Knut Dundas Morå
fysikk@dundasmora.no, he/him

School of Underground Physics at Bertinoro

Statistics and Inference

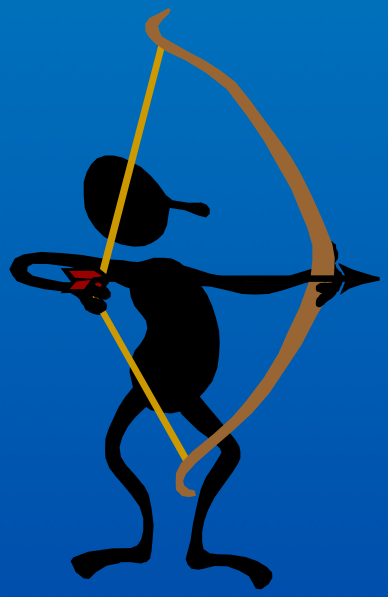
for rare event searches

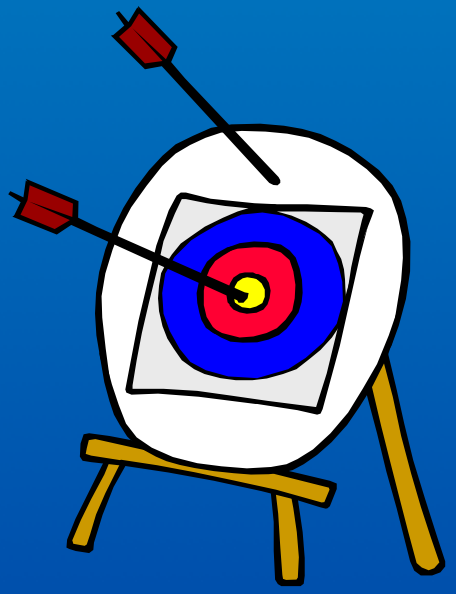
What is a statistical model?
Does it describe your data?
What kinds of conclusions can we draw?

24

GOALS!

This course should teach you:





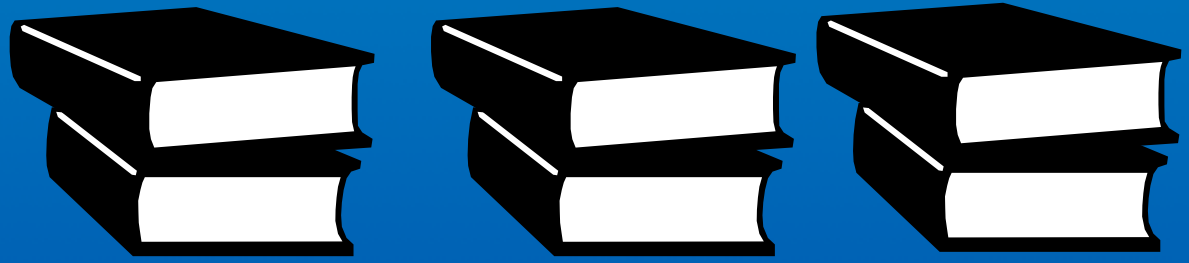
- To construct a statistical model for your experiment
- To consider how to test whether your data is compatible with your statistical model
- To use your statistical model make statements about physics
- And to interpret others' statements

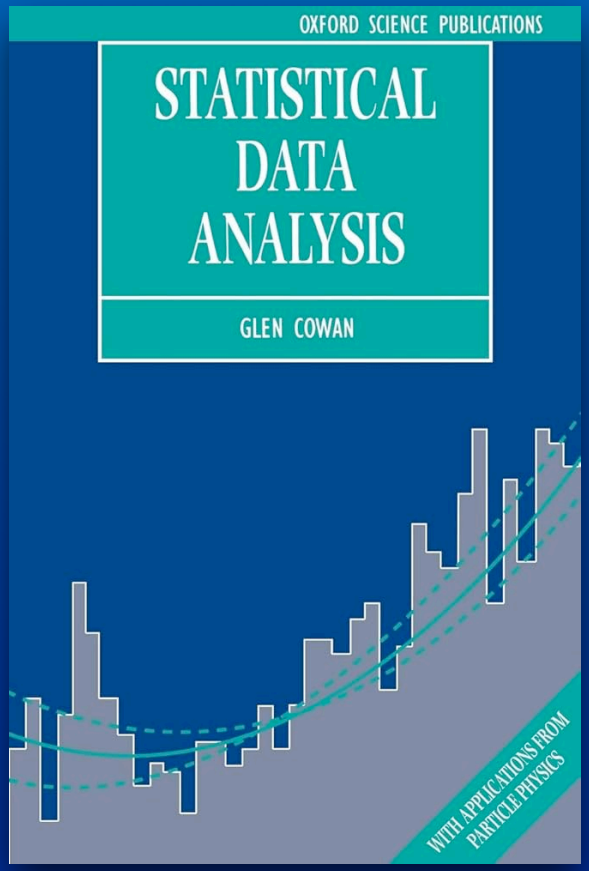
Structure

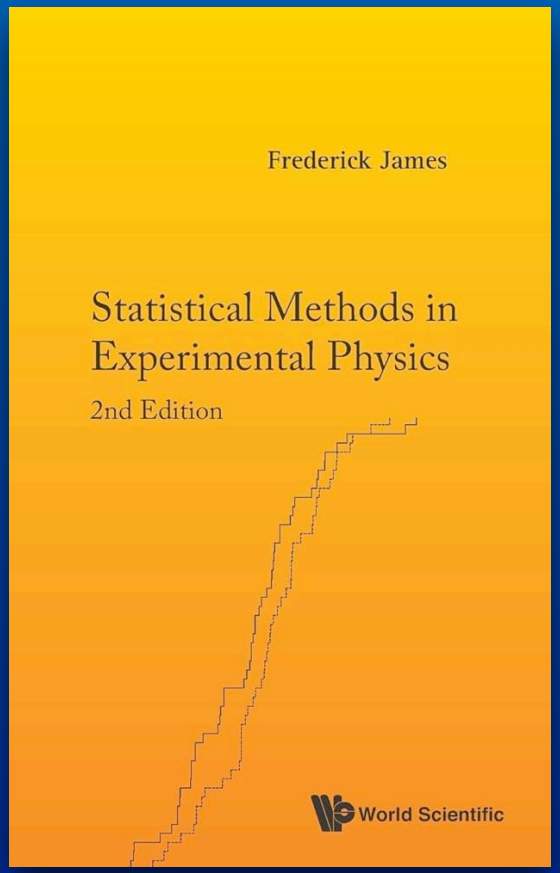
Three lectures and exercise sessions

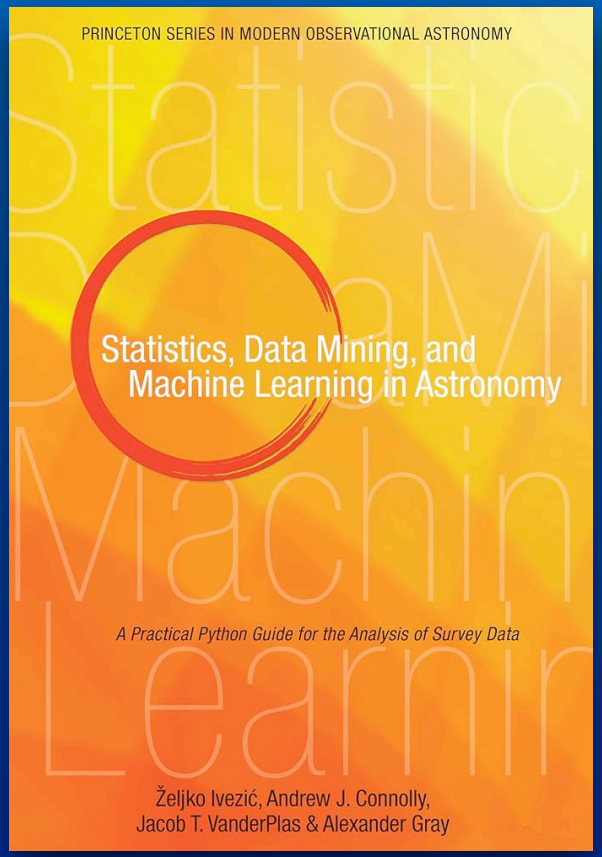
- TUE 1645-1830
  - Introduction
  - Hypothesis Testing
  - Goodness of Fit
- WED 1115-1300
  - Example analyses
  - Profile Likelihood
  - Asymptotic distributions
  - Look-Elsewhere effect
- THUR 1645-1830
  - Confidence Interval construction
  - Nonasymptotics
  - Bayesian credible intervals
  - Unfolding
  - Tools

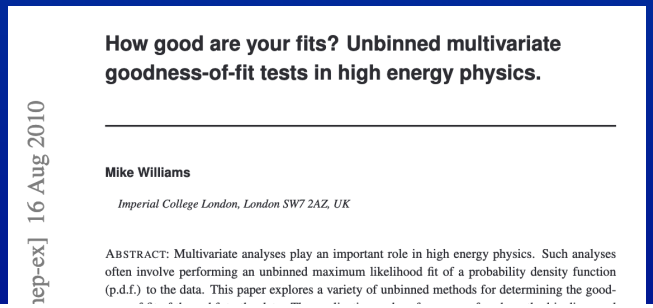
Resources











How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics.

Mike Williams

Imperial College London, London SW7 2AZ, UK

ABSTRACT: Multivariate analyses play an important role in high energy physics. Such analyses often involve performing an unbinned maximum likelihood fit of a probability density function (p.d.f.) to the data. This paper explores a variety of unbinned methods for determining the good-

hep-ex] 16 Aug 2010

https://arxiv.org/pdf/1006.3019

- useful to think about the goodness-of-fit challenge

Notation

SOUP 24

I'll follow Frederick James' Statistical Methods

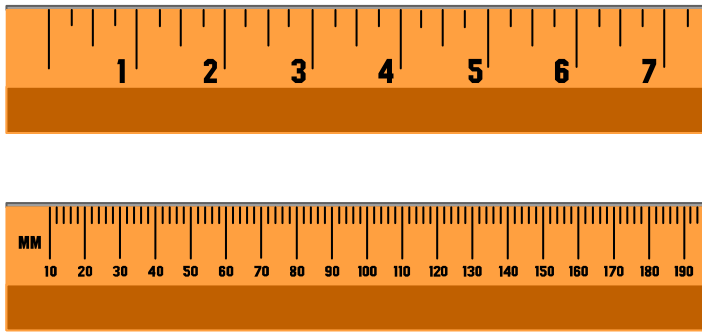
- A random variable, or several are $X, X_i, \mathbf{X}$
- The probability of an event A is $P(A)$
- Parameters of a model are $\boldsymbol{\theta}$
- Conditional probabilities are $P(A \,|\, B)$
- The likelihood is $\mathscr{L}(\boldsymbol{\theta} \,|\, \boldsymbol{X}) = P(\boldsymbol{X} \,|\, \boldsymbol{\theta})$
- Expectation value(s) for counting experiments are $\mu, \boldsymbol{\mu}$
- Expectation values, variance $\mathrm{E}(X), \mathrm{V}(X)$
- best-fit parameters or point estimates are $\hat{\theta}$

Observed data is random variables

SoOP 24

or is it are

- Our measured data is a result of processes both truly and practically random (e.g. quantum processes, me reading a ruler crooked)
- In some cases, the data itself is close to what we wish to measure, and we hardly think of ourselves doing statistics
- However, in particular when looking for small or subtle effects, the random noise may be significant, and the relationship between physics parameters and the measured quantity less straightforward
  - You'll need to make a statistical model for how your data came to be,
  - And methods to make sound conclusions



Test statistics are functions of your observations

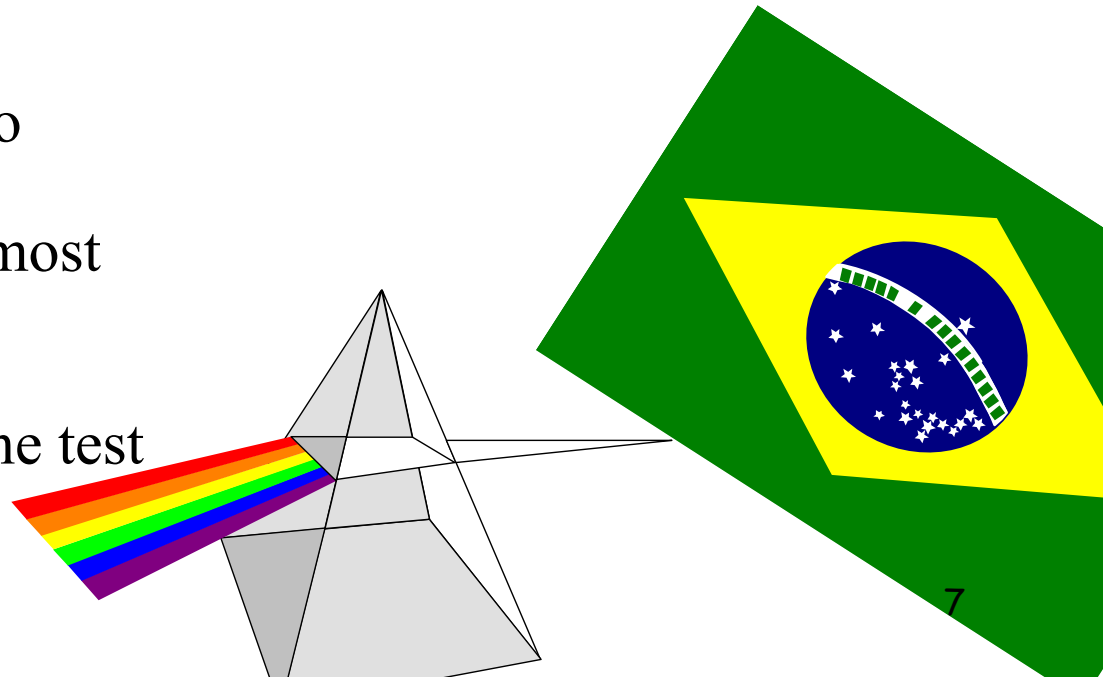
- Any function of your observed data will be a random variable

$$\hat{\mu} = \frac{1}{N}\sum_i x_i$$

- By using the right function, we can gather all the information gathered into one number
  - E.g. estimators ($\hat{s}$) which directly give a measurement of some parameter

$$\hat{\sigma} = \sqrt{\frac{\sum_i (x_i - \hat{\mu})^2}{N-1}}$$

- The tricky part will most often be to
  - choose the function to give the most information from the data, and
  - Understand the *distribution* of the test statistic



Probability distributions/densities

SoOP 24

- if X is a continuous variable, we may define a probability *density* function (PDF) to describe the distribution

$$f(X) = \mathrm{Lim}_{\epsilon \to 0} P(x_0 < X < x_0 + \epsilon)/\epsilon$$

- The cumulative density function (CDF), $F(X)$, is often also useful

$$F(X) = \int_{-\infty}^{X} f(X')\mathrm{d}X'$$

  - and its inverse!

$$P(X_0 < X < X_1) = F(X_1) - F(X_0)$$

Expectation values, variance

Useful Summaries of location: $\mathrm{E}(X) = \int_{-\infty}^{\infty} X \cdot f(X)\mathrm{d}X$

and spread: $\mathrm{V}(X) = \int_{-\infty}^{\infty} (X - \mathrm{E}(X))^2 \cdot f(X)\mathrm{d}X$

Linear: $\mathrm{E}(a \cdot y(X) + b \cdot z(X)) = a \cdot \mathrm{E}(y(X)) + b \cdot \mathrm{E}(z(X))$

If x_i are identically distributed independent random variables:

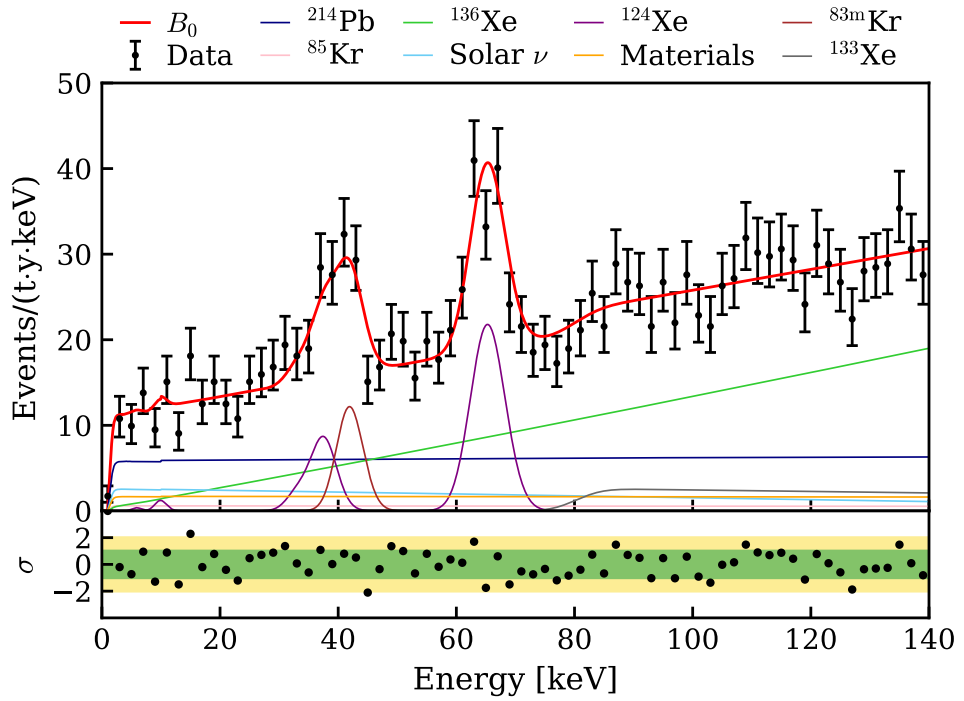
the mean estimator $\hat{\mu} = \frac{1}{N}\sum_i x_i$, has the correct expectation

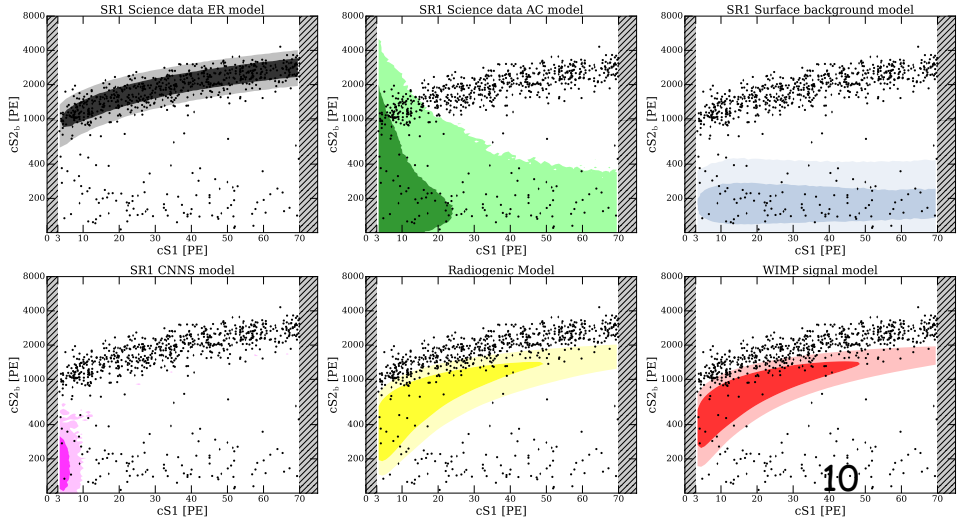
$E(\hat{\mu}) = E(X)$

as doers the variance estimator $\hat{\sigma^2} = \frac{\sum_i (x_i - \hat{\mu})^2}{N-1}, E(\hat{\sigma^2}) = V(X)$

Any number of distributions!

