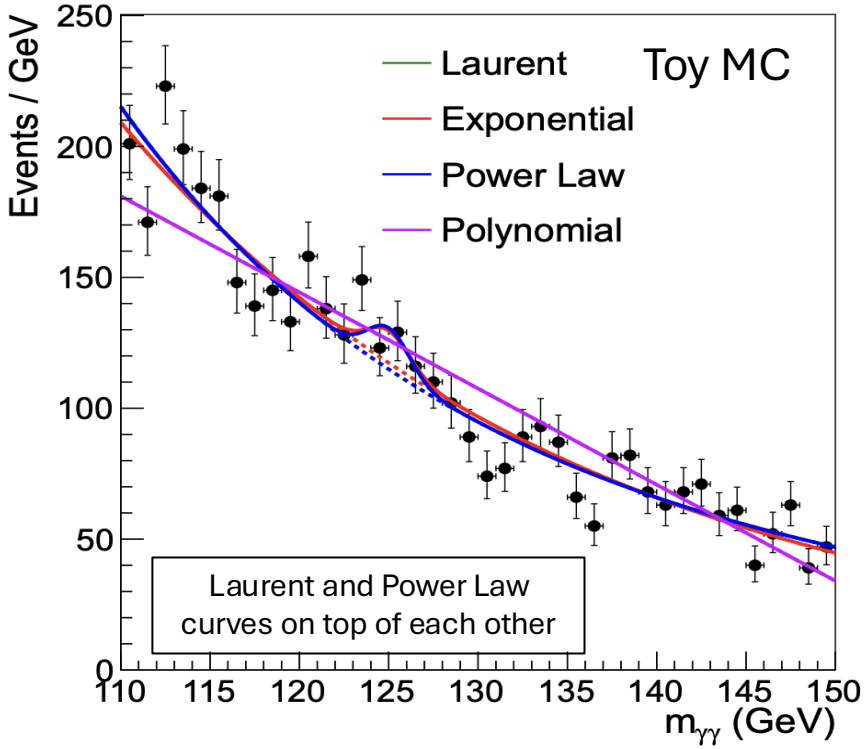
Discrete Profiling (Envelope Method)

What is it for?

- You do not know the functional form from 1st principles
- You do not trust MC to predict the shape

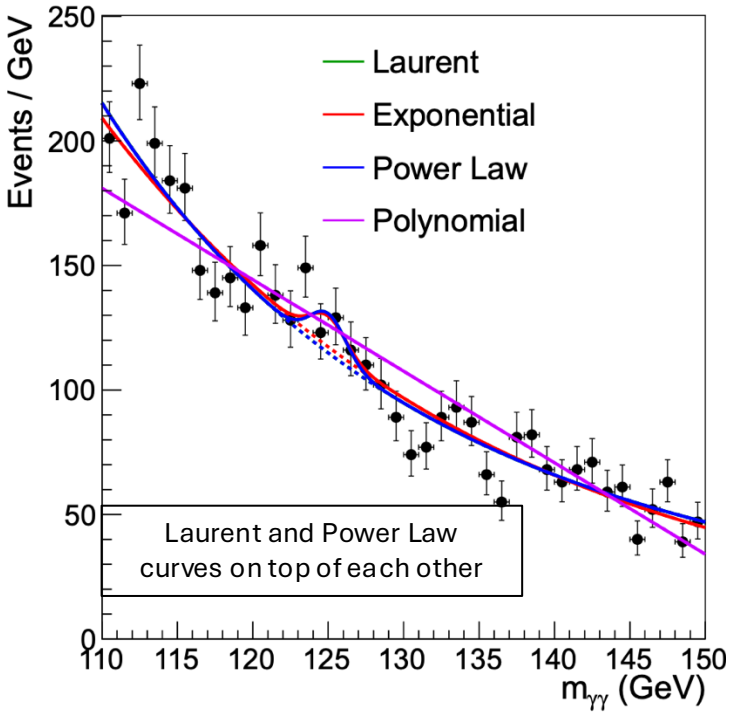
What's the solution?

- You try different analytic functions
- The choice of function is a nuisance parameter
  - But it is a **discrete** (not continuous) nuisance parameter
  - No pdf associated with the choice
  - **Standard Profiling of the likelihood does not work**
- **Now what?**

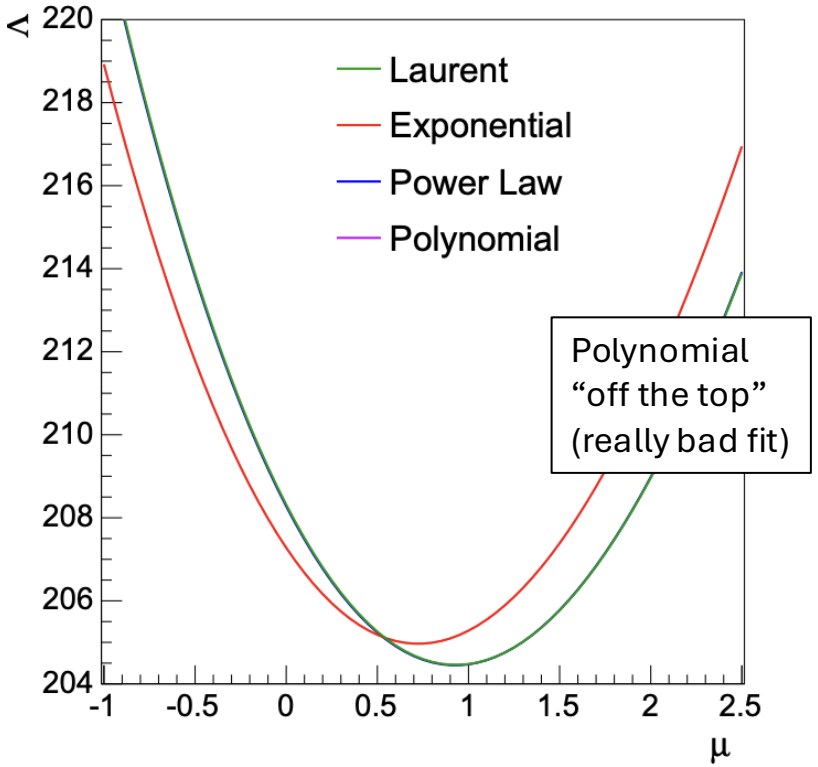


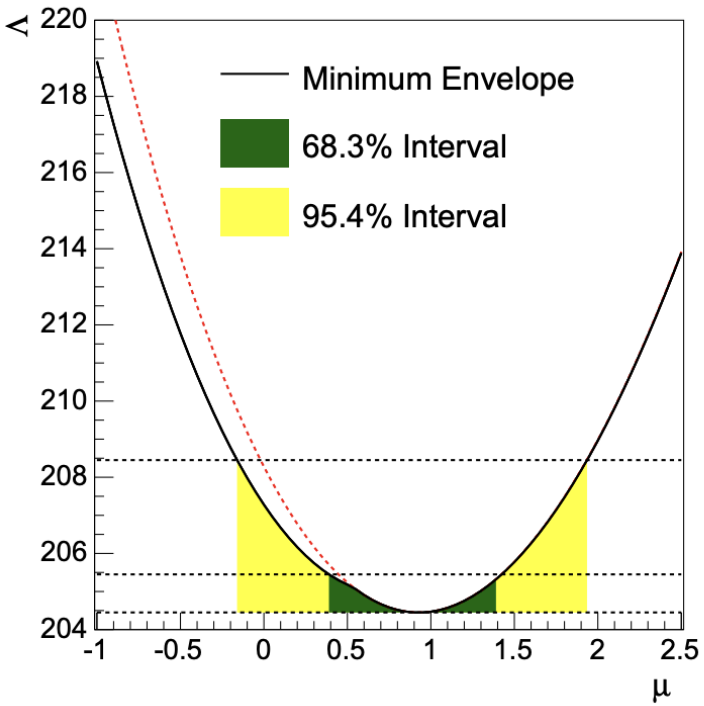
1. "Power law"; $f(x) = p_0 x^{p_1}$.
2. "Exponential"; $f(x) = p_0 e^{p_1 x}$.
3. "Laurent"; $f(x) = p_0/x^4 + p_1/x^5$.
4. "Polynomial"; $f(x) = p_0 + p_1 x$.

Dauncey et al.



NLL curves for different choices





Take "envelope" of green and red curves

What if you have templated histogram pdfs from poor MC statistics?

- **Best to avoid it.**
- Want MC statistics >> Data statistics (factor 10 maybe?)
- To formally include MC statistics, each bin in each template has a nuisance with its own pdf
- Looks intractable!
- Barlow-Beeston showed that it can be done with a clever treatment of these nuisances which are handled inside the NLL calculation.
- John Conway simplified this further (Barlow-Beeston lite) also to avoid numerical issues
- Dembinski et al. recently came up with a refined proposal
- Barlow-Beeston (or Barlow-Beeston-lite?) available in RooFit

Constraining Parameter(s) in a Fit

- Minuit allows you to put numerical bounds on fit parameters
- Might want it to impose "physics", e.g., a cross-section cannot be negative
  - **But if your fit prefers an unphysical result, wouldn't you want to know?**
- The Minuit manual warns against numerical bounds
  - The covariance matrices for parameters near a boundary are unreliable
  - Internally Minuit remaps the parameter into a $\sin^{-1}$ function of the parameter which can cause numerical instabilities
- OTOH, sometime allowing Minuit to wander into unphysical regions can cause attempts to take logarithms of negative numbers
  - Think about taking *S* or *B* very negative in the expression below

$$\mathrm{NLL} = S + B - \sum d_i \log\left(S\, p_{iS} + B\, p_{iB}\right)$$

From the Minuit manual

1.2.1 The transformation for parameters with limits.

For variable parameters with limits, Minuit uses the following transformation:

$$P_{\text{int}} = \arcsin\left(2\,\frac{P_{\text{ext}} - a}{b - a} - 1\right) \qquad P_{\text{ext}} = a + \frac{b - a}{2}\left(\sin P_{\text{int}} + 1\right)$$

so that the internal value P_{int} can take on any value, while the external value P_{ext} can take on values only between the lower limit a and the upper limit b. Since the transformation is necessarily non-linear, it would transform a nice linear problem into a nasty non-linear one, which is the reason why limits should be avoided if not necessary. In addition, the transformation does require some computer time, so it slows down the computation a little bit, and more importantly, it introduces additional numerical inaccuracy into the problem in addition to what is introduced in the numerical calculation of the `FCN` value. The effects of non-linearity and numerical roundoff both become more important as the external value gets closer to one of the limits (expressed as the distance to nearest limit divided by distance between limits). The user must therefore be aware of the fact that, for example, if he puts limits of $(0, 10^{10})$ on a parameter, then the values 0.0 and 1.0 will be indistinguishable to the accuracy of most machines.

Constraining Parameter(s) in a Fit

Minuit does not allow to impose constraints among parameters

- For example: needed to simultaneously fitting two tracks to come from the same point (vertex constraint)
  - Express constraint as $f(\vec{a}) = 0$

Three solutions

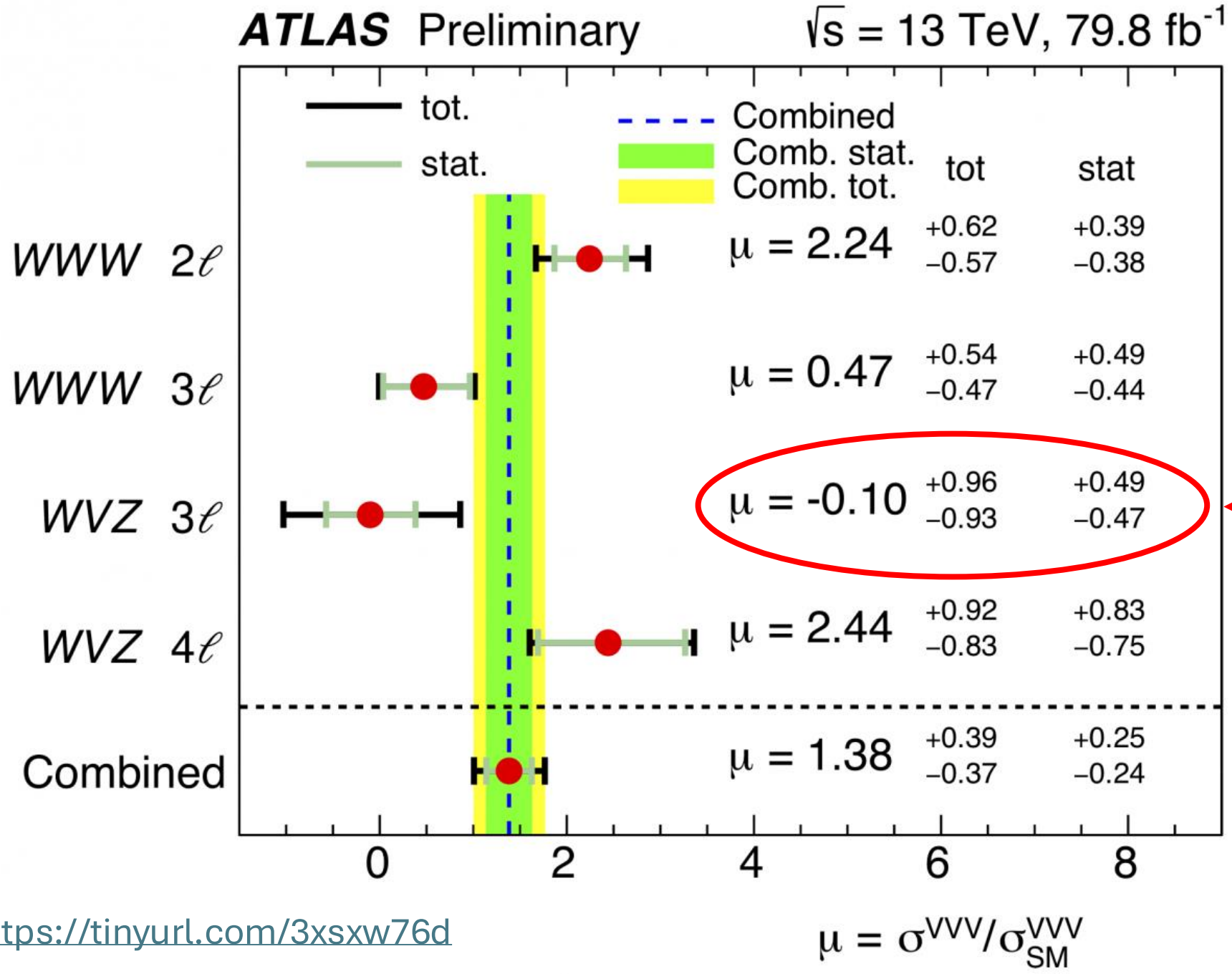
1. Solve each constraint equation to eliminate an a_i from the set $\vec{a}$
2. Approximate by adding a penalty term $\chi^2 \rightarrow \chi^2 + \frac{f^2(\vec{a})}{s^2}$.
   - Choose s wisely
3. Lagrange multiplier method $\chi^2 \rightarrow \chi^2 + 2\lambda f(\vec{a})$
   - Minimize with respect to $\vec{a}$ **and** λ
   - $\frac{\partial \chi^2}{\partial \lambda} = 0 \rightarrow f(\vec{a}) = 0$
   - **Does not work with Minuit. Only with linear algebra techniques**

Physical Region

Should you build into your analysis physical region requirements on the thing that you are trying to measure?

- e.g. require cross-section > 0 in your fitting procedure

- **My strong opinion: NO**
- If your answer is $\sigma = -100 \pm 10$ nb…..your analysis is broken!
  - If you were to impose σ>0 requirement, your fit might give you $\sigma = 0^{+10}_{-0}$nb and you would never know that you have a big problem!
- OTOH $\sigma = -5 \pm 10$ nb is a perfectly fine result (at least in principle), You can/should report it in the paper.
  - How to interpret it is a different story
- Sometime reporting results outside the physical region is **mandatory**



Outside physical region, but needed for unbiased combination!