- If we are certain about the outcome, is it really an experiment?
- Depending on what you measure, your distributions may be as simple or as complicated as can be imagined
- However, for many problems, physical considerations or your experience may lead you to have a look at some of the most common ones used— they are useful building blocks!
- Some (student T, F-test, χ^2) are also useful because they describe the behaviour of some useful test statistics

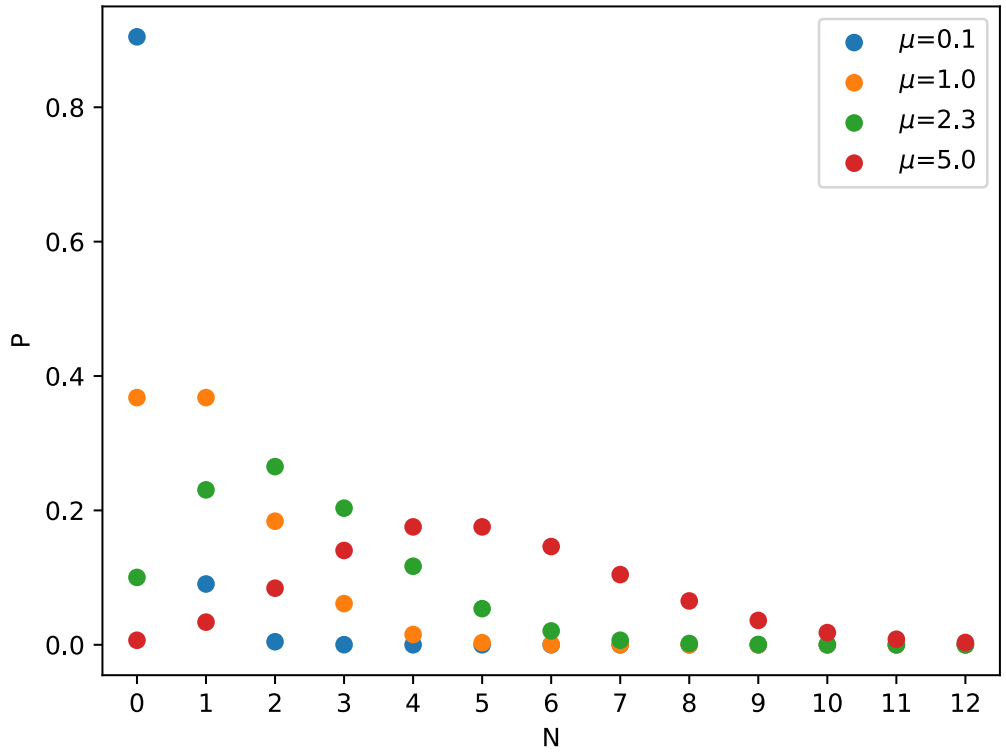




The Poisson distribution

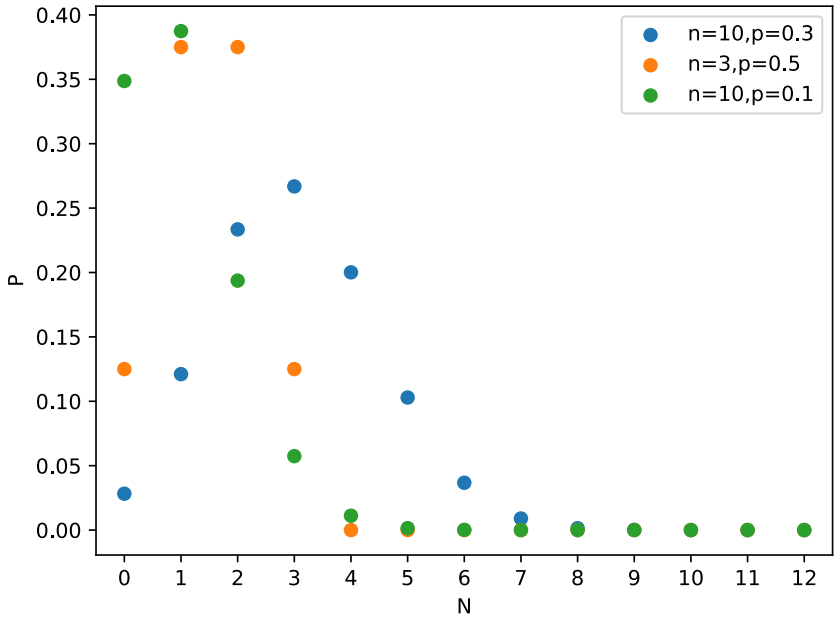
$$P(N) = \frac{\mu^N e^{-\mu}}{N!}$$

- If you count events that happen in a certain period, you'll end up with a Poisson distribution
- Expectation value and variance are both μ



Bi/multi-nomial distributions

- If we count how many times each of a finite set of outcomes happens, we get the multinomial distribution
  - M total tries, n_i events in each category, with probability p_i
  - And if the number of possible outcomes $k = 2$, we get the Binomial distribution
- Examples: Histogram bin counts, classification

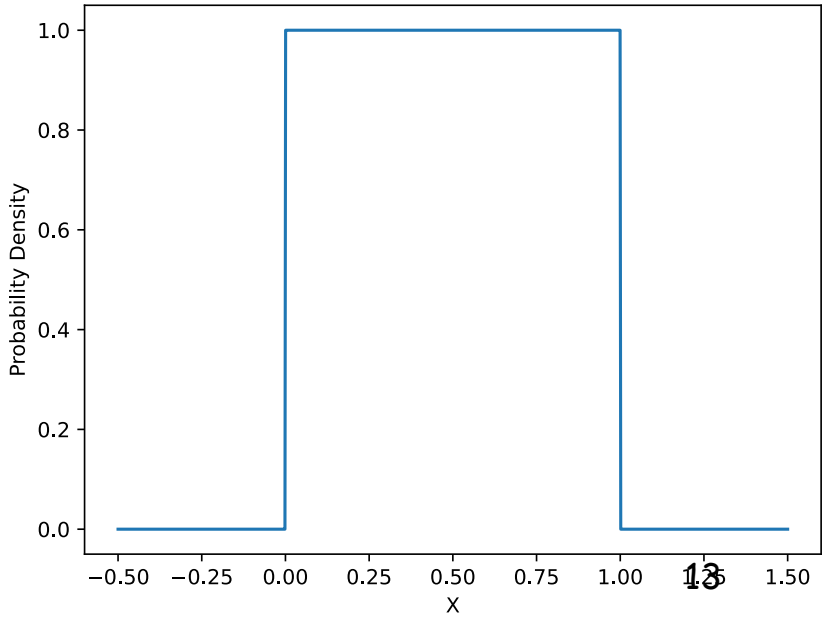


Uniform distributions

- Turns up in e.g.
  - Spatial distribution of dark matter events?
- But more importantly, it is often very often useful to convert another distribution into a uniform distribution (*Y* here) between 0 and 1

$$f(X) = \begin{cases} \frac{1}{b-a} & \text{if } a < X < b \\ 0 & \text{else} \end{cases}$$

$$Y(X) = \int_{-\infty}^{X} f(X')\mathrm{d}X'$$

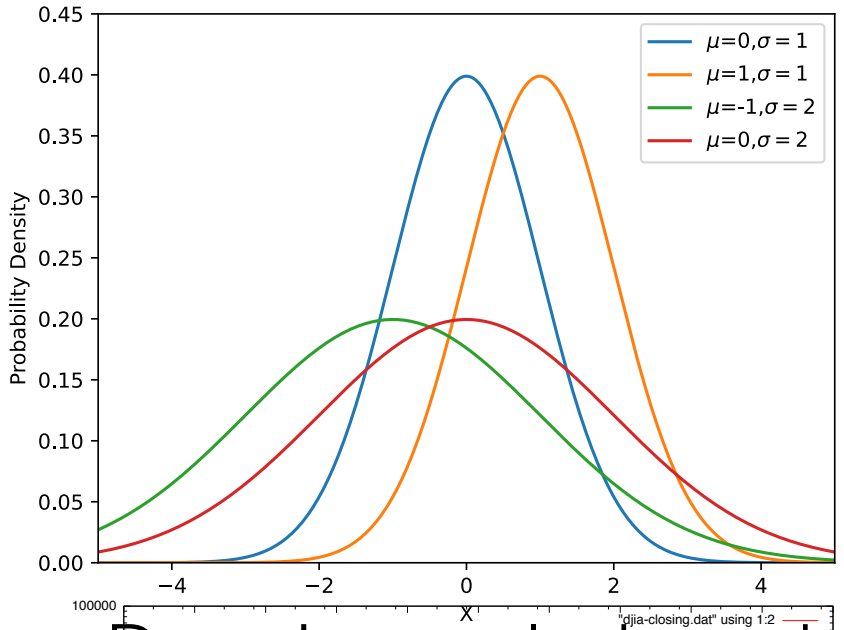


The Gaussian distribution

The industry default. AKA bell curve, normal distribution

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$$

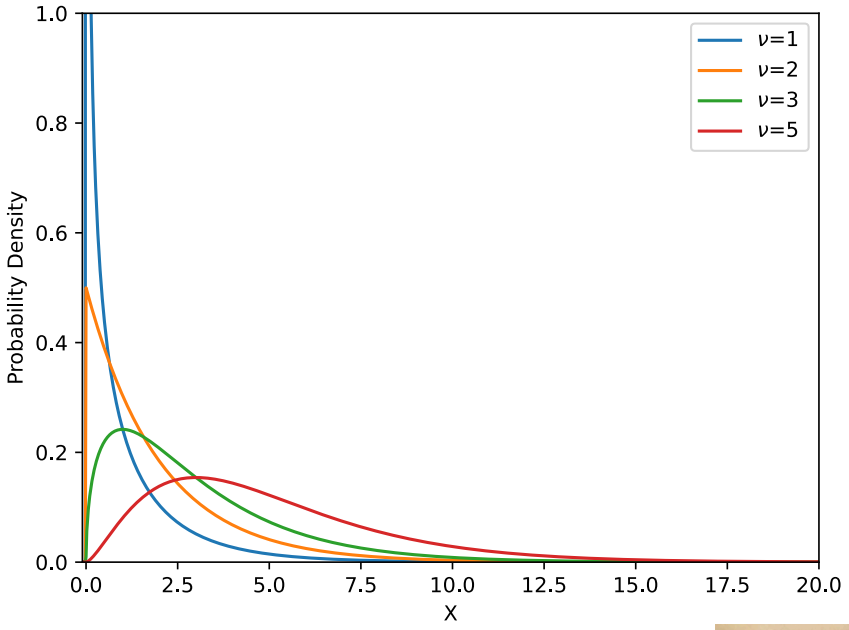
- The Gaussian distribution is the limit of sums of random numbers with finite mean and variance— the Central Limit Theorem
  - E.g. — diffusion!
- For this reason, it is often the "default" assumption for a continuous distribution
- However, by using this (or many other analytical distributions) you may be assuming to know the behaviour for even very extreme outliers



The χ^2-distribution

$$f(X\,|\,\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} X^{\nu/2-1} e^{-x/2}$$

- The sum of the square of ν standard normal distributed numbers is distributed according to the χ^2 -distribution
- We'll see later that this means that you'll encounter this distribution frequently when computing confidence intervals



$$\sum_{i=1}^{N} \frac{(X_i - \mu_i)^2}{\sigma_i^2} \sim \chi^2_{\nu=N}$$

Histograms as distribution estimates

- If you wish to characterise the distribution of, for example, the distribution of energy deposited by electrons and photons in a calorimeter, or the total path length of all tracks, you may never find an analytical estimate
- Higher dimensionality can challenge this approach
- and you'll need to check you have enough samples or include the uncertainty

Fitting using finite Monte Carlo samples (Barlow and Beeston)

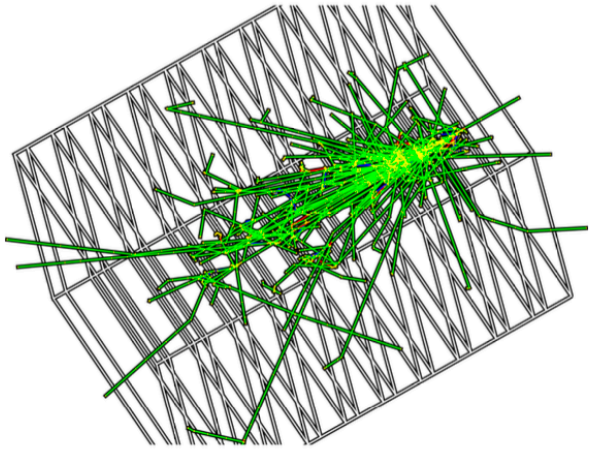
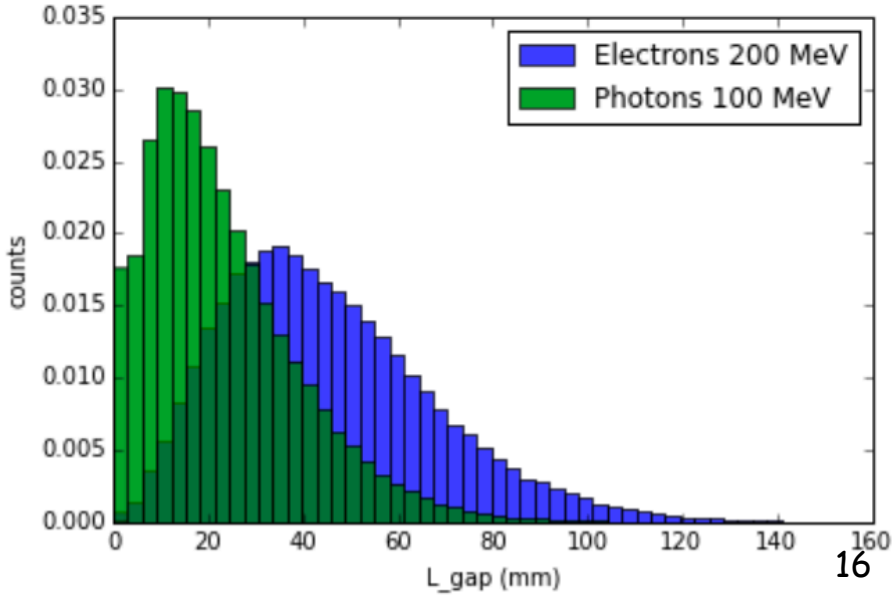


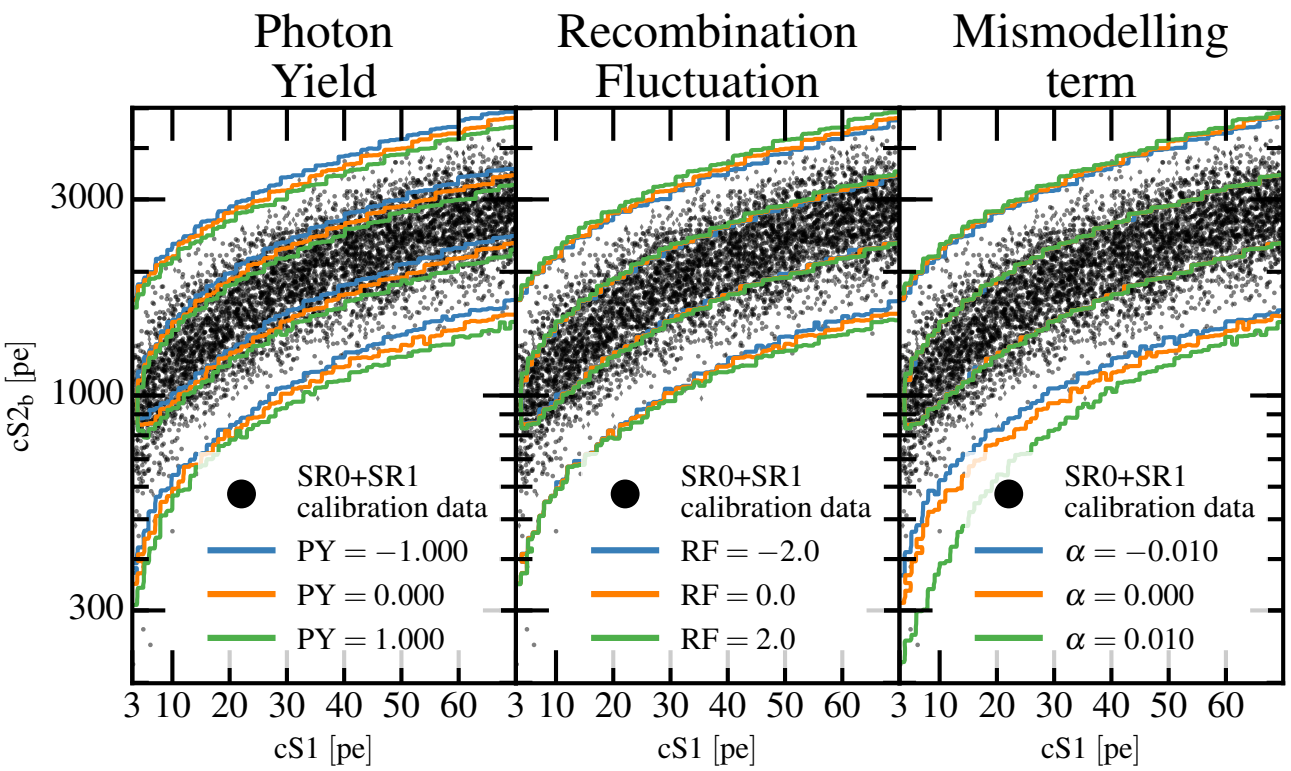
Figure 1: Electromagnetic shhower in calorimeter induced by photon



Technical aside 1: Shape and Rate?