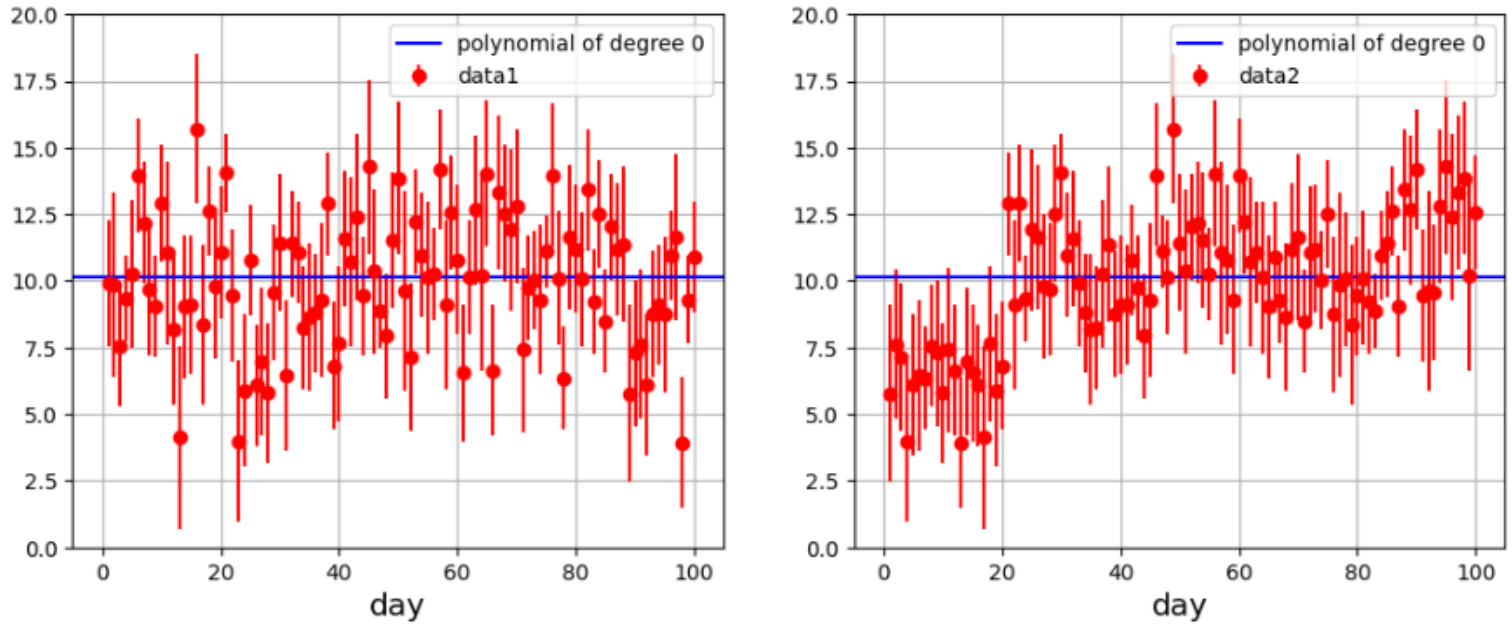
https://tinyurl.com/3xsxw76d

Goodness of fit (GOF)

- How can you quantify if your fit is “good”?
- Sadly, there is no unique way
- Individual judgement is probably best.
  - But your colleagues will question it and want a number 😂

χ^2 is only for binned, assumes Gaussian uncertainties, and is not sensitive to shapes

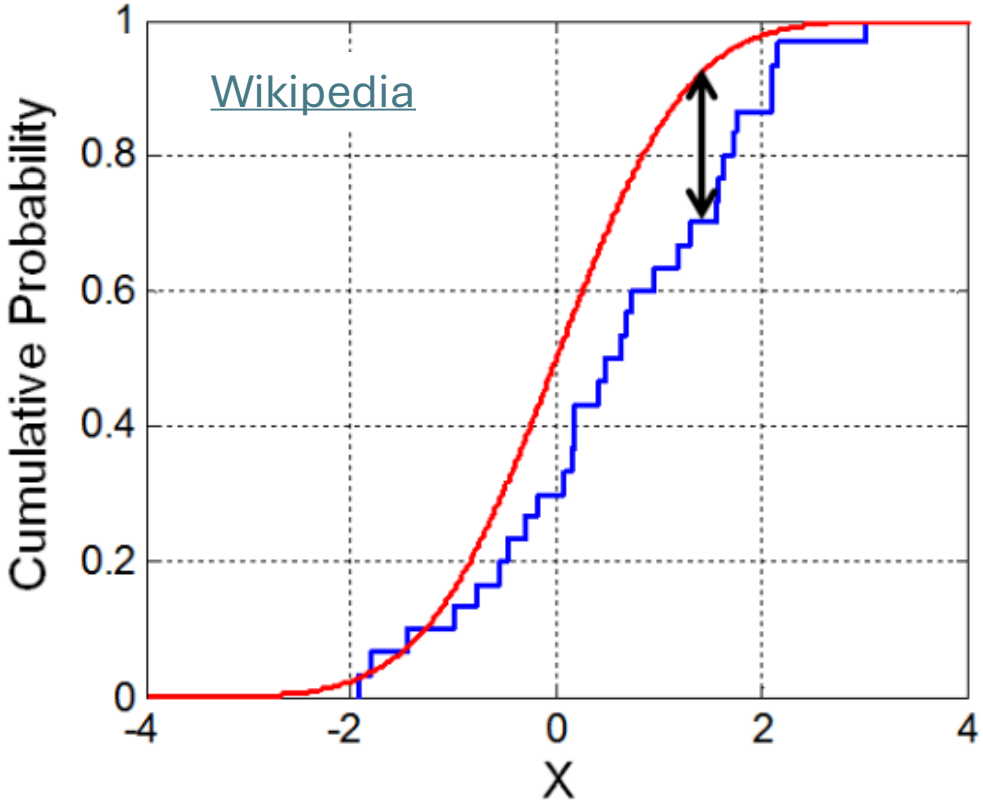
$$\chi^2 = \sum \frac{(d_i - \mu_i)^2}{\sigma_i^2}$$



Two (fake) datasets, with exactly the same χ^2 to a straight line fit.

Saturated model as GOF test is based on ratio of likelihoods and is ~ extension of χ^2 in Poisson regime

Kolmogorov-Smirnov (KS)test is sensitive to shape, but (mostly) only for 1D



Anderson-Darling and Cramer-von Mises are other variants of CDF-based tests

There are ad-hoc 2D implementations, including one in root, but they make assumptions

What definitely not to do…..

Compare the data NLL at its minimum with expectations from MC ensemble experiments

- Easy to see why with an example
- Earlier we wrote down the NLL for a lifetime experiment $p(t_i|\tau) = \frac{1}{\tau} e^{-t_i/\tau}$

$$\mathrm{NLL} = \frac{1}{\tau}\sum t_i + N\log\tau$$

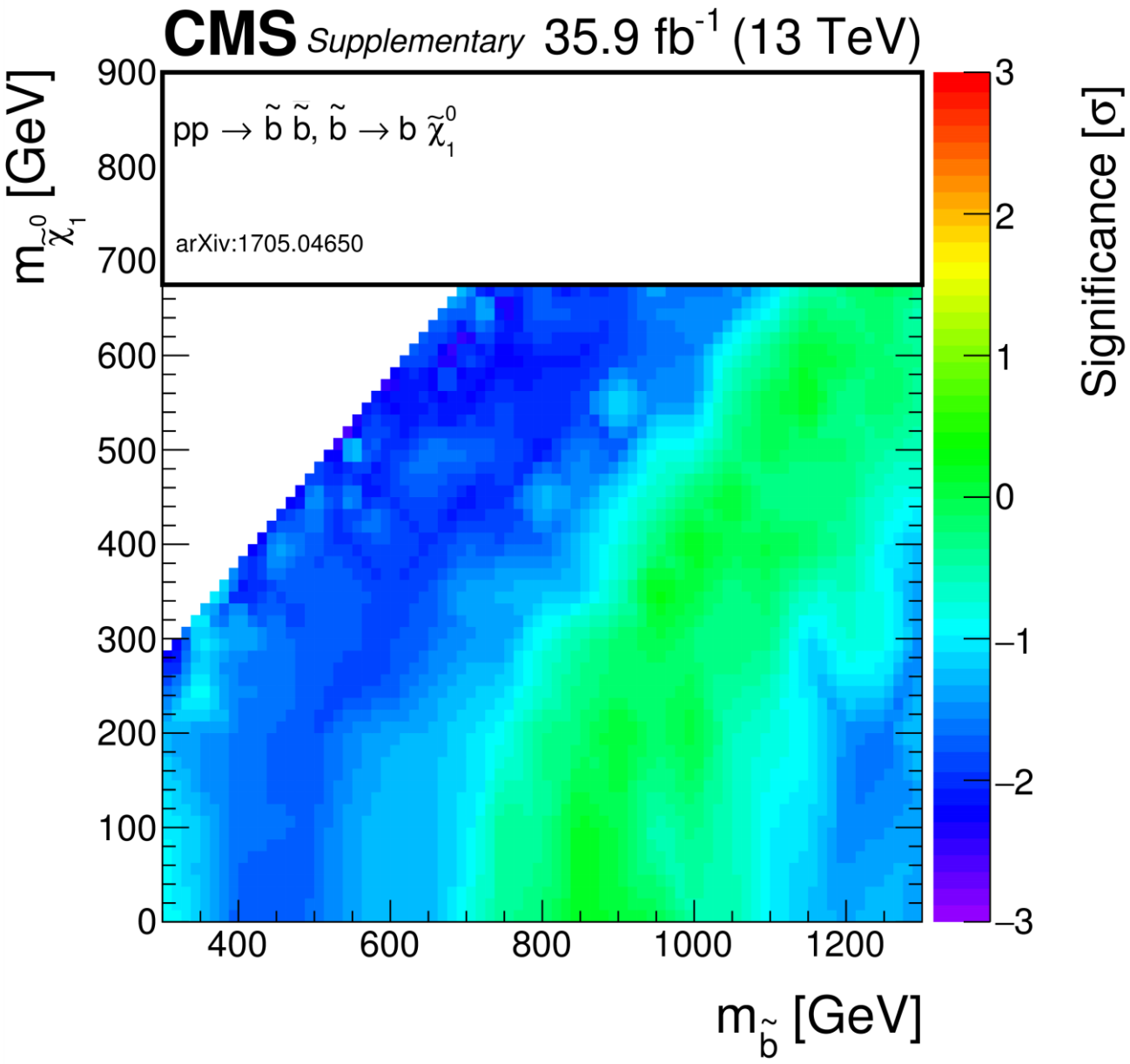
$$\hat{\tau} = \frac{\sum t_i}{N} = \bar{t}$$

$$\mathrm{NLL}_{\mathrm{min}} = N(1+\log\hat{\tau}) = N(1+\log\bar{t})$$

Any dataset with the same average *t* will yield the same minimum of NLL regardless of whether it is distributed exponentially or not

Test Statistics

- Back to single bin counting experiment
- Expect B background events from SM, see N
- p-value for background-only = prob of seeing anything as extreme or more, ie, prob of seeing $\geq N$ events when B are expected
  - N here is the "test statistics"
- Multiple bins $\{B_1, B_2, B_3, ...\}$ and $\{N_1, N_2, N_3, ...\}$. How to define "more extreme"?
  - Not unique
- Define test statistics $t = f(\vec{B}, \vec{N})$ that maximizes discrimination for a possible signal
- **The p-value (and significance) then depends on the signal assumption**
- This is a classic problem of hypothesis testing, H_0 vs. H_1
  - H_0 = background-only hypothesis
  - H_1 = signal+background hypothesis



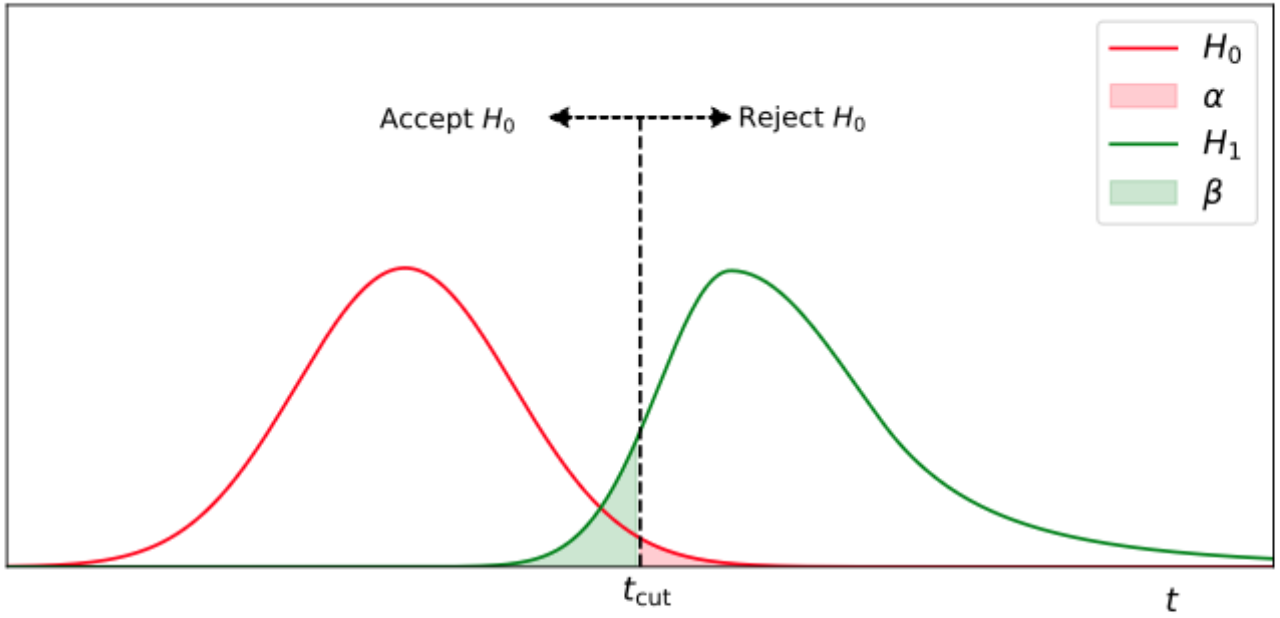
This analysis had > 100 bins

At each (quantized) point in the 2D plane the S+B vs B-only hypotheses were tested with a different test statistic

Because the kinematics of S and how it populates the bins changes in the plane

From each test statistic a p-value and a significance were calculated
(as well as a limit on the cross-section, which is not shown in this plot)

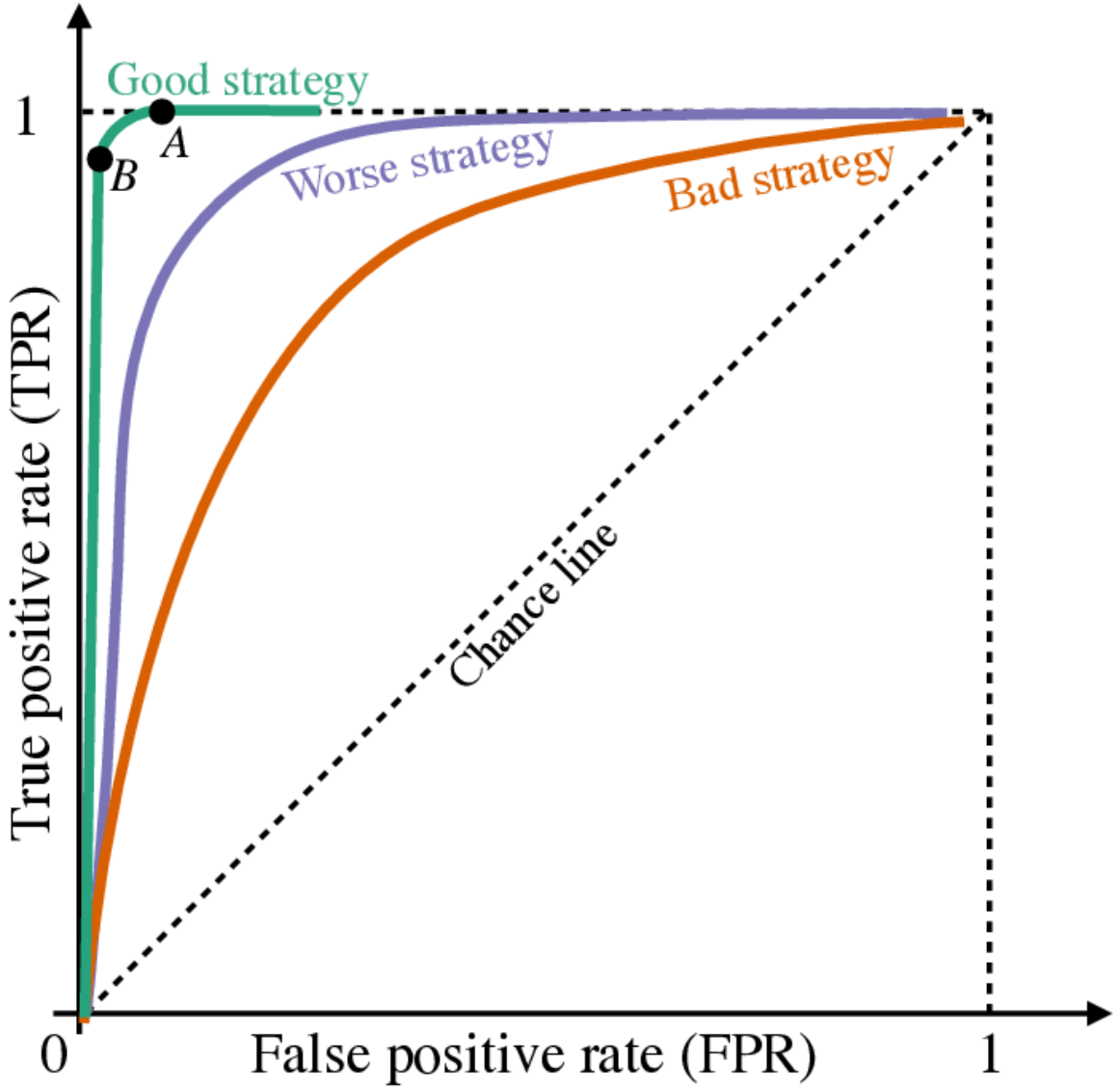
Hypothesis Testing



A sketch of the PDFs of the test statistics t for the two hypotheses, with a cut that is used to make the decision. The color-coded shaded areas have probabilties α or β for the H_0 and H_1 hypothesis, respectively.