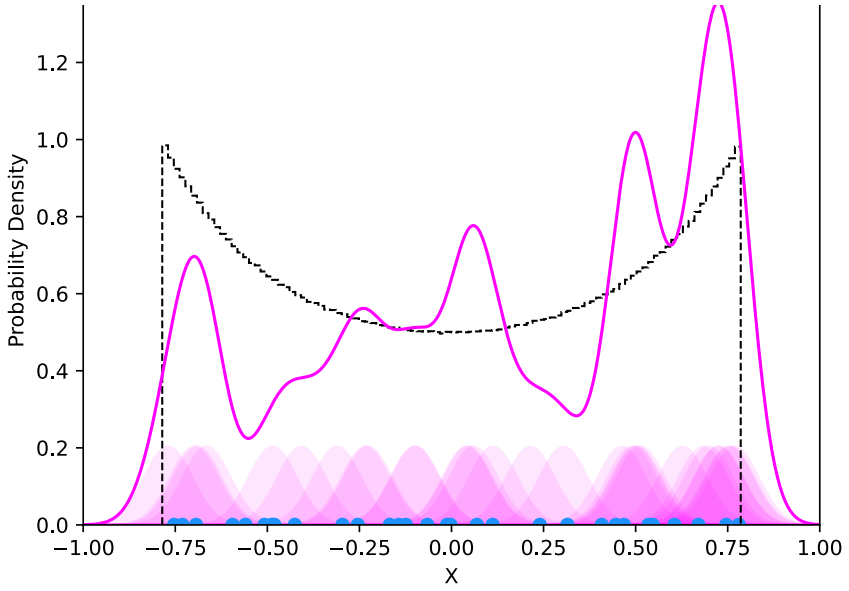
SoUP 24

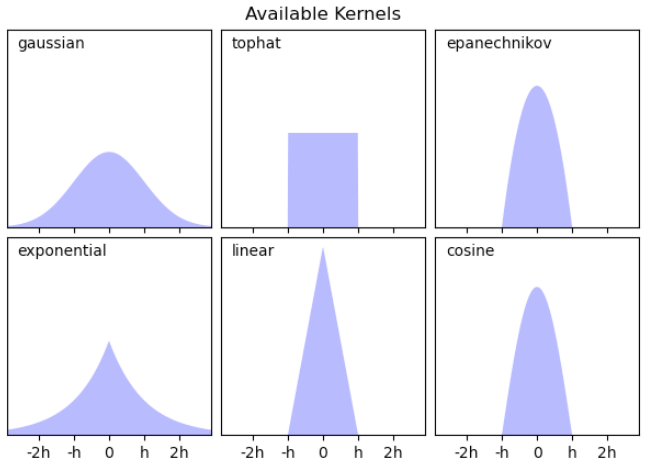
- When using histograms to estimate the distribution, nuisance parameters are well-named
- To have a continuous nuisance parameter, “template morphing"-- linear interpolation between some points in parameter space is often used
- Since this is computationally tricky, there will often be a divide between "rate parameters”-- those that only affect expectation values, and therefore are “easy” and "shape parameters”— those that require modifying the PDF of one or more signal/background model



KDEs

- Another method to estimate densities, or too make a distribution estimate smoother is to use a kernel density estimate— adding a kernel, a PDF centred on each event in the sample
  - To choose the width of this kernel, you may have to split your dataset in a fit and validation dataset
- If your distribution has sharp edges, or areas with very dissimilar densities, you may wish to use an adaptive KDE





scikit-learn provides extensive KDE functionality

Frequentist interpretation of probability

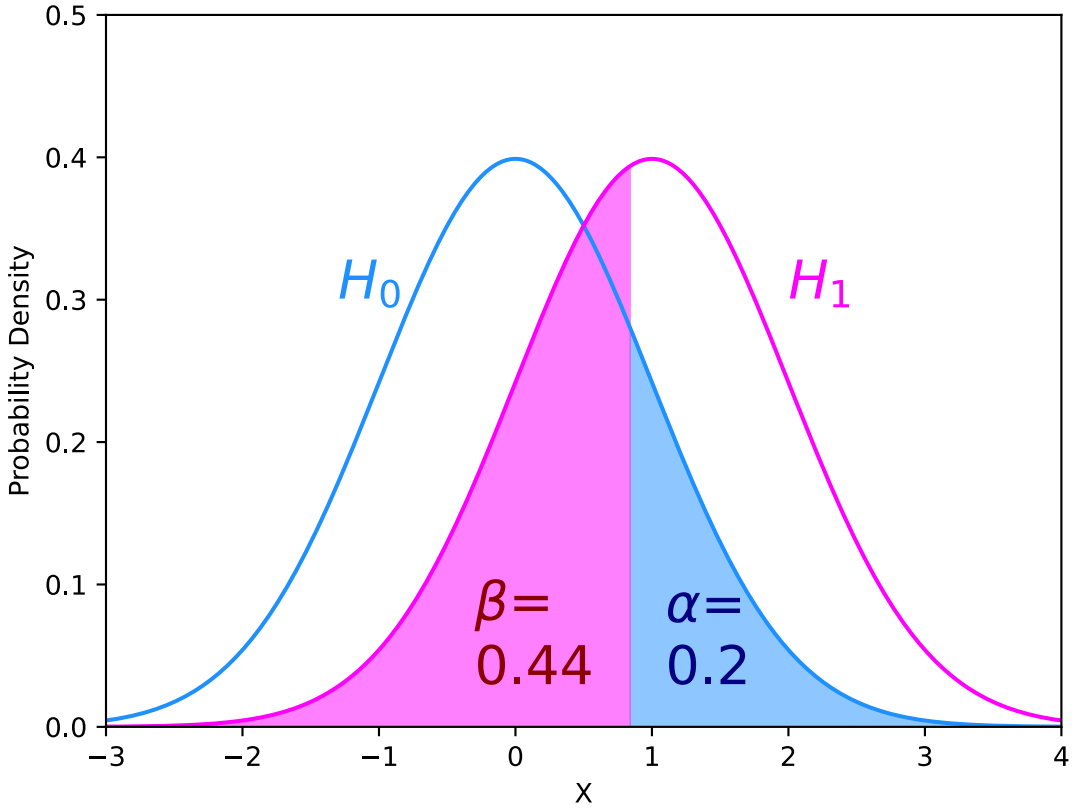
24

- The frequentist interpretation of probability is the relative frequency of some outcome in the limit of infinite number of repetitions
  - This limit needs only be in principle— valid frequentist inference can occur for a single experiment as l*ong as that experiment is repeatable*
- Views the data as random outcomes of fixed processes
  - In some sense— a very particle physics way of looking at the world
- Dominant in particle physics

Hypothesis testing

SoOP 24

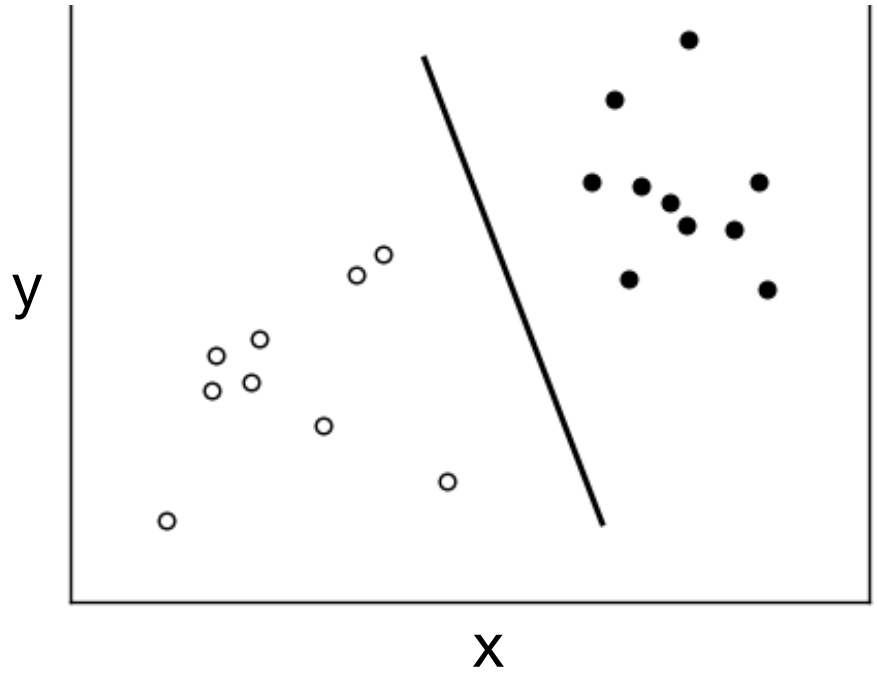
- Frequentist hypothesis testing: make a decision between the two alternatives
- You get to choose:
  - What test statistic you use to separate the two hypotheses!
  - And, the decision boundary, either explicitly
    - Or implicitly by demanding a certain probability to reject H_0

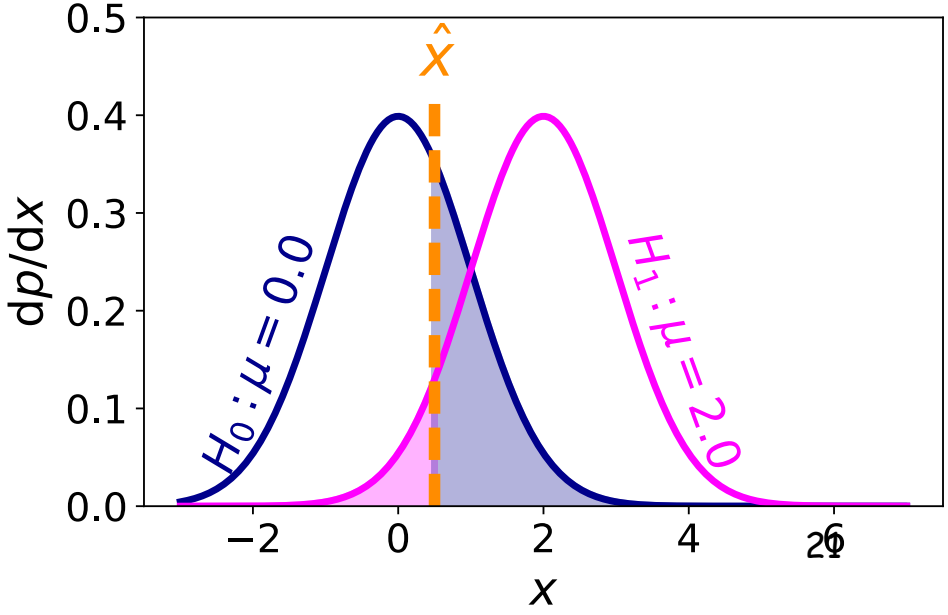


	P(accept H0)	P(accept H1)
H0 is true	1-α	α (test size)
H1 is true	β	1-β (power)

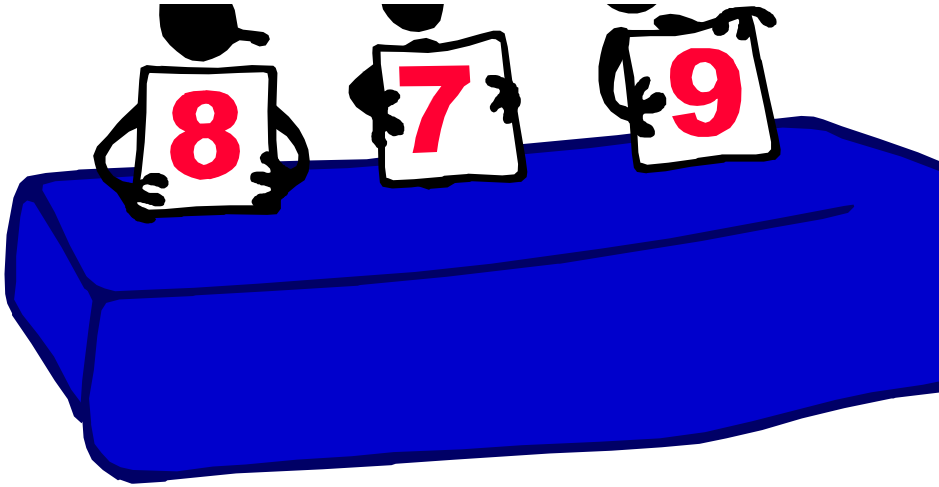
Test Statistics

- From the collected data, we wish to find a function of the data that expresses a direction or ordering of the data in a more H0 or H1 direction
- Typical examples; mean, median etc.
- For the example to the right, y would be a poor test statistic if we wish to distinguish the two, x would be better, and a combination would provide very good separation





What is a p-value?