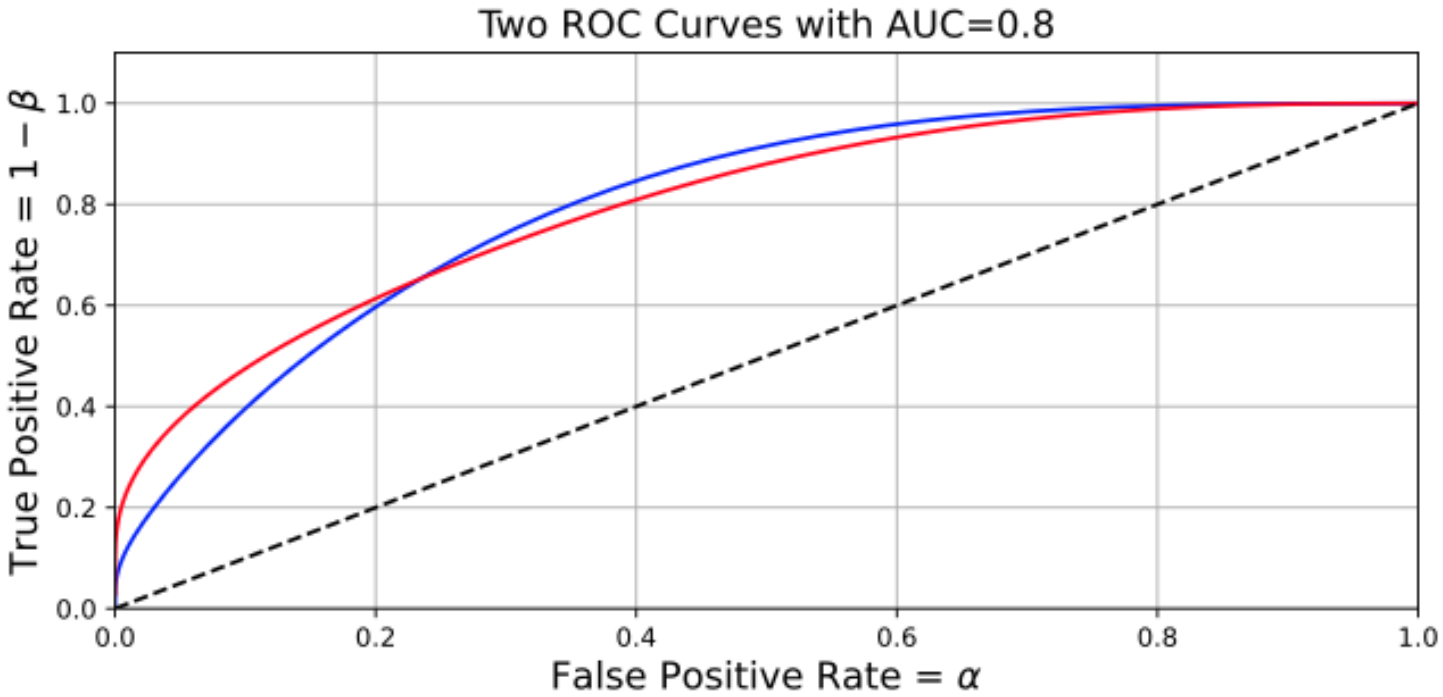
- False Positive Rate (FPR) = α
- True Positive Rate (TPR) = $1-\beta$

Hypothesis Testing: ROC curve and AUC



- Receiver Operating Characteristic (ROC) curve is FPR vs. TPR as the cut on the test statistics is changed
- Area Under the Curve (AUC) is 1 for perfect separation, 0.5 for “coin toss”
- Value of AUC is often used to characterize how good the hypothesis testing is

AUC does not tell the full story



Sample ROC curves for two toy discriminants.

In practice you pick an operating point, e.g, to separate electrons from π^0.
Depending on whether you care more about TPR or FPR the red curve may be better or worse than the blue one.

Keep track of what you care about, do not blindly only look at AUC

Hypothesis Testing: Jargon

- Type I error: reject H_0 when it is correct

- Type II error: accept H_0 when it is not

What is the “best” test statistics

- Best means best TPR at fixed FPR
- The best test statistics is the likelihood ratio

$$t(\vec{x}) = \frac{p(\vec{x}|H_1)}{p(\vec{x}|H_0)}$$

Figure of Merit (FOM)

- How do you optimize your analysis selection
- Carrying out the full analysis in MC for many different selections is often impractical
- Optimize an FOM based on *S* and *B* expectations in your signal region
- In Gaussian regime $S/\sqrt{B}$ is a FOM for discovery or limit setting mode
  - Expect to see *N*=*S*+*B* events.
  - Excess over background is *S*=*N*-*B*.
  - Background fluctuates as $\sqrt{B}$
  - Z-score is $S/\sqrt{B}$
- In Poisson regime equivalent FOM is $\sqrt{2[(S+B)\log\left(1+\frac{S}{B}\right)-S]}$

Figure of Merit (FOM) (continued)

- In Gaussian regime $S/\sqrt{S+B}$ is a good FOM when S is well established and we want to make measurements of its properties
  - Measurements will be made on *N*-*B*=*S* with stat uncertainty $\sqrt{S+B}$
  - So statistical power is $S/\sqrt{S+B}$
- Include syst. uncertainties on BG with $S/\sqrt{B+\sigma_B^2}$ and $S/\sqrt{S+B+\sigma_B^2}$
- G. Punzi suggested FOM= $\frac{\varepsilon}{\frac{n}{2}+\sqrt{B}}$ where ε is signal efficiency and *n* is the number of sigmas that is targeted for discovery.

What if you are looking for a signal whose rate you cannot predict?

- Optimize for signal upper limit in eg μ_{95} is the 95% expected limit in number of events, FOM= $\frac{\mu_{95}}{\varepsilon}$

3σ "evidence for" and 5σ "observation of"

- Standard procedure in HEP. **Enforced by journals**.
- Justified by protecting against statistical fluctuations, mistakes, look-elsewhere-effect, etc
- IMHO almost unhealthy obsession with these standards.
- For a "measurement" of an expected rare SM process, what we care is the rate (cross-section, branching-ratio, coupling....)
  - Even a 2.5σ measurement can be interesting.
  - "Evidence" and "Observation" are just for bragging rights.
- For a BSM effect, even 5σ is not enough.
  - Extraordinary claims require extraordinary evidence (corroboration! x-checks!)
  - New Physics is not discovered with statistics alone.
- 5σ is p-value=1-in-3.5-million. Hard to trust any estimates of statistical fluctuations with systematics at that level

VOLUME 73, NUMBER 2 PHYSICAL REVIEW LETTERS 11 JULY 1994

Evidence for Top Quark Production in $\bar{p}p$ Collisions at $\sqrt{s} = 1.8$ TeV

counts from background alone. We find $\mathcal{P}_{\text{combined}} = 0.26\%$. This corresponds to a 2.8σ excess for a Gaussian probability function.

Ooops....

What happens in other fields?

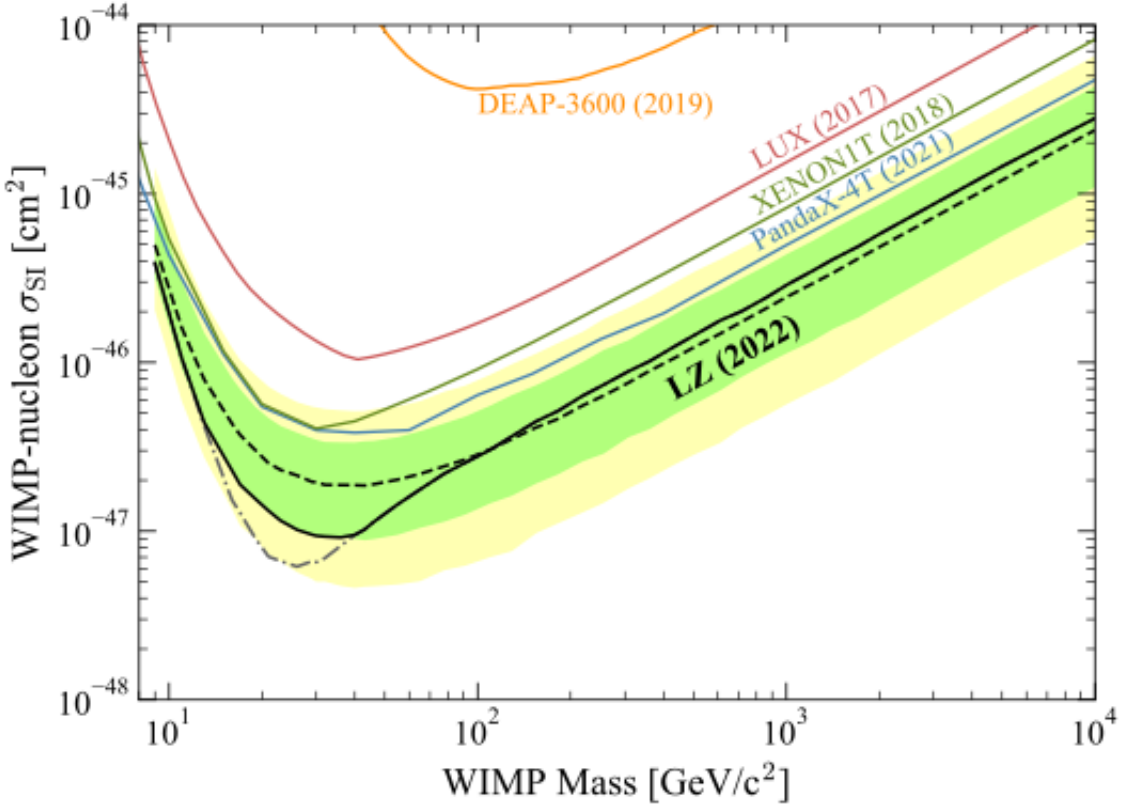
Social Sciences, Medicine….

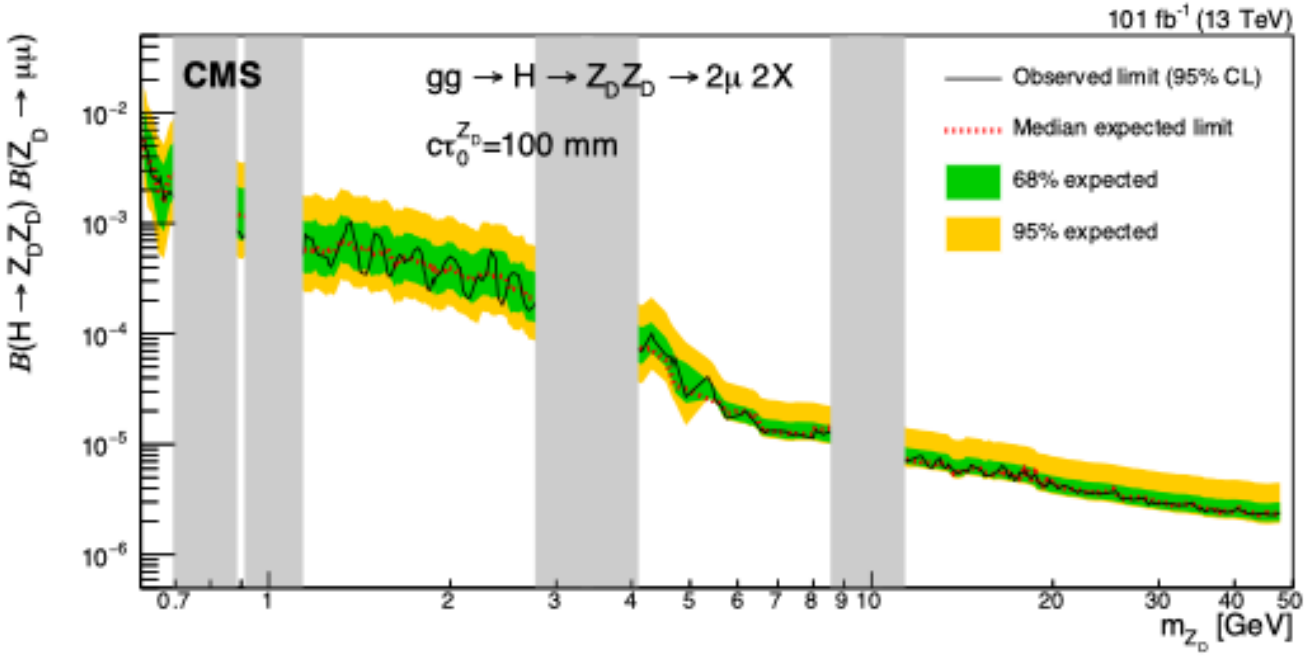
- p-value<0.05 → Significant (in practice: Accept the effect is true)
- Means 1/20 studies with no effect are reported as having an effect
- There is no culture to report “no effect”. Unpublishable.
- Huge incentive to “play around” until p-value<0.05
- Has led to a ”Reproducibility Crisis”
  - “p-hacking”, “garden of forking paths”
- For your amusement, try this game

Is experimental HEP immune from these problems?

- Not so sure….
- Blind analyses. Good, but not perfect.
- Analyses very complicated. If consistent with SM scrutiny can be “light”
  - Not out of malice. Reality and human nature.
  - Cynically I say that many of our results are “about right”.
- OTOH if inconsistent with SM much more scrutiny, task forces,….
  - It’s OK, extraordinary claims…..

Expected Limits





Publishing "expected" limits and "Brazilian Flags" is a relatively new (~15 year?) development in HEP. Allows for comparison of experiments (bragging rights!) taking away the luck of getting a limit better or worse than "deserved" because of statistical fluctuations.

Expected Limits, some dirty laundry

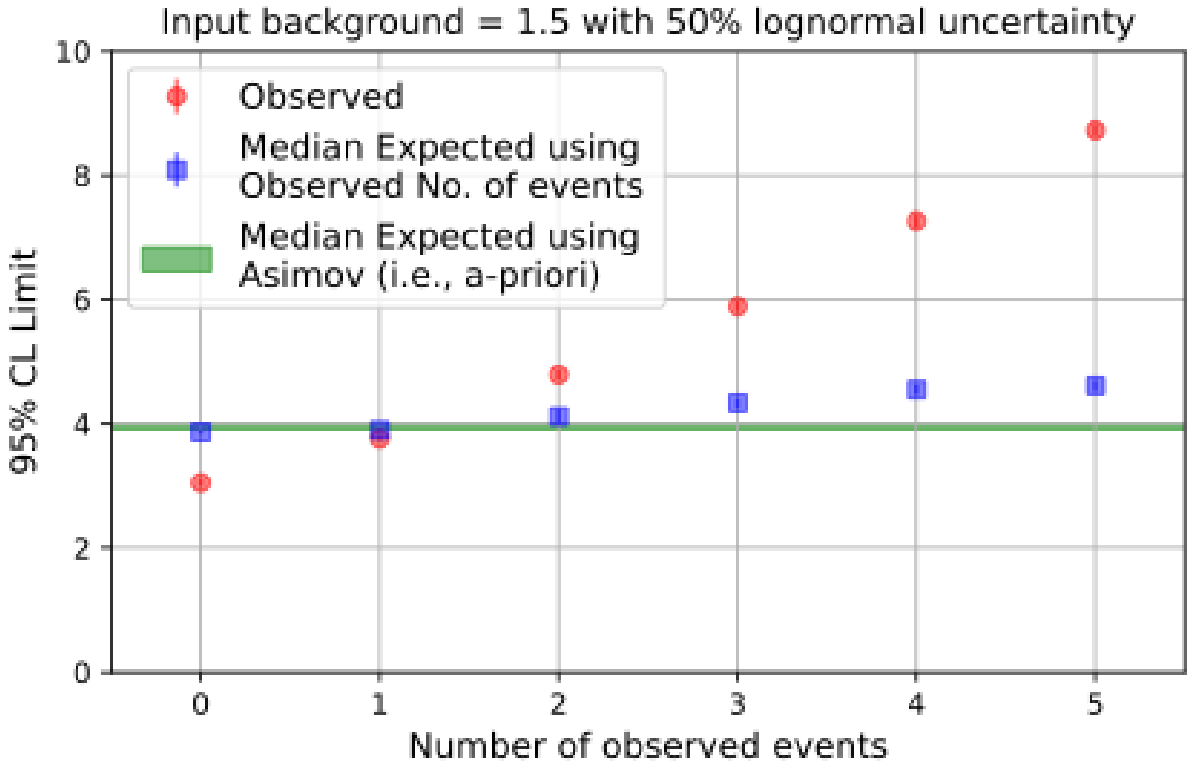
In a counting experiment the result is quantized (N=0,1, 2,…)

- It is not really possible to define accurately 1σ and 2σ expected bands
- But it is done through some (questionable?) gymnastics with the nuisances.