- Since we want to use the best test statistic for each case, we could have many ways of measuring agreement with a hypothesis
- However, we can transform all our rulers into the same space by using p-values, which works with the integral of the distribution of T
- all p-values are between 0 and 1, and are defined by deciding on:
  - a test statistic
  - and a decision of what direction that test statistic expresses more tension with H_0
- Under H_0, p is uniformly distributed between 0 and 1

$$p(T_{\text{obs}}) = \int_{T_{\text{obs}}}^{\infty} f(T \mid H_0)\mathrm{d}T$$

p-values are the probability to observe a dataset equally or more extreme* than the one observed, given a certain (null) hypothesis

*ordering by a test statistic**

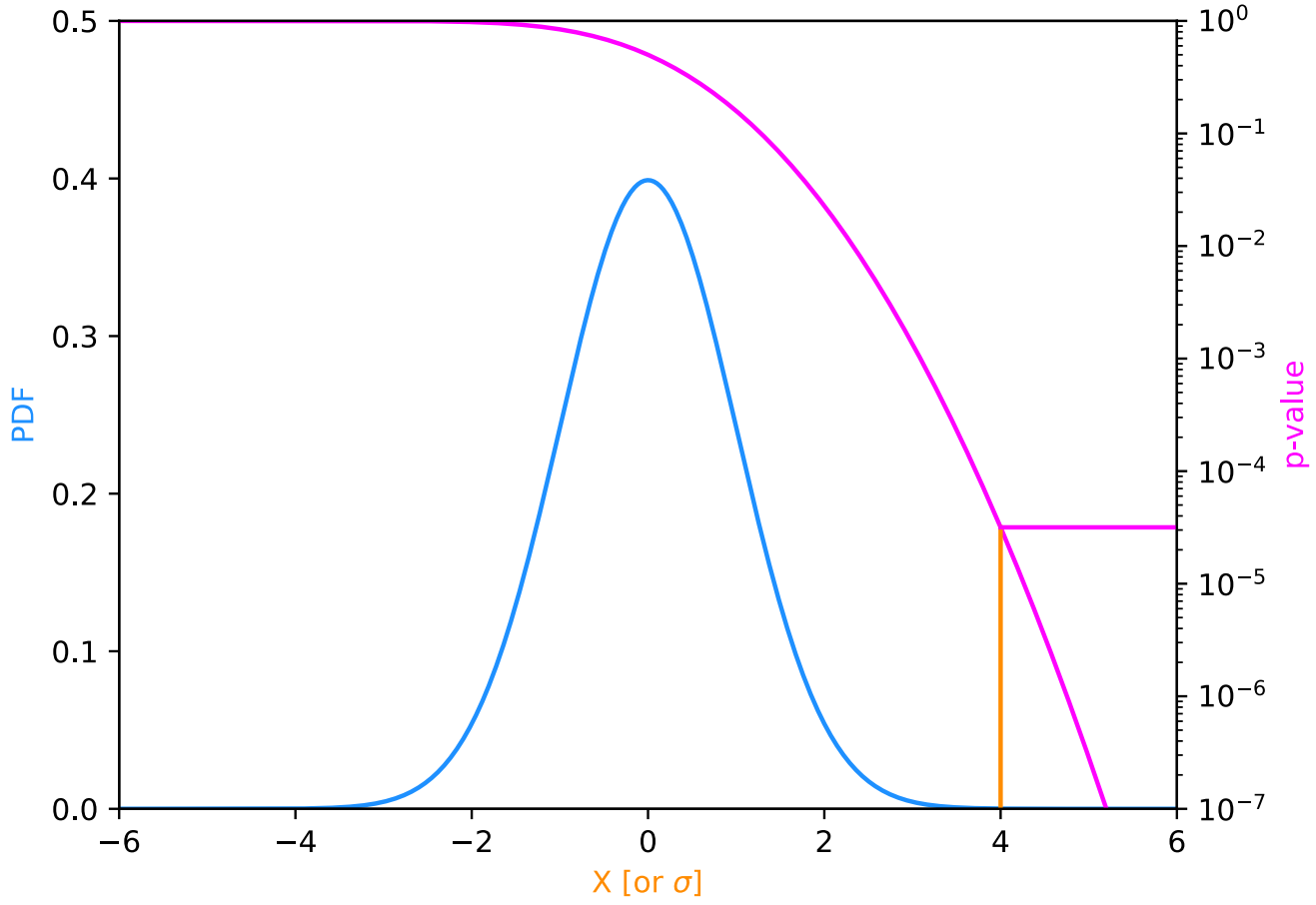
**usually chosen to separate the null and alternative hypothesis as well as possible

"Counting Sigmas"

This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

$$\sigma = \Phi^{-1}(1 - p)$$

- As a yardstick for p-values, you can often see "sigmas", or σ (or Z-score) used.
- "Five sigma", or 3×10^{-7} is the "standard" for discovery
  - Though you should consider what is the appropriate threshold in your field
- Be wary that you often also see the 2-sided version!



Why 5?

SoUP 24

Search	Degree of surprise	Impact	LEE	Systematics	Number of σ
Higgs search	Medium	Very high	Mass	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/low	Medium	Δm	No	4
Neutrino oscillations	Medium	High	$sin^2(2\theta), \Delta m^2$	No	4
$B_s \to \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/very high	M, decay mode	Medium	7
$(g-2)_\mu$ anomaly	Yes	High	No	Yes	4
H spin $\neq 0$	Yes	High	No	Medium	5
4^{th} generation q, l, ν	Yes	High	M, mode	No	6
$v_\nu > c$	Enormous	Enormous	No	Yes	>8
Dark matter (direct)	Medium	High	Medium	Yes	5
Dark energy	Yes	Very high	Strength	Yes	5
Grav waves	No	High	Enormous	Yes	7

Table 1: Summary of some searches for new phenomena, with suggested numerical values for the number of σ that might be appropriate for claiming a discovery.

https://arxiv.org/abs/1310.1284, Louis Lyons

The Likelihood

- A very useful test statistic is likelihoods—the probability of the data *given* a model
  - Likelihoods are central to most of both Bayesian and Frequentist methods
- As an example, the likelihood as a function of expected events for a counting experiment that sees 3 events is:
- We often deal with independent events (e.g. number of events in different histogram bins); we can build up a total likelihood by multiplying (or, using logarithms, adding) terms
- The well-loved χ^2-statistic is what you get if you combine Gaussian likelihood terms

$$\mathcal{L} = P(\text{data}|H)$$

$$\mathcal{L}(\mu|N=3) = \text{Poisson}(N=3|\mu)$$

$$\mathcal{L}(\vec{\mu}|\vec{N}) = \prod_i \text{Poisson}(N_i|\mu_i)$$

$$\log(\mathcal{L}(\vec{\mu}|\vec{x},\vec{\sigma})) =$$

$$\sum_i \log(\text{Gaussian}(x_i|\mu_i,\sigma_i)) =$$

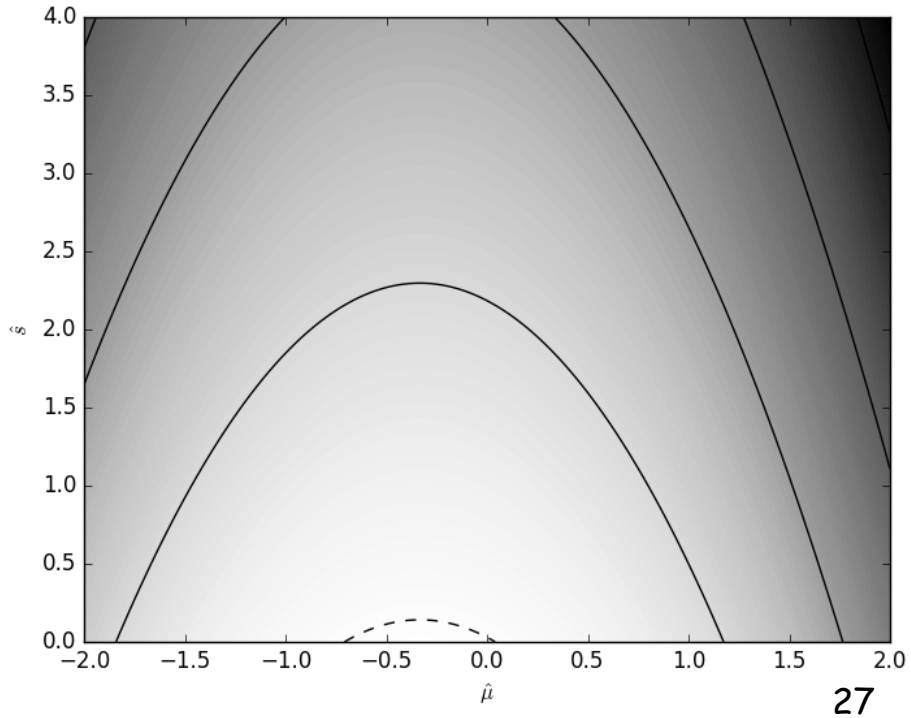
$$\sum_i (\frac{(x_i-\mu_i)^2}{\sigma_i^2}) + K$$

The Neyman-Pearson Lemma

$$\lambda = \frac{\mathscr{L}(H_1))}{\mathscr{L}(H_0)}$$

- IFF H_0 and H_1 are completely specified, the likelihood ratio between the two turns out to be the solution to the test statistic problem— it is the *uniformly most powerful test*.

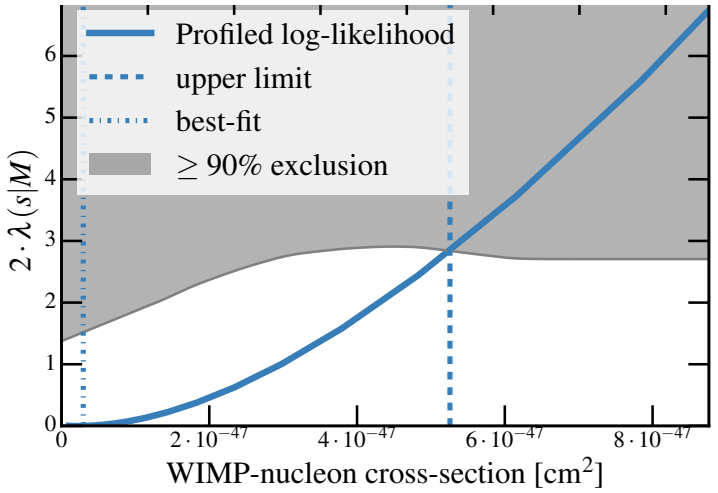
- For example, the plot to the right shows the NP ratio between two Gaussian hypotheses, one with $\mu, \sigma = 0,1$ and one 1,2.

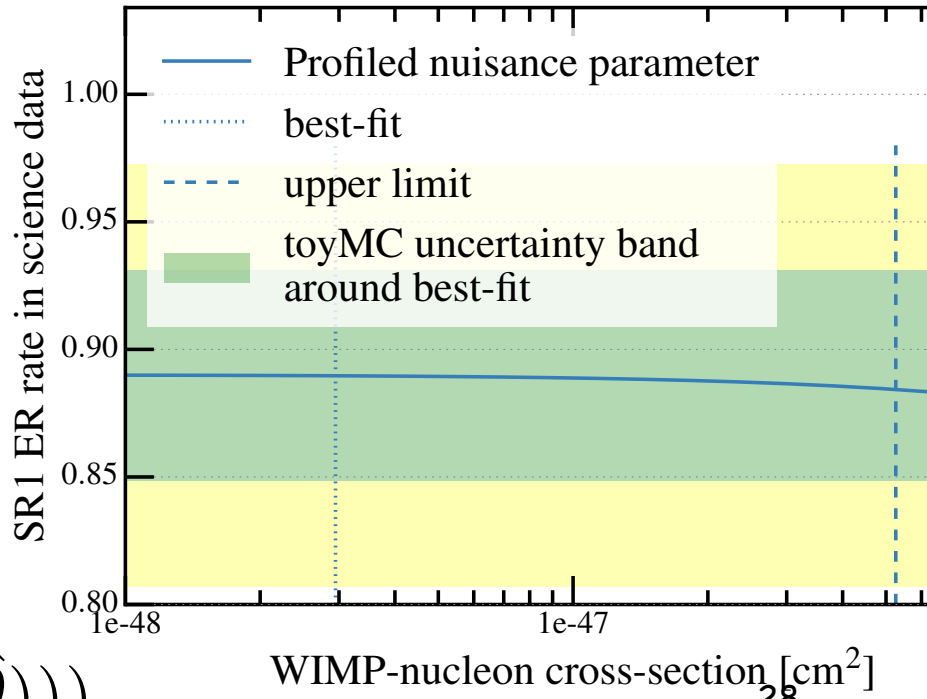


Profiling Likelihoods

SoUP 24

- We seldom have completely specified hypotheses
- Our background and signal models have uncertainties, parameterised by nuisance parameters (theta below)
- Unlike the Neyman-Pearson case, we are not guaranteed that this is the best possible test, but it very often performs well.

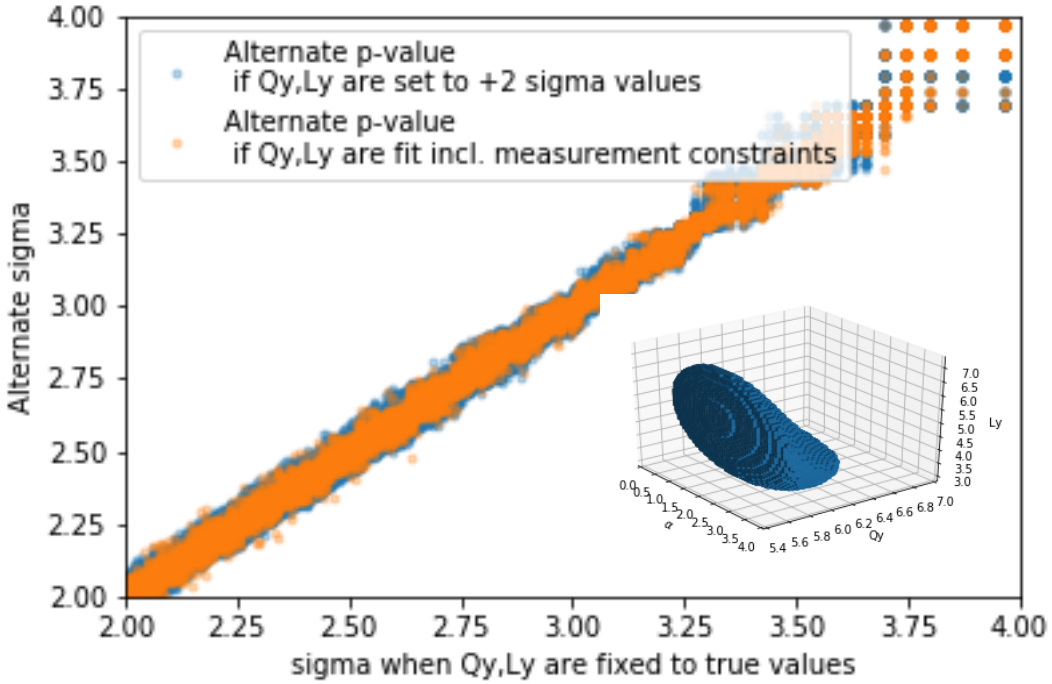


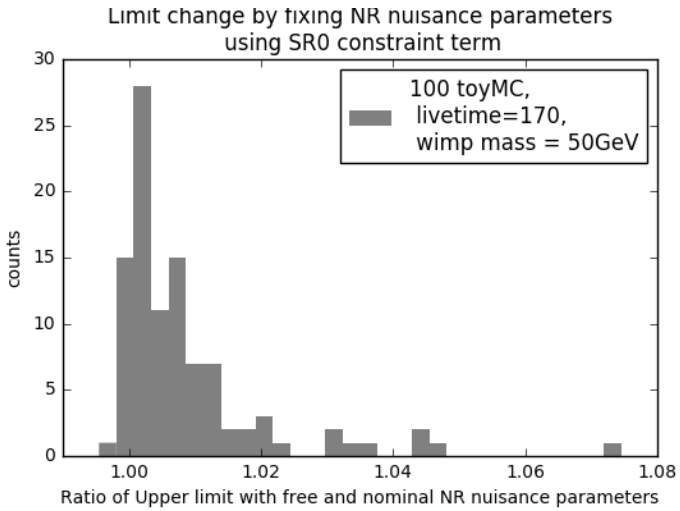


$$\lambda(s) = -2 \cdot (\log(\mathcal{L}(s, \hat{\hat{\theta}})) - \log(\mathcal{L}(\hat{s}, \hat{\theta})))$$

Follow-up question: What parameters may be ignored?

- We are rarely (never) able to include every possible uncertainty in our inference frameworks
  - And it is not likely that every parameter is important
- Need ways to decide which parameters are unimportant enough
- To my knowledge, no standards or consistency in how these questions are treated.
- To the right, two toy investigations in XENON1T— signal shape parameters often have very little impact on confidence intervals





E. Aprile et. al (XENON). Search for Coherent Elastic Scattering of Solar ^{8}B Neutrinos in the XENON1T Dark Matter Experiment. Phys. Rev. Lett., 126:091301, 2021. doi: 10.1103/PhysRevLett.126.091301.

E. Aprile et. al (XENON). Dark Matter Search Results from a One Ton-Year Exposure of XENON1T. Phys. Rev. Lett., 121(11):111302, 2018. doi: 10.1103/Phys- RevLett.121.111302.