What does "expected" really mean?

- "expected a-priori" = calculated without looking at the signal region
- "expected a-posteriori" = calculated after profiling (i.e. fitting) nuisances (eg: background) after having looked at the signal region
  - **This is what is done at the LHC and I disagree with it**

a-prior and a-posteriori limits for simple counting experiment using LHC algorithm



Makes no sense to me to quote an expected limit based on the number of observed events in the signal region…

Feldman-Cousins (FC)

- A frequentist limit setting method that does not have the problem that occur when the background fluctuates low

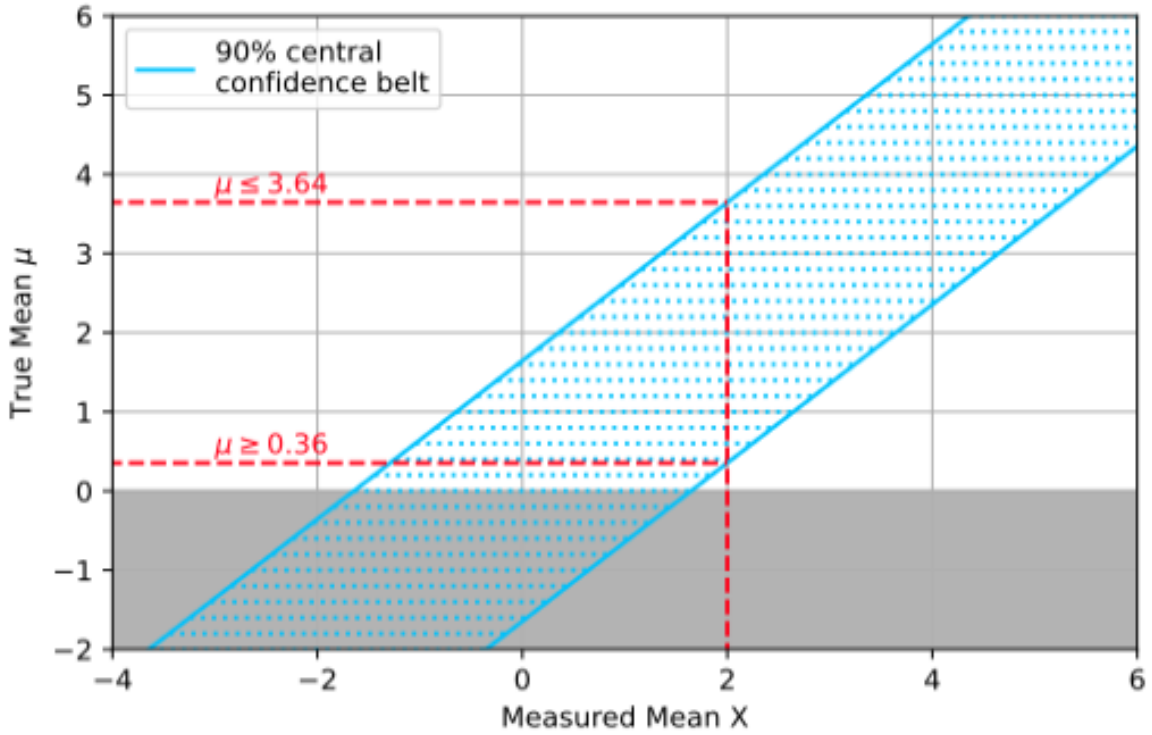
- CL_s fixes the problem, but then the ”coverage” is not what it is claimed to be. Power-constrained is ad-hoc. In contrast, the FC coverage is OK.

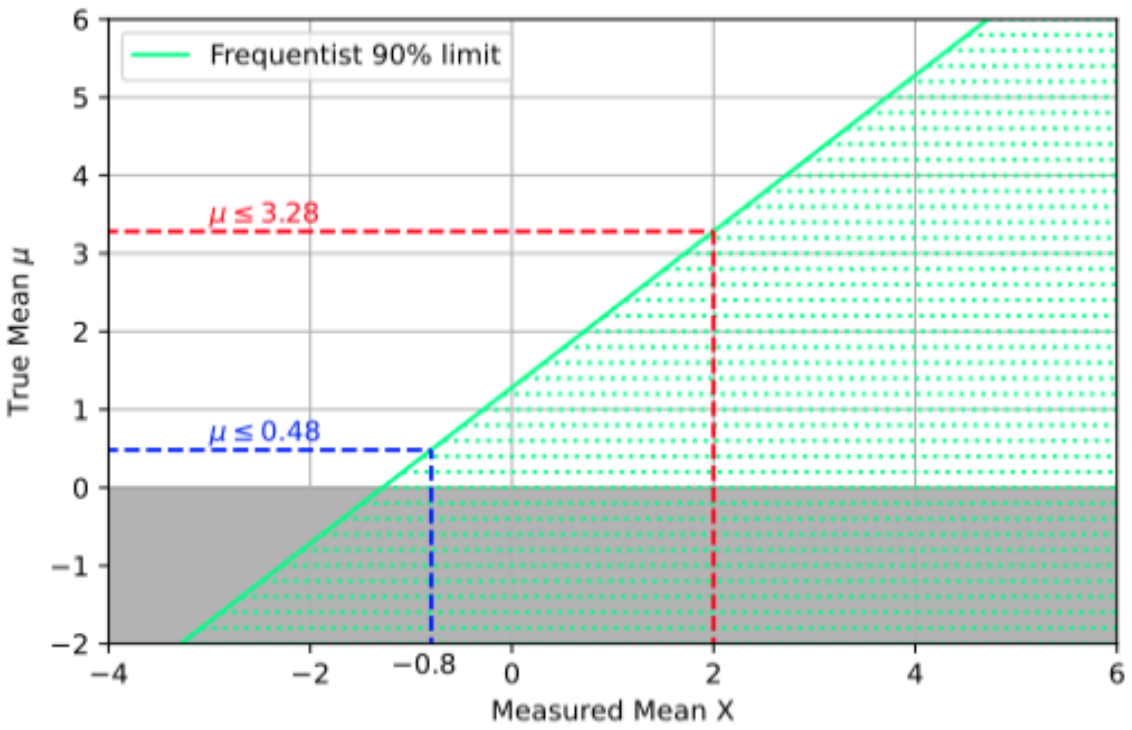
- It solves the (perceived?) problem of switching between “limit setting” and “observation”. FC call it the “flip-flopping” problem.

- Recall frequentist approach. 90% CL for concreteness.
- Gaussian measurement with $\sigma=1$
- x = measurement in units of sigma, μ = truth
- Central 90% belt, find $x_1 - x_2$ interval such that
  $p(x_1|\mu) = p(x_2|\mu)$ and $\int_{x_1}^{x_2} p(x|\mu)dx = 0.9$
- For 90% CL $\int_{x_0}^{\infty} p(x|\mu)dx = 0.9$

- Recall frequentist approach. 90% CL for concreteness.
- Gaussian measurement with $\sigma=1$
- x = measurement, μ = truth
- Central 90% belt, find $x_1 - x_2$ interval such that
  $p(x_1|\mu) = p(x_2|\mu)$ and $\int_{x_1}^{x_2} p(x|\mu)dx = 0.9$
- For 90% CL $\int_{x_0}^{\infty} p(x|\mu)dx = 0.9$

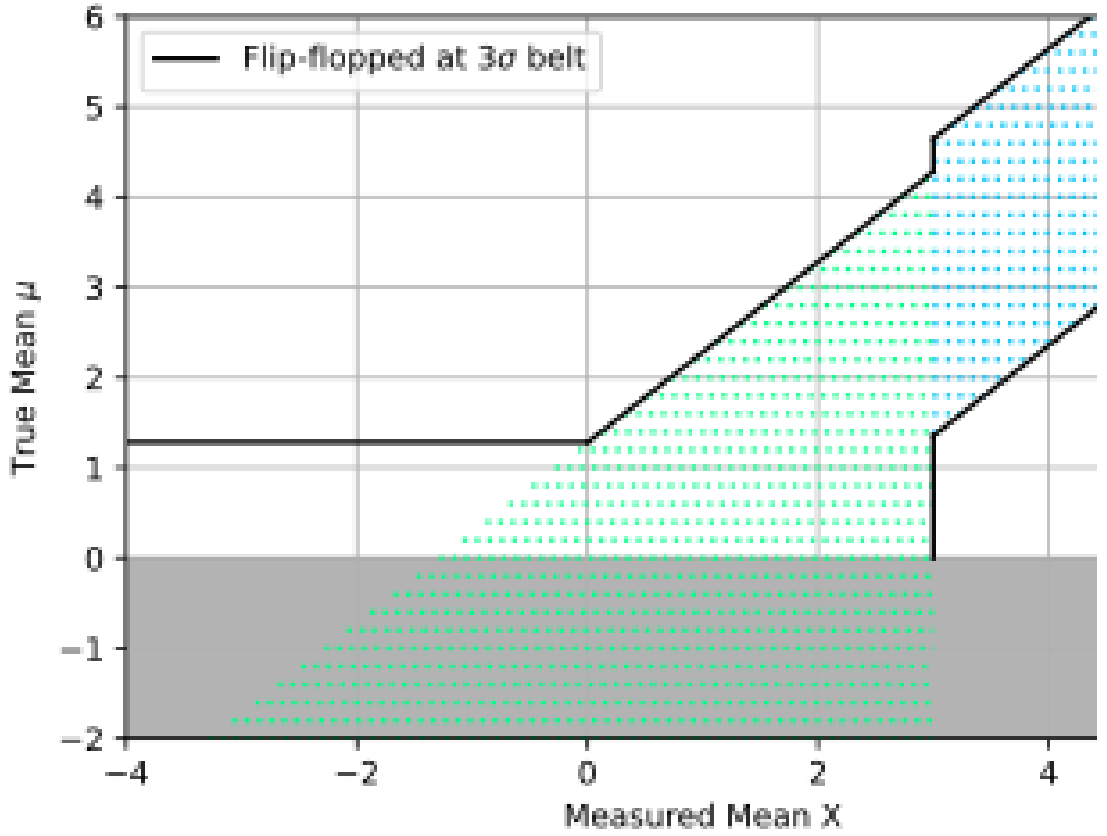
Here we show examples for measurement x=2 and, in case of limit, x=2 and x=-0.8
The physical region is $\mu>0$
For x<-1.3 the frequentist limit does not exist



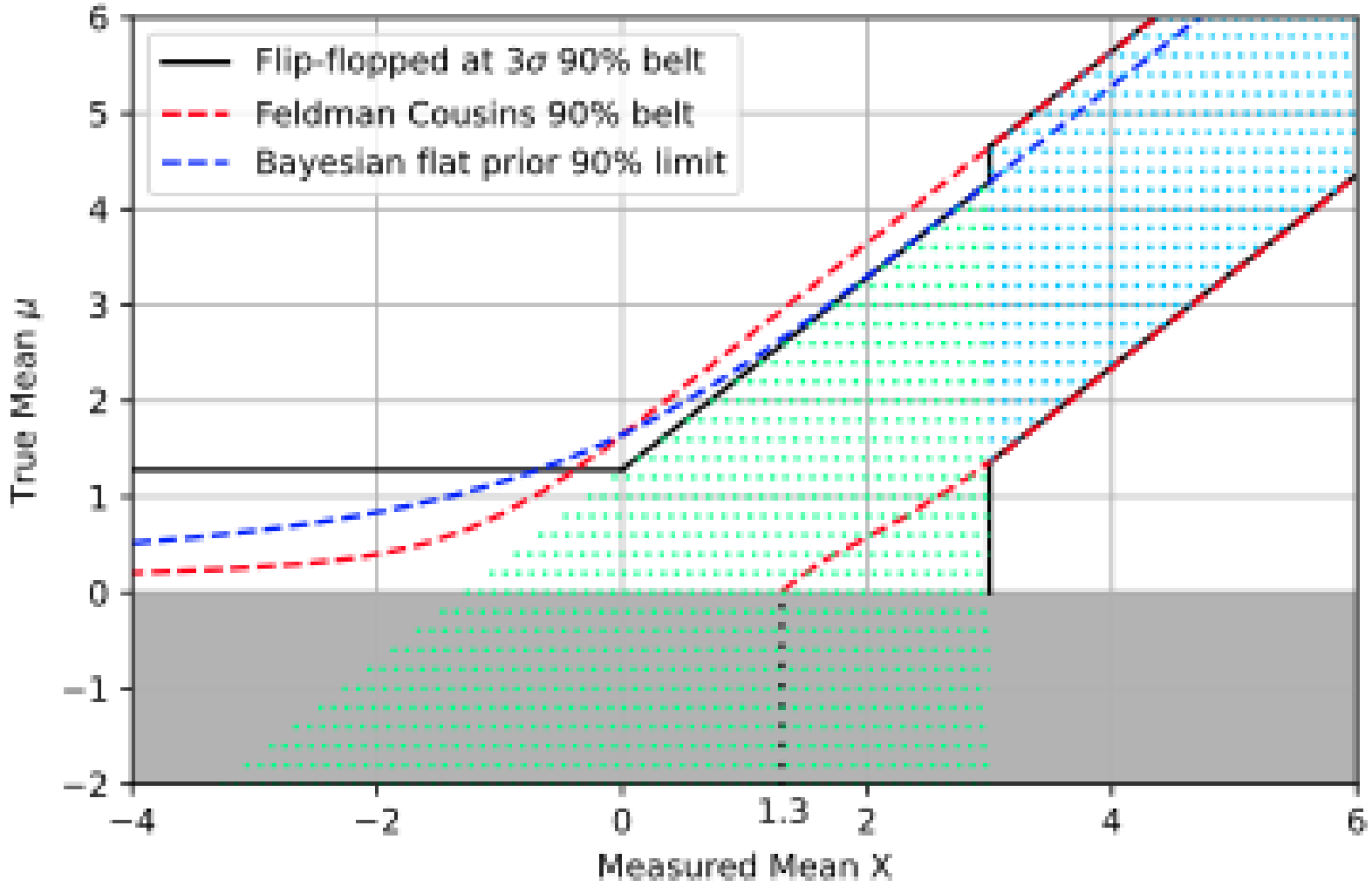


Flip-flopping problem, in case of an excess

- Suppose I say: if I get a greater than 3σ (x>3) result, I quote a measurement, otherwise a limit
- Then the belt is discontinuos



FC: No flip-flopping and works even when *x* negative



How they determine this new belt?

- Define ranking parameter $R(x|\mu) = \frac{p(x|\mu)}{p(x|\hat{\mu})}$ where $\hat{\mu}$ maximizes $p(x|\mu)$ subject to $\mu>0$
- Ratio of likelihoods.
- At any value of μ, order the values of x from largest lo smallest R.
- Add more and more values of x until one covers the required probability (eg 90%)
- Works all the way to $x<<0$. Unifies limits and measurements

Final Remarks

- Statistical Methods are playing a large role in modern HEP analysis
- It is important to have some understanding of what is going on
  - I do not like complete black boxes….do you?
- Common sense should rule
- I hope that you found these two quick lectures useful
  - Take them as an inspiration to dig a little deeper

Homeworks ☺

- David had asked me for some “exercises” as part of the school
- I picked a few from the homework assignments from my class
- If you are interested, try them out, either here or when you go back home when you have more time
- I have the “solutions” including the associated code (python jupyter notebooks) and I plan to make them available
  - But first I need to repackage them in a convenient way.

Some resources for you to look at

- Glenn Cowan and Luca Lista
  - Statistics in the 2026 PDG
- Roger Barlow
  - Statistics: A guide to Use of Statistical Methods in the Physical Sciences. (Manchester Physics Series)
- Glenn Cowan
  - Statistical Data Analysis (Oxford Science Publications)
- Frederick James
  - Statistical Methods in Experimental Physics, 2nd Edition (World Scientific)
- Luca Lista
  - Statistical Methods for Data Analysis, 3rd Edition (Springer)

Also: there are links inside these slides to various papers if you want to dig deeper