Estimators

SoOP 24

- Estimators are test statistics we wish to use to understand some physical parameter.
- The ideal estimator has zero bias ($E(\hat{\theta}) = \theta$) and as low variance as possible
  - And most importantly, that it is *consistent*— that it converges to the true value with increasing observations
- A simple method to construct an estimator is to compute the expected mean or higher moments of the distribution, and invert that expression
- *The maximum likelihood* will, in the limit of a large sample be ideal: it is consistent, and is asymptotically normally distributed with the minimal possible variance

$$\delta \log \mathscr{L}(\hat{\boldsymbol{\theta}}) / \delta \theta_j = 0;$$

From the earliest days of statistics, statisticians have begun their analysis by proposing a distribution for their observations, and then, perhaps with somewhat less enthusiasm, have checked whether this distribution is true

- Ralph B. D'Angostino and Michael A. Stephens, Goodness-of-Fit Techniques, 1986

Goodness-of-Fit (GOF)

SoOP 24

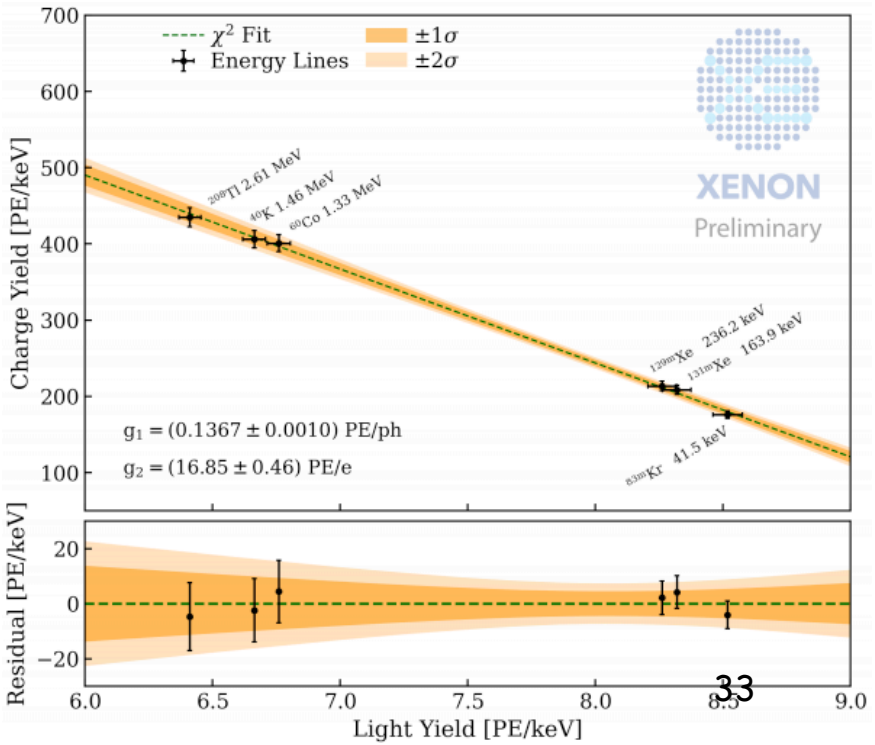
- The conclusions we draw from our data depends on our statistical model
- Unless we have a strong physical argument for a certain distribution to hold (e.g. Poission for counting events) we should probe the correctness of our model or fit to the data
- Unlike other hypothesis testing, GOF tests must consider every possible other alternative as a competitor to the model we test
- The conclusion to a failed goodness-of-fit test may therefore sometimes just be "worry more"

*"I am powerful. And I am only the most lowly gatekeeper. But from room to room stand gatekeepers, each more **powerful** than the other. I can't endure even one glimpse of the third."*

$$\chi^2 = \sum (x_i - E(X_i))^2 / \sigma_i^2$$

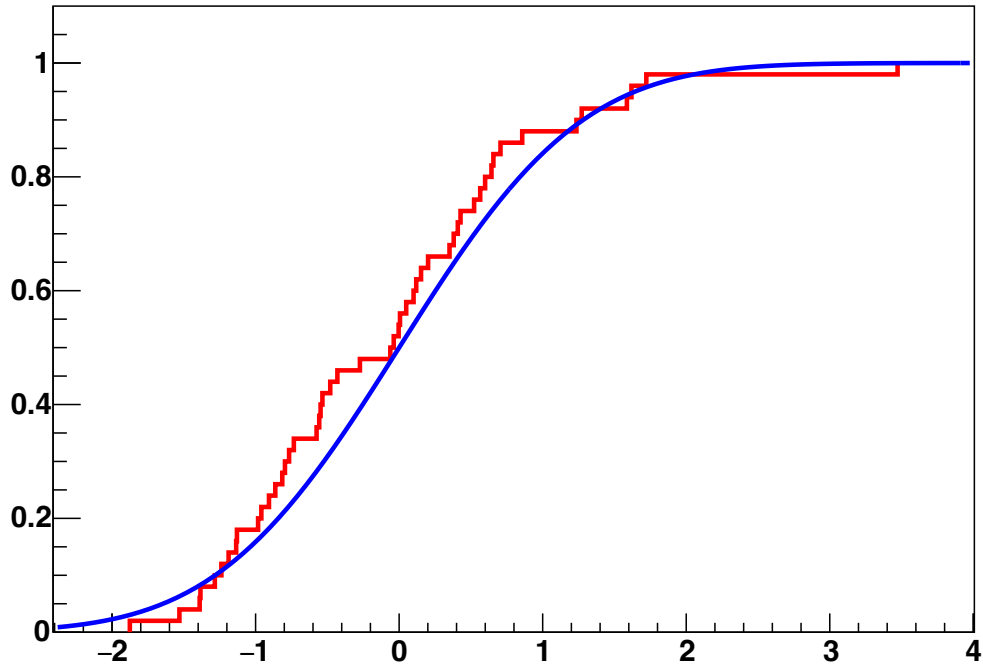
- The sum of ν standard normal-distributed numbers is χ^2_ν DOF-distributed
- Often encountered fitting curves
  - If there are errors in both x and y, you may transform it into an effective total error on y
- or histograms with large enough counts that they approach a Gaussian
- If one or more parameters are fit, the effective number of degrees of freedom is reduced accordingly (this assumes that the parameters are independent)

$\nu \approx$ number of observations - number of fitted parameters



KS and AD

- Kolmogorov-Smirnov and Anderson-Darling are two tests that rely on comparing the Empirical Distribution Function (the cumulative fraction of events) and the tested distribution
- Useful since no binning is assumed
- The KS test considers the maximal distance between the two, and manages to be *distribution-free*— the distribution of the test statistic does not depend on F
- Alternatives include the Cramér-von Mises test, which is also distribution-free and Anderson-Darling which is not



$$D_{KS} = \max |EDF(X) - F(X)|$$

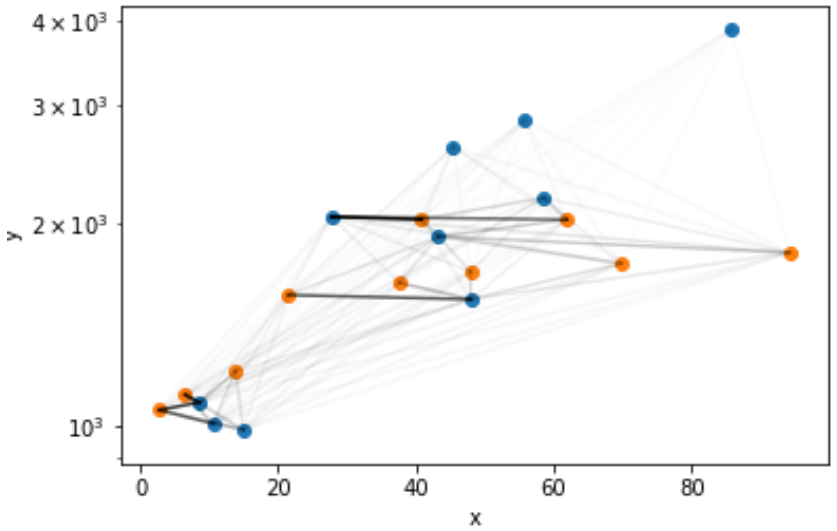
$$W^2 = \int_{-\infty}^{\infty} (EDF(X) - F(X))^2 f(X) dX$$

Other alternatives exist!

SoUP 24

https://arxiv.org/abs/hep-ex/020301

- Ideally, you should consider what sorts of mismodelling you are most worried about and choose goodness-of-fit tests to target these with the most *power*
  - Often, a projection on the dimension you care about will be a good start
- Some neat ideas exist to try to tackle high dimensionality by considering an analogue of electrostatic energy between point clouds
- One caution: the likelihood itself may seem tempting, but turns out to be a poor GOF test statistic



8. CONCLUSIONS

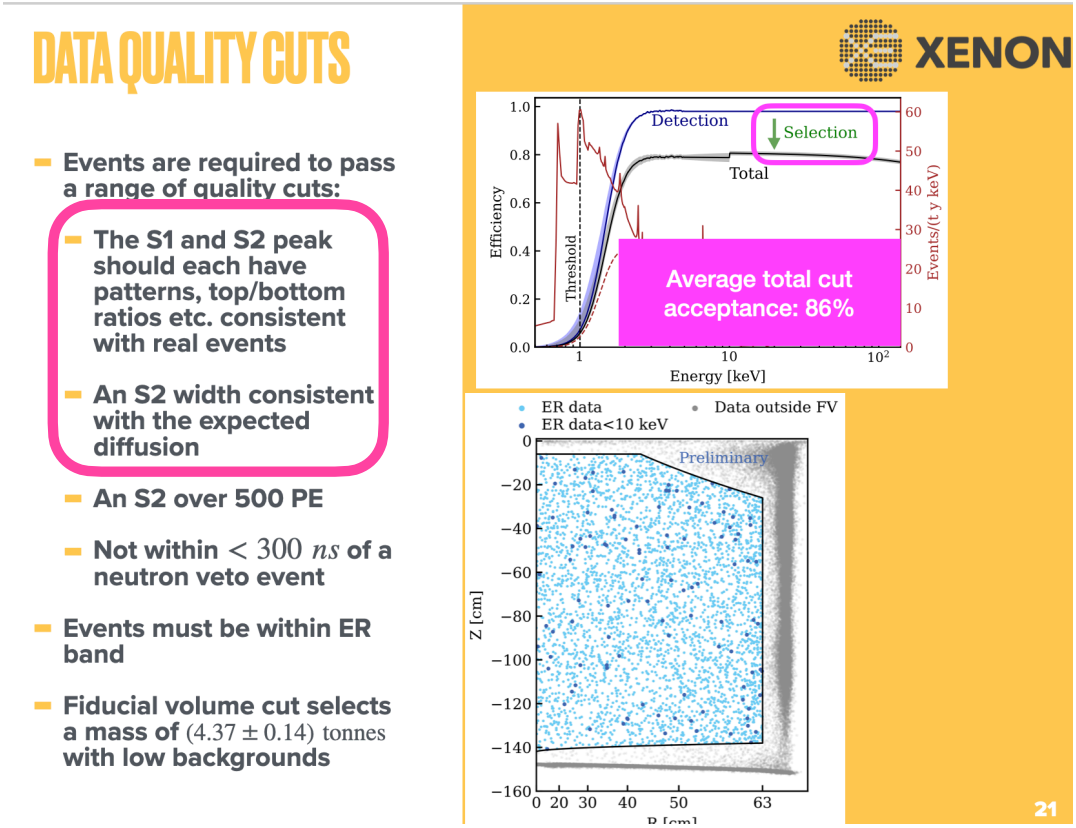
- This "g.o.f." method is fatally flawed in the unbinned case. Don't use it. Complain when you see it used.
- With fixed p.d.f.'s, the method suffers from test bias, and is not invariant with respect to change

https://arxiv.org/abs/physics/0310167

Cuts are often GOF tests!

SoUP 24

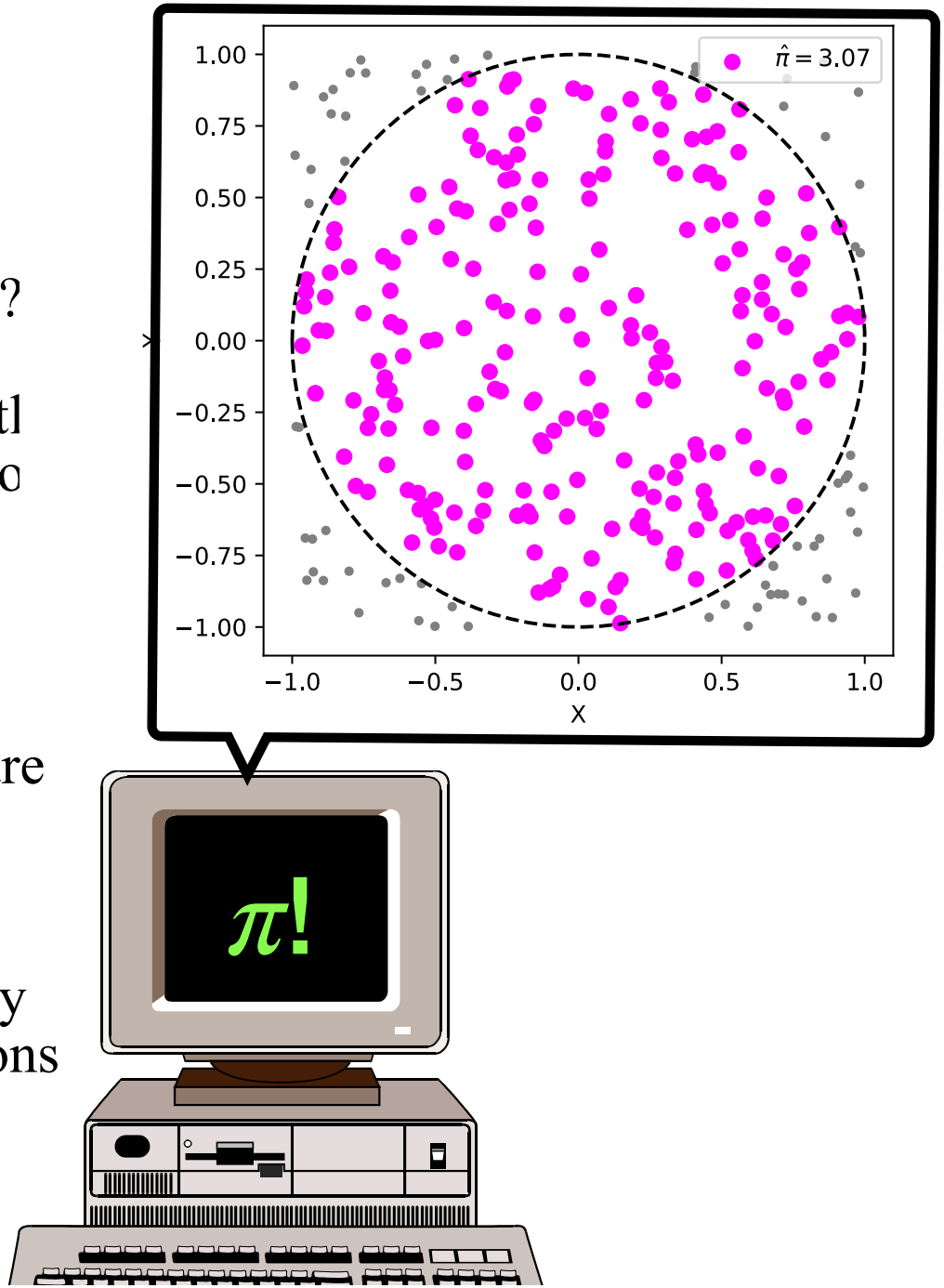
- Many event selections may be considered goodness-of-fit tests— asking whether they are compatible with coming from a signal
- Others are more standard hypothesis tests, if the background model is specified
- But we often define some cuts first and only model what remains!



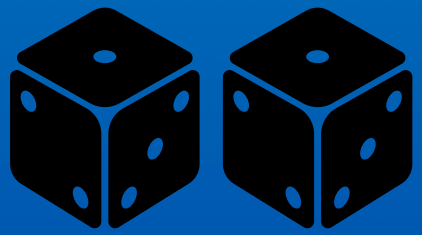
What are Toy Monte-Carlo methods?

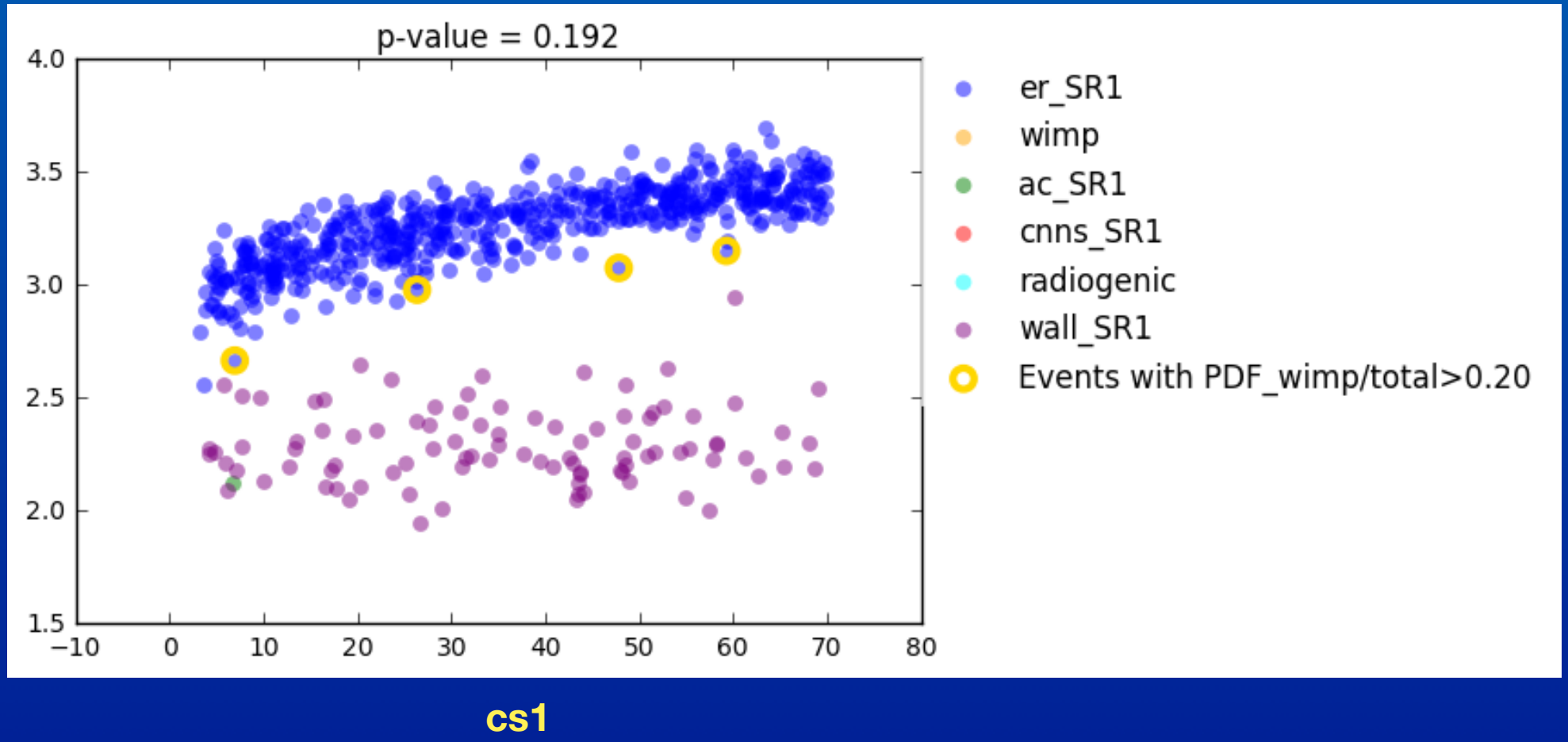
SoOP 24

- What is the area of a circle?
- Or, often equally importantl — what is the distribution o our estimate for π, or any other test statistic you can imagine?
  - In this case you can figure out the distribution,
- But for many more complicated cases, you may either rely on approximations or simulated results

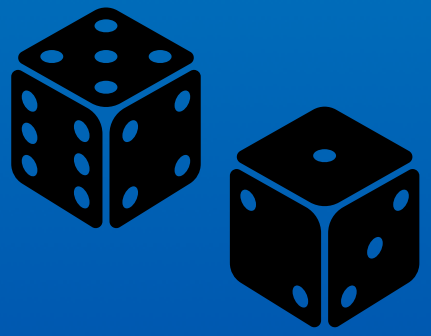


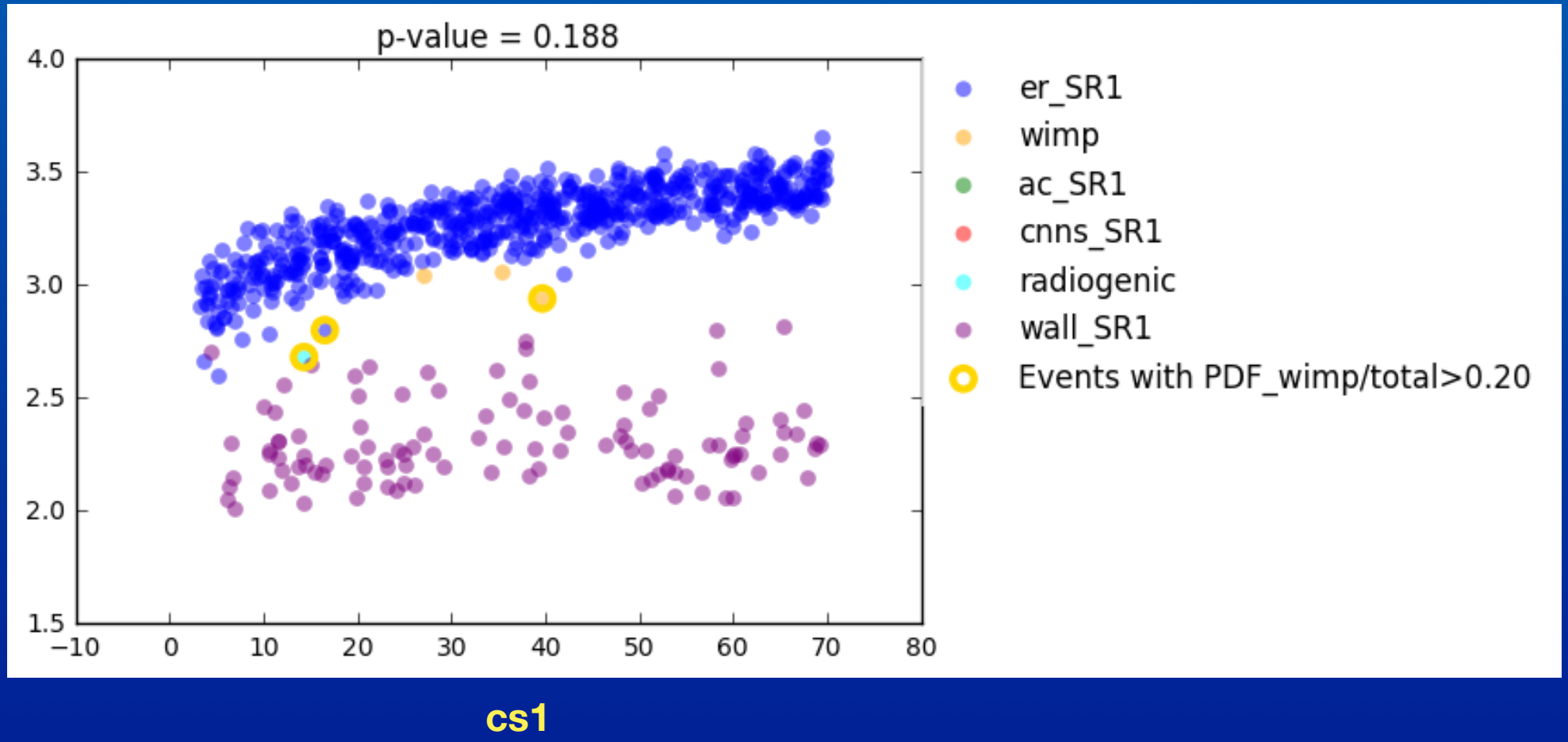
Searching for rare events is a matter of luck:



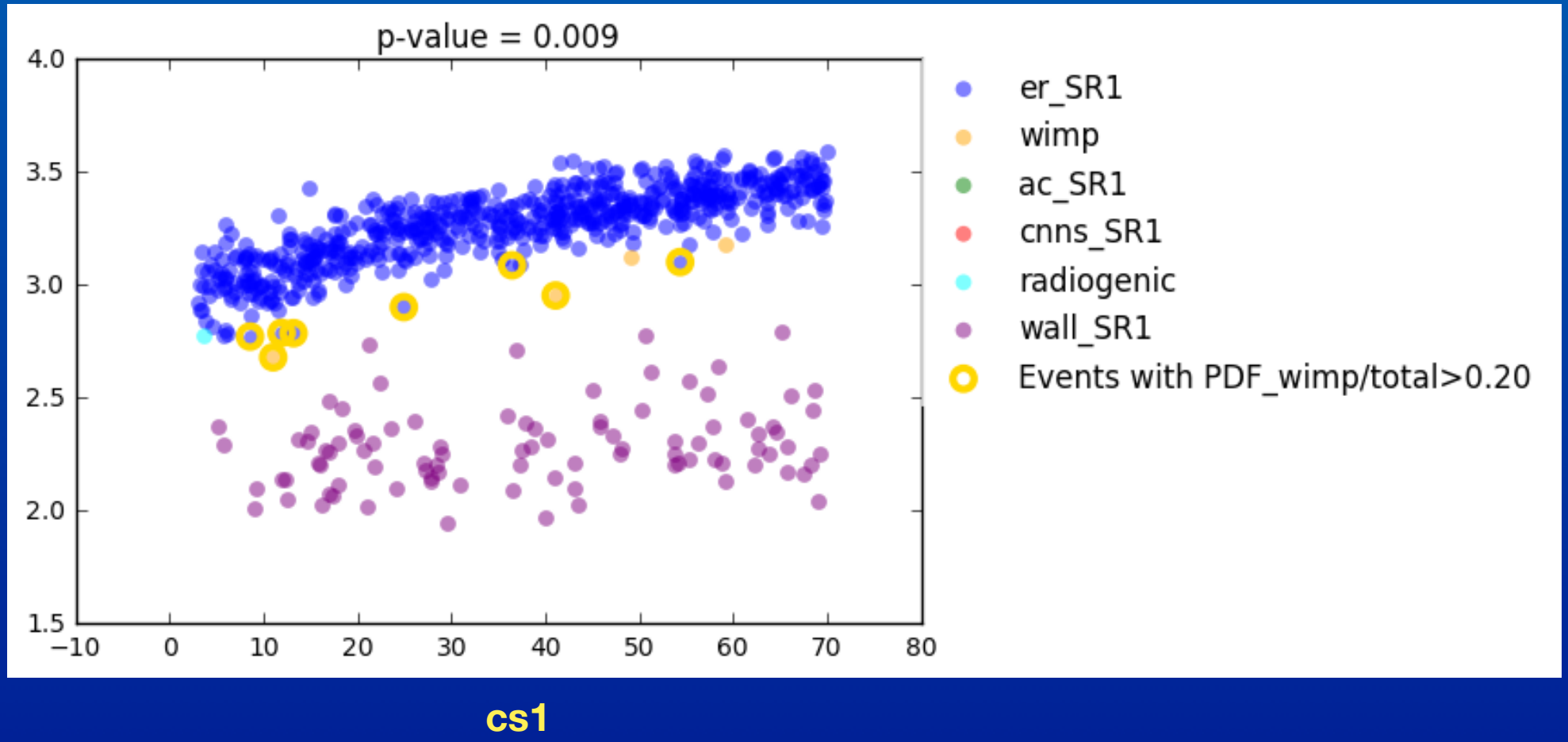


Searching for rare events is a matter of luck:





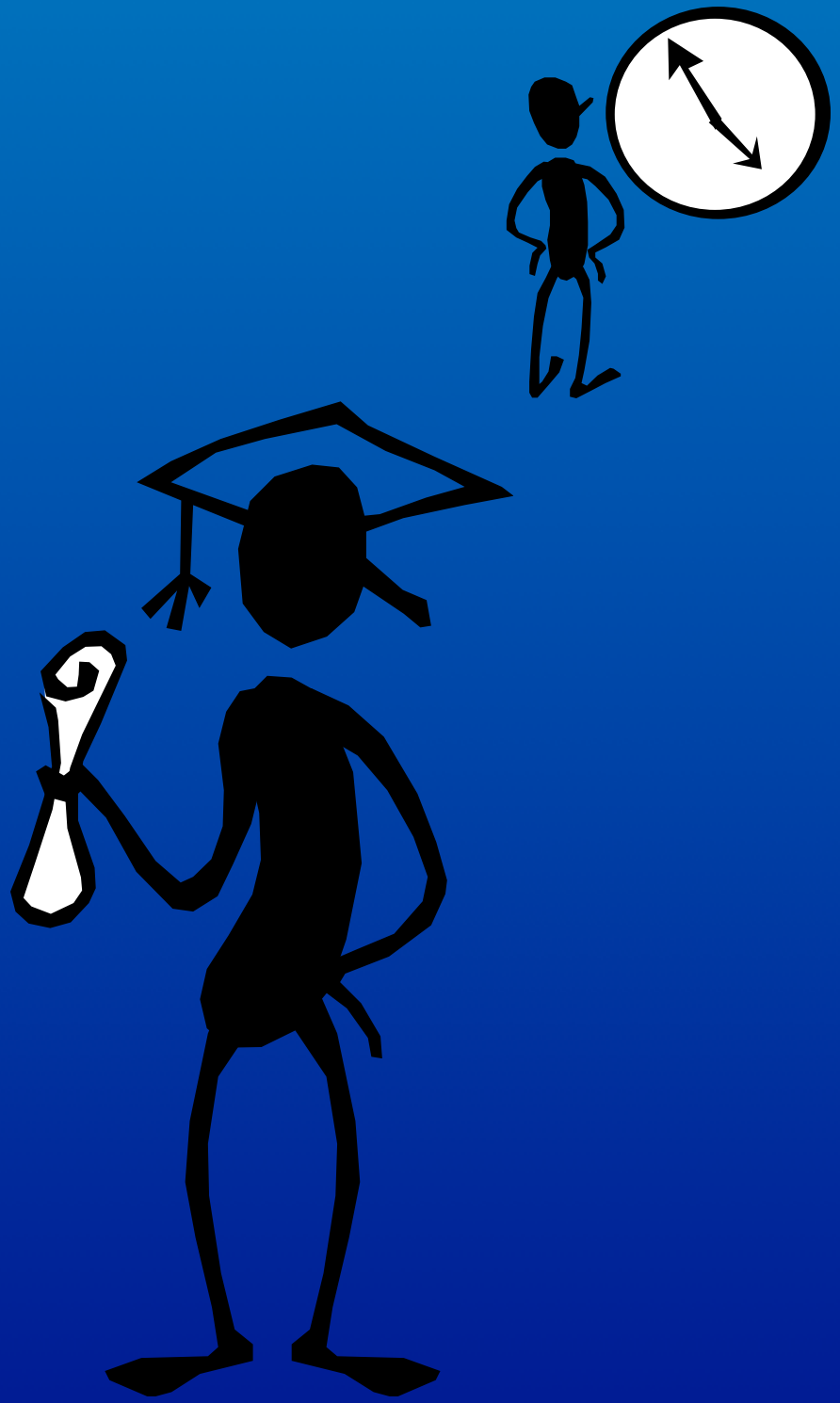
Searching for rare events is a matter of luck:



Questions?

Introduction to statistics

- We model our observations with a statistical model, usually in terms of probability distributions.
- We choose test statistics that distil the information we wish to learn from the data
- and often formulate questions in terms of hypothesis tests— given the data, should we favour one or the other?
- A particularly important hypothesis test is whether your data agrees with the distribution you use!



Knut Dundas Morå
fysikk@dundasmora.no, he/him

School of Underground Physics at Bertinoro

Statistics and Inference

for rare event searches

What is a statistical model?
Does it describe your data?
What kinds of conclusions can we draw?

24

Summary of first topic

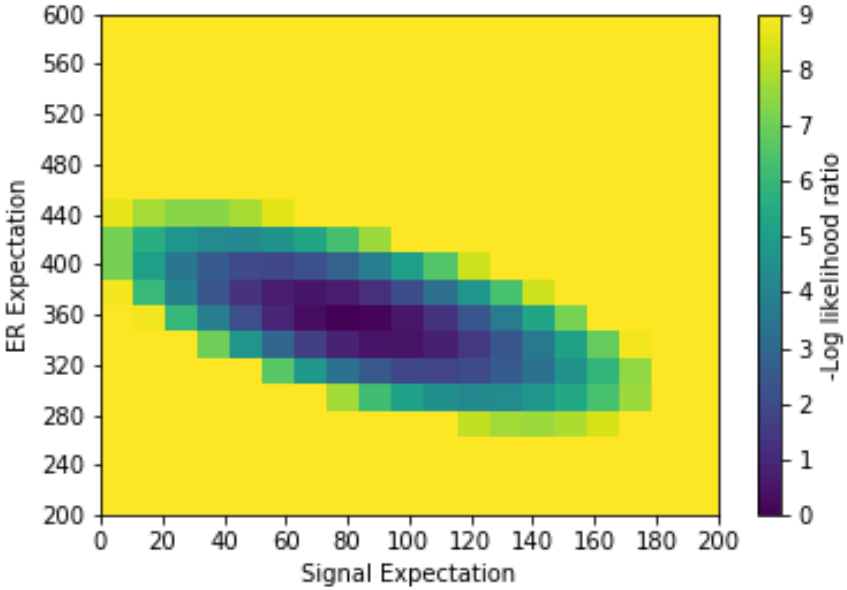
- We model our observations with a statistical model, usually in terms of probability distributions.
- We choose test statistics that distil the information we wish to learn from the data
- and often formulate questions in terms of hypothesis tests— given the data, should we favour one or the other?
- A particularly important hypothesis test is whether your data agrees with the distribution you use!

For today

- Example analyses
- Profile Likelihood
- Asymptotic distributions
- Look-Elsewhere effect

Profiling Likelihoods & nuisance parameters

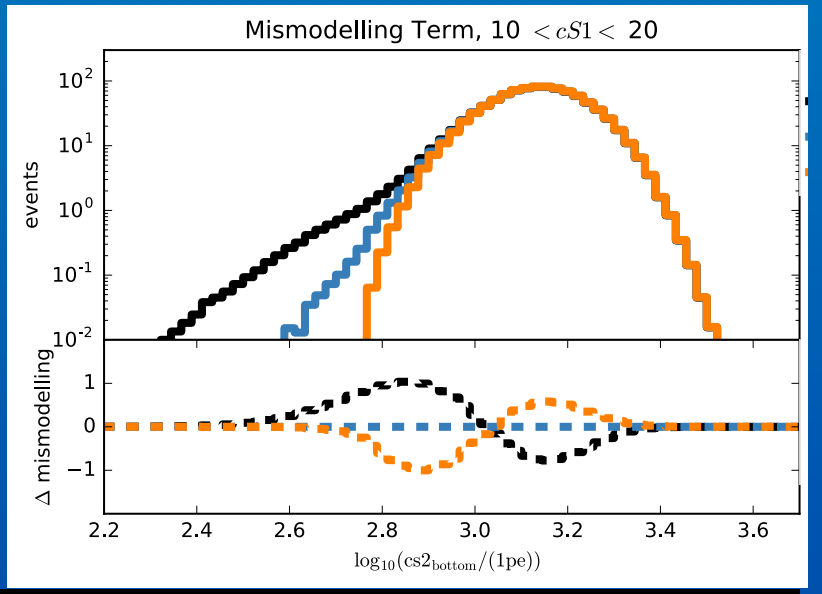
- We seldom have completely specified hypotheses
- Our background and signal models have uncertainties, parameterised by nuisance parameters (θ)— you'll see some examples in the next slides.
- The global best fit we denote with $\hat{s}, \hat{\theta}$
- However, we also want to test other s — for example s=0 for discovery significance or a range of s for confidence intervals.
- In these cases, we set the other nuisance parameters to their conditional best-fit $\hat{\hat{\theta}}$.



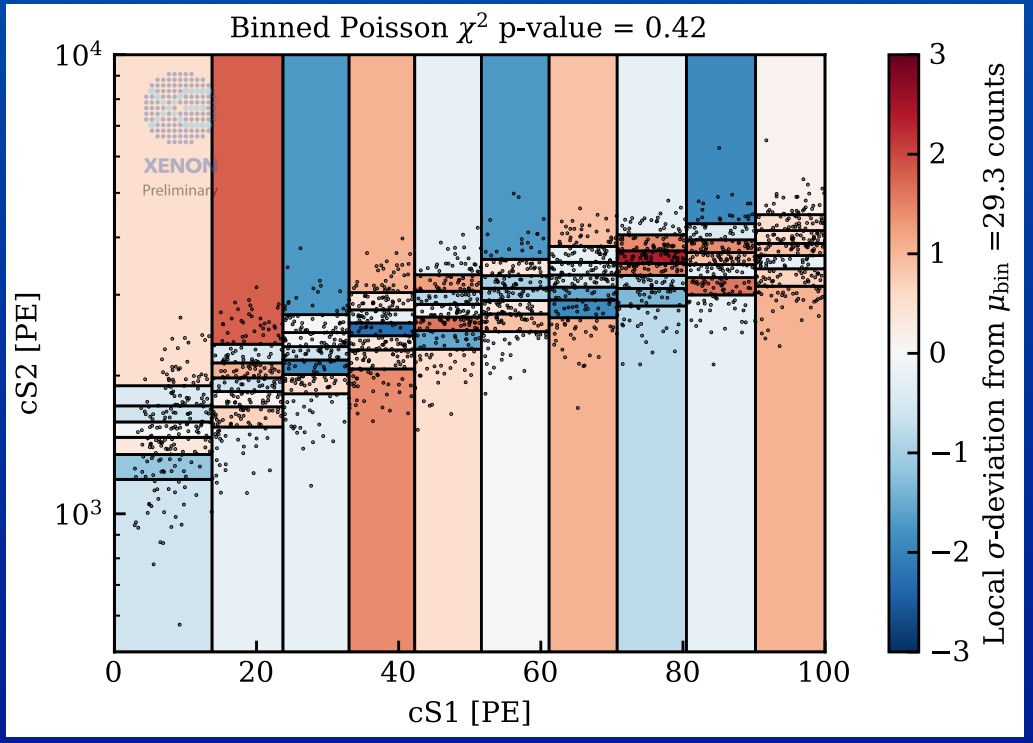
$$\delta \log \mathscr{L}(s, \hat{\hat{\boldsymbol{\theta}}}) / \delta \theta_j = 0;$$

The likelihood relies on the model

- The validity of the inference relies on the underlying model
- The signal model may be quite forgiving— if an excess is 10-20 events, far tails are less significant
- Experiments typically include uncertainties on background rates, but not always on the distribution used.
- XENON1T added a "signal-like" background shape to its ER background model to lower the chance of overconstraining the model.
- For XENONnT, this was replaced by a more careful selection of nuisance parameter directions, and a stronger focus on pre-defined goodness-of-fit tests chosen for their power to discover mismodelling

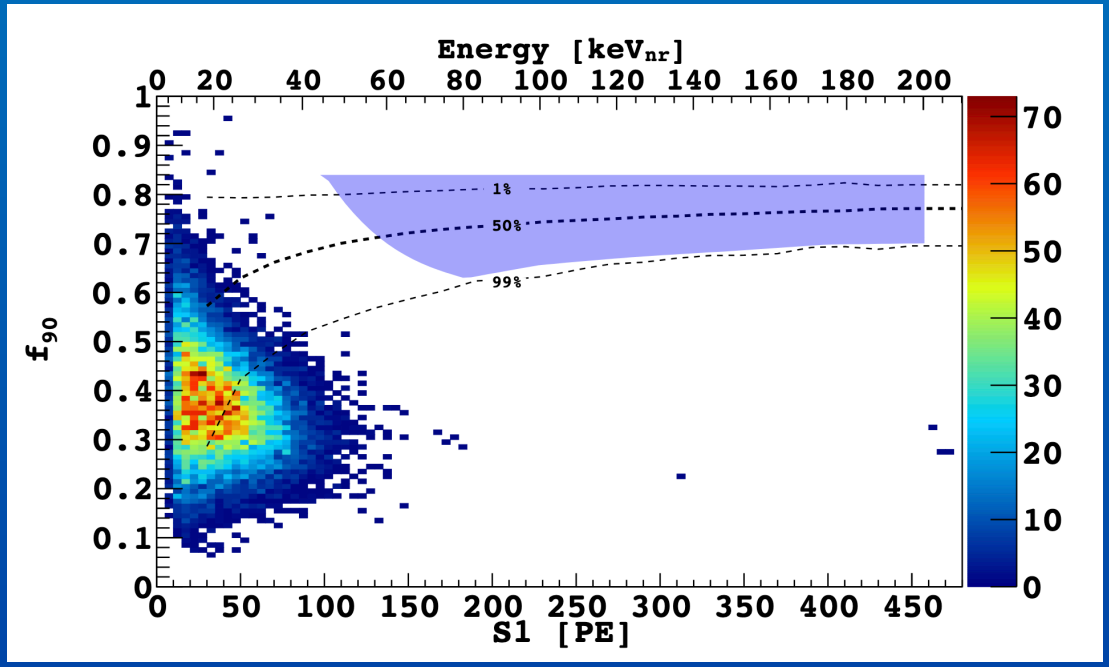


N. Priel et al. A model independent safeguard against background mismodeling for statistical inference. 2017(05):013–013, may 2017. doi: 10.1088/1475-7516/2017/05/013.



Counting Experiments

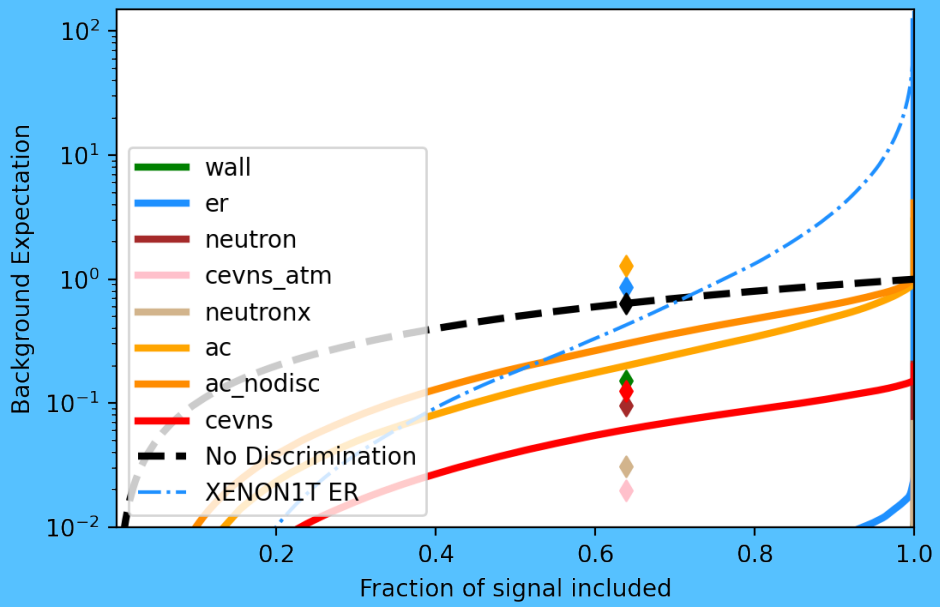
- "just" counting events— but the estimate of the background rate and acceptance can be as complicated as anything
- If there is no signal/background overlap *or* complete overlap, this may be the optimal sensitivity
- Otherwise, it might still be a worthwhile compromise if you're worried about whether you can model your background correctly

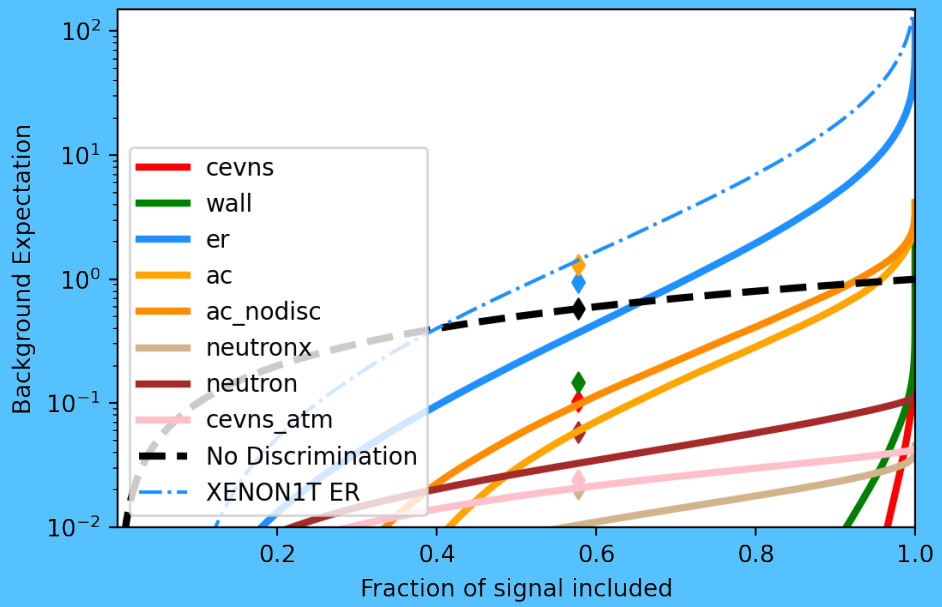


DarkSide-50 532-day https://arxiv.org/pdf/1802.07198

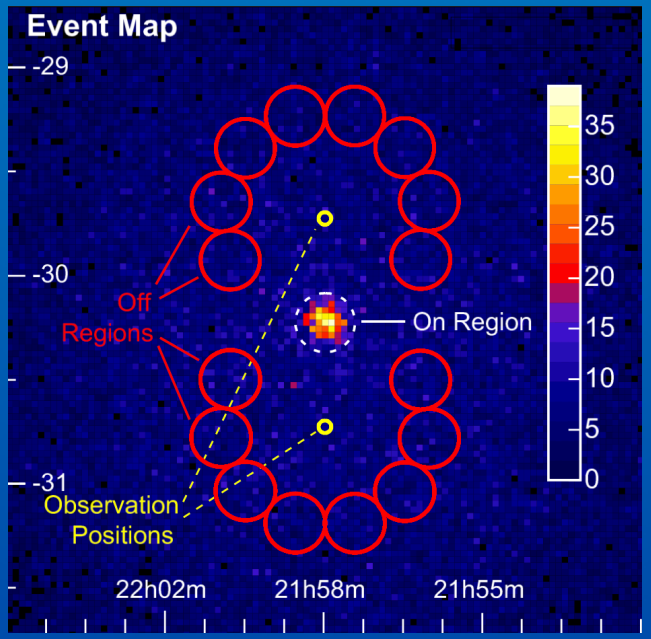
$$\mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) = \text{Poisson}(N_{\text{sci}} \,|\, \mu_b(\vec{\theta}_b) + \mu_s(s, \vec{\theta}_s, \vec{\theta}_b))$$

However, shapes often matter





On-Off likelihoods

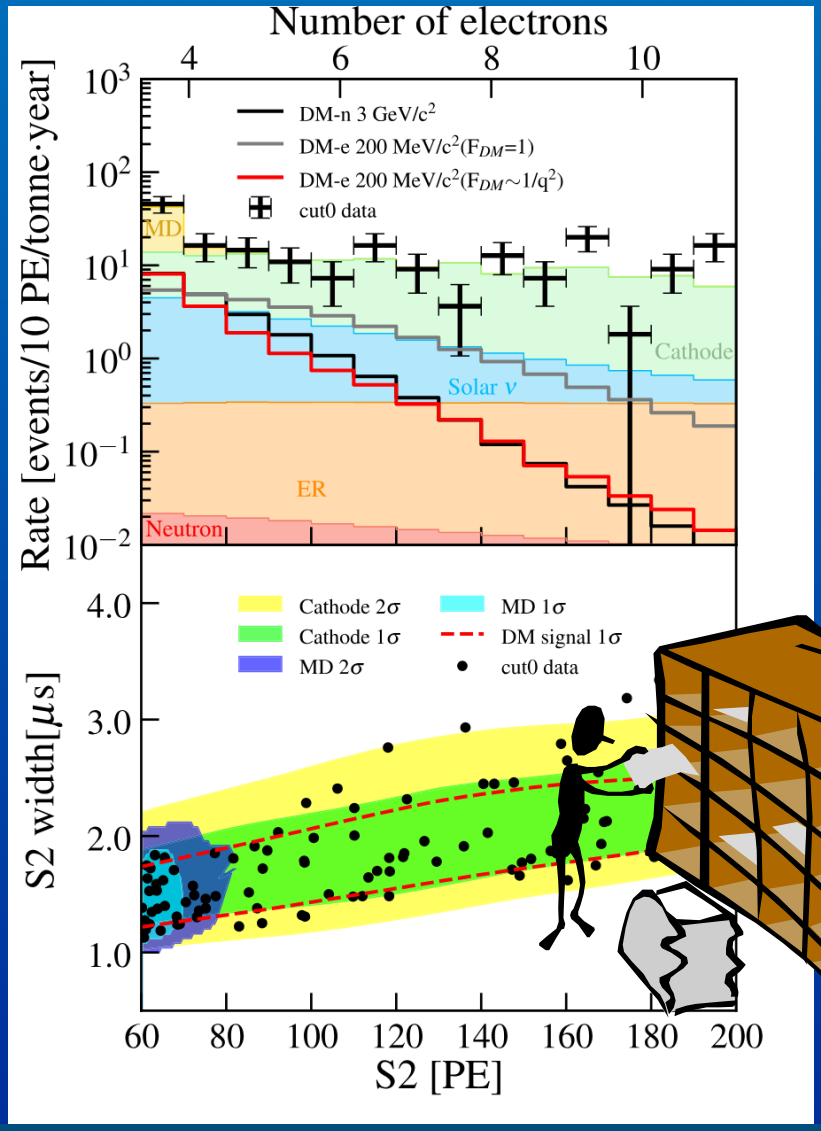


- WIMP searches rarely get to turn off their signal completely
- Directional dark matter searches and some axion searches, on the other hand can take representative data in a no/low signal and high signal state
- Also common in indirect detection

$$\mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) = \text{Poisson}(N_{\text{sci}} \mid \mu_b(\vec{\theta}_b) + \mu_s(s, \vec{\theta}_s, \vec{\theta}_b)) \times \text{Poisson}(N_{\text{cal}} \mid \alpha \times \mu_b(\vec{\theta}_b))$$

Binned Likelihood

- With more than ~ 5 events in each bin, you can use computationally efficient methods to compute test statistic distributions
- Eases visualisation and goodness-of-fit
- And simpler to share results
- Minimal sensitivity loss if the bin width is small compared to the detector resolution

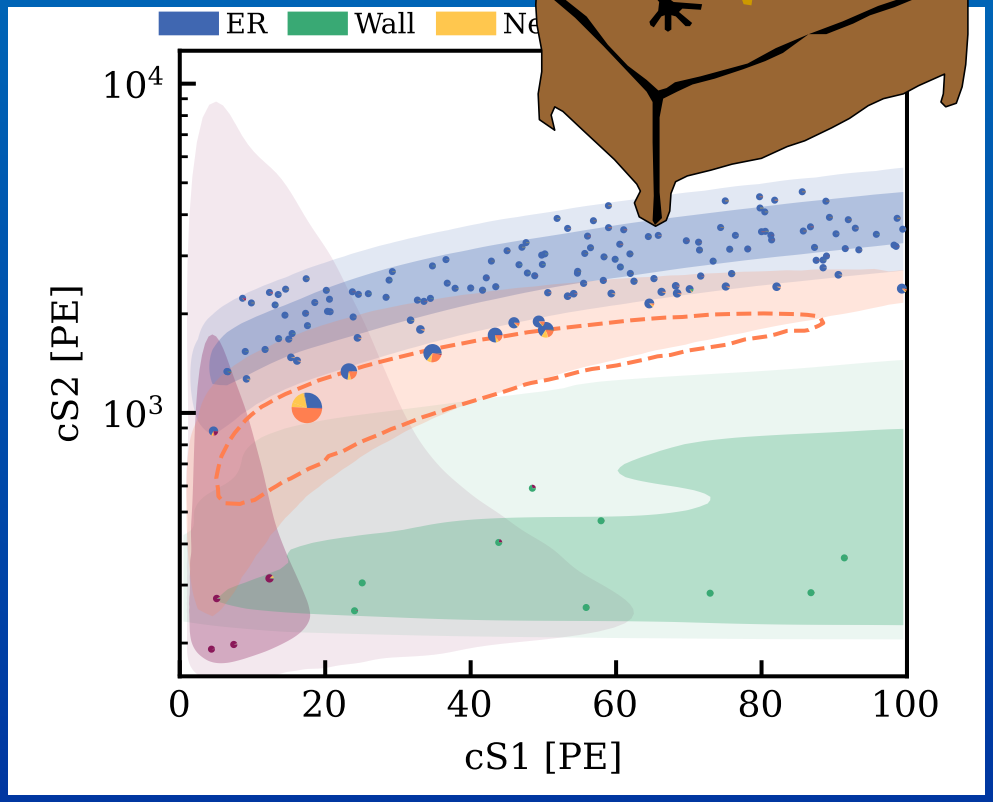


PandaX ionisation-only search, https://arxiv.org/abs/2212.10067

$$\mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) = \prod_{i=1}^{N_s} \left[\text{Poisson}(N_i \,|\, \mu_{b,i}(\vec{\theta}_b) + \mu_{s,i}(s, \vec{\theta}_s, \vec{\theta}_b)) \right]$$

Unbinned (extended) likelihood

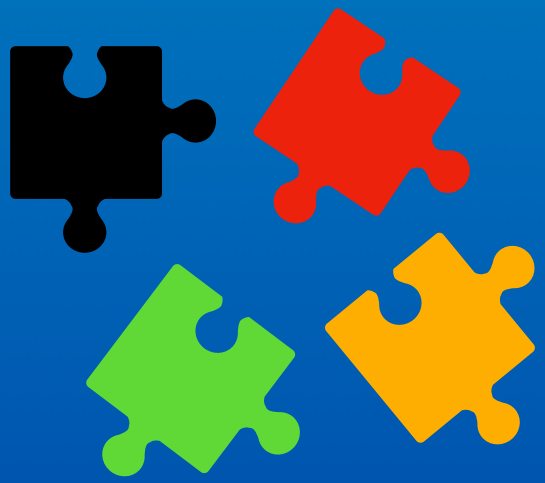
- If the events are too few to fill bins, the unbinned likelihood promises the best performance
- Might still have to rely on binned methods for goodness-of-fit
- if you rely on Monte Carlo methods to generate distributions, that can require a lot of statistics and be harder to validate



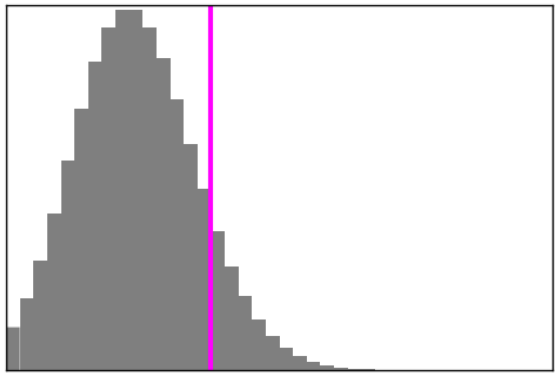
XENONnT first WIMP search

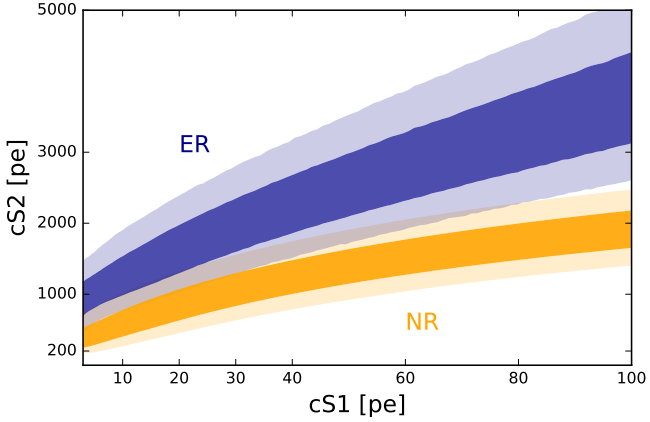
$$\mathscr{L}_{\text{sci}}(s, \overrightarrow{\theta_s}, \vec{\theta}_b) = \text{Poisson}(N_{\text{sci}} \,|\, \mu_b(\vec{\theta}_b) + \mu_s(s, \overrightarrow{\theta_s}, \vec{\theta}_b)) \times \prod_{i=1}^{N_s} \left[\frac{\mu_s}{\mu_s + \mu_b} f_s(\vec{x}_i \,|\, s, \overrightarrow{\theta_s}, \vec{\theta}_b) + \frac{\mu_b}{\mu_s + \mu_b} f_b(\vec{x}_i \,|\, \vec{\theta}_b) \right]$$

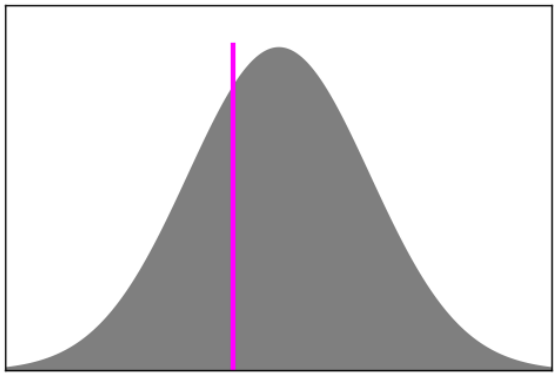
Likelihoods can be composed



$$\mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b)_{\textbf{Science run}} = \mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) \times \mathscr{L}_{\text{cal}}(\vec{\theta}_b) \times \mathscr{L}_{\text{anc}}(\vec{\theta}_b)$$





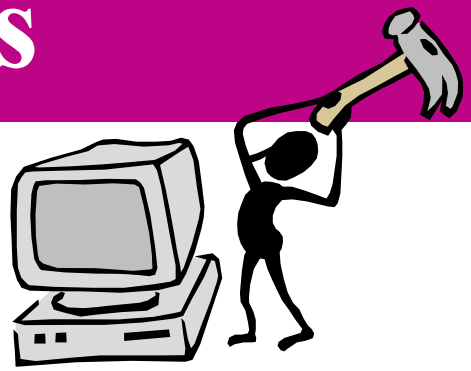


$$\mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b)_{\textbf{tot}} = \mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b)_{\textbf{tot}} \times \mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b)_{\textbf{tot}} \times \mathscr{L}_{\textbf{shared}}(\theta)$$

Asymptotic Distributions

SoUP 24

A massive shortcut if you're careful/lucky



- The log-likelihood for a number of gaussian-distributed numbers has the same form as the χ^2 -formula (Wilks' theorem)
- It turns out that if a set of conditions that are quite often^{TM} fulfilled, the distribution of the likelihood ratio converges to a χ^2 -distribution with some number of free parameters
- This can massively simplify your computations, and so it is worth to look through in detail

$$q(s) = -2 \cdot \log\left(\frac{\mathscr{L}(s, \hat{\hat{s}})}{\mathscr{L}(\hat{s}, \hat{s})}\right)$$

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

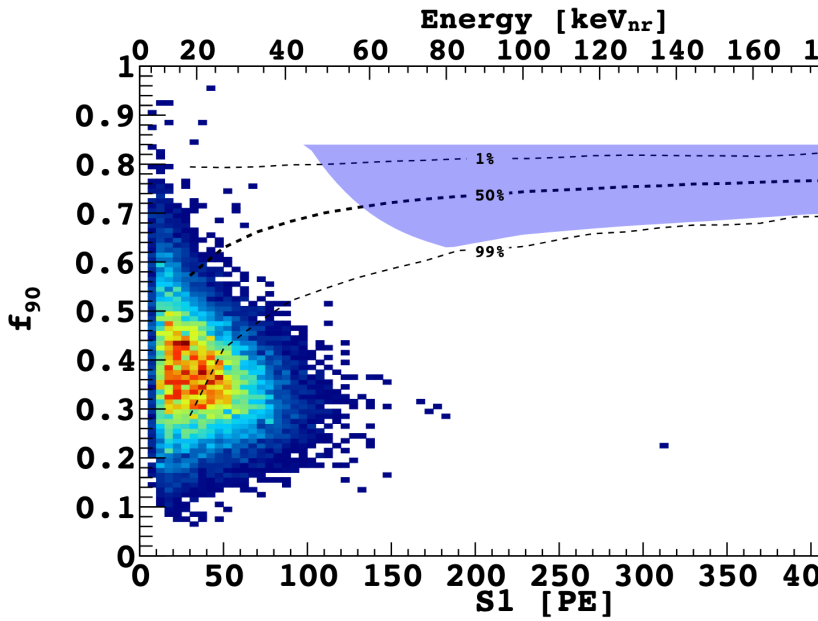
IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

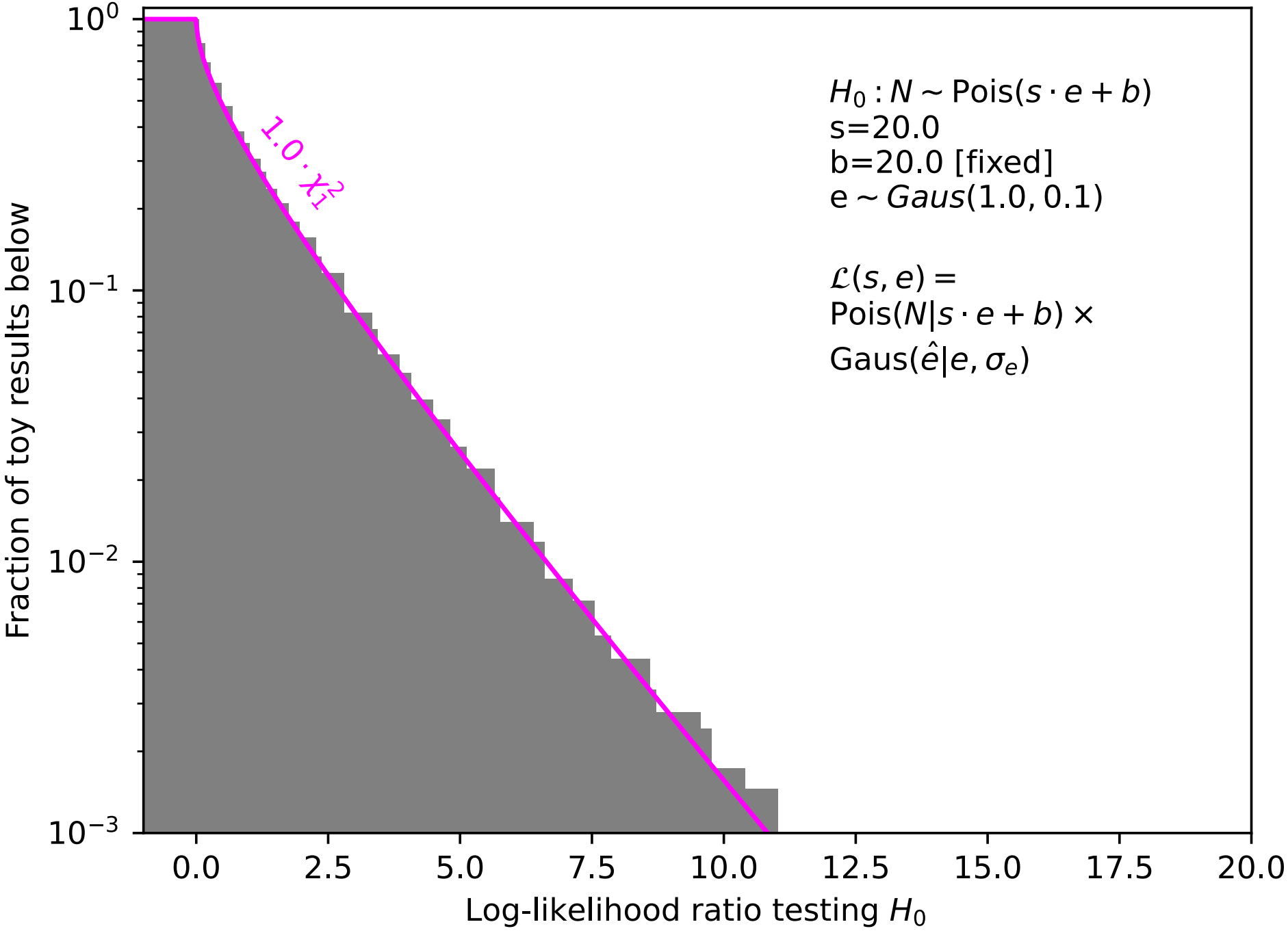
https://arxiv.org/abs/1911.10237

Asymptotic Distributions



- As our example: the profile log-likelihood ratio test for a counting experiment with a known background but uncertain efficiency
- Parameters:
  - Signal s
  - efficiency e
- Fixed, known parameters:
  - Background expectation *b*
  - efficiency uncertainty σ_e
- Data:
  - Number of events N
  - efficiency estimate e_{meas}

$$(L)(s,e) = \text{Pois}(N \,|\, s \cdot e + b) \times \text{Gaus}(e_{\text{meas}} \,|\, e, \sigma_e)$$


10^0
10^{-1}
10^{-2}
10^{-3}
Fraction of toy results below
$1.0 \cdot \chi^2_1$
$H_0 : N \sim \text{Pois}(s \cdot e + b)$
s=20.0
b=20.0 [fixed]
e $\sim$ *Gaus*(1.0, 0.1)
$\mathcal{L}(s, e) =$
$\text{Pois}(N|s \cdot e + b) \times$
$\text{Gaus}(\hat{e}|e, \sigma_e)$
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Log-likelihood ratio testing H_0

Asymptotic Distributions

What does "sufficiently data" mean?

- Wilks' theorem holds in the asymptotic case of infinite data, but convergence can often be quick:
  - Poisson counting with more than ca. 10 events
  - Gaussian measurements
- However, if you have an unbinned likelihood, the important consideration is *signal-like* background events— for example seen with lXe TPC searches

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

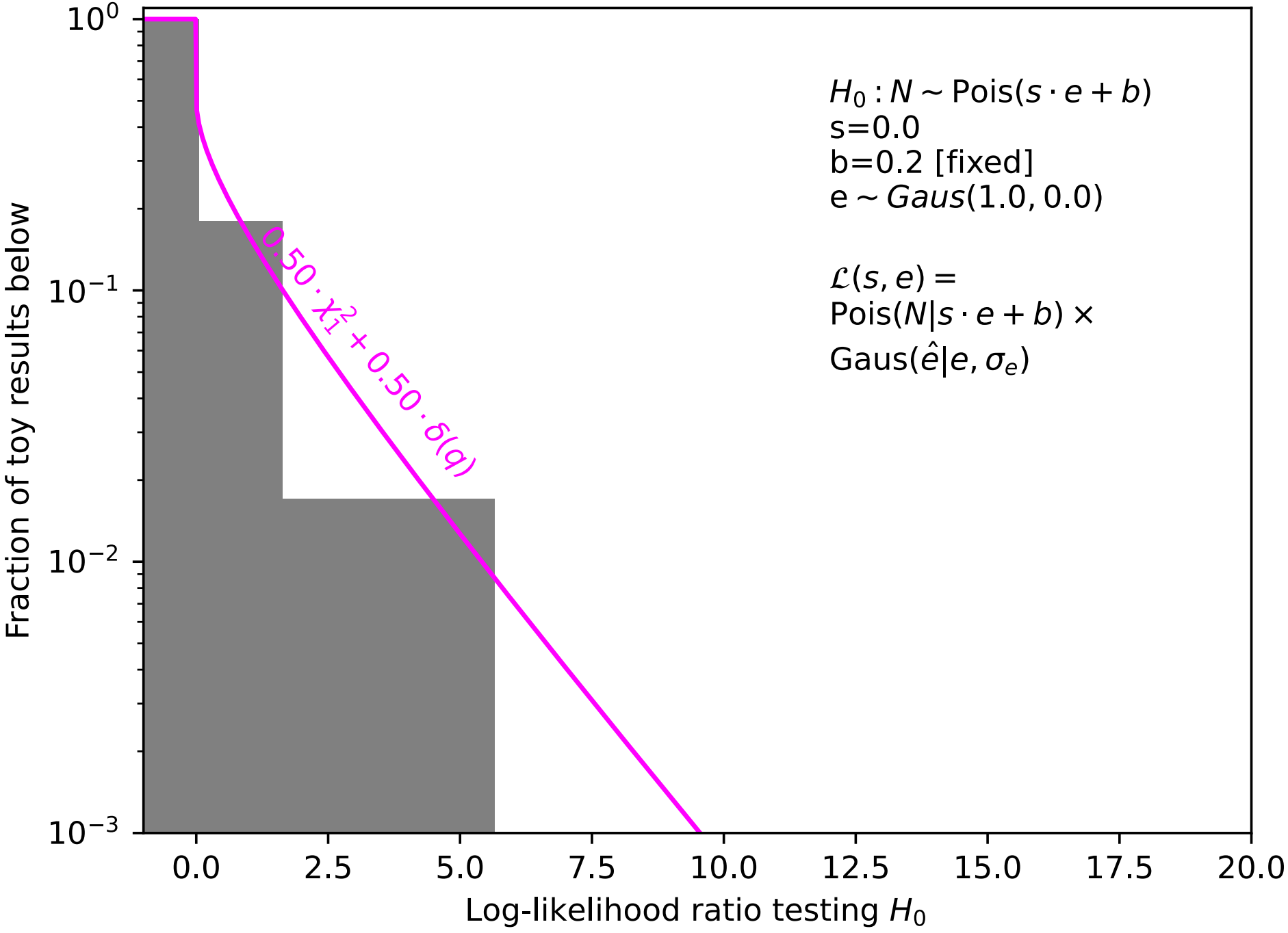
INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

https://arxiv.org/abs/1911.10237


10^0
10^{-1}
10^{-2}
10^{-3}
Fraction of toy results below
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Log-likelihood ratio testing H_0
$0.50 \cdot \chi^2_1 + 0.50 \cdot \delta(q)$
$H_0 : N \sim \text{Pois}(s \cdot e + b)$
s=0.0
b=0.2 [fixed]
e $\sim Gaus(1.0, 0.0)$
$\mathcal{L}(s, e) =$
$\text{Pois}(N|s \cdot e + b) \times$
$\text{Gaus}(\hat{e}|e, \sigma_e)$

Asymptotic Distributions

SoOP 24

What does "interior of the parameter space" mean?

- As a mental shortcut— if under your null or signal hypothesis, parameters sometimes or often goes to a physical boundary, it will not behave asymptotically
- This is very often the case e.g. if you're looking for a signal with expectation value ≥ 0
- If you are testing the hypothesis that the model that has the parameter *at the boundary*— for example that the signal is 0, you may be able to use *Chernoff's theorem* if all other conditions are met

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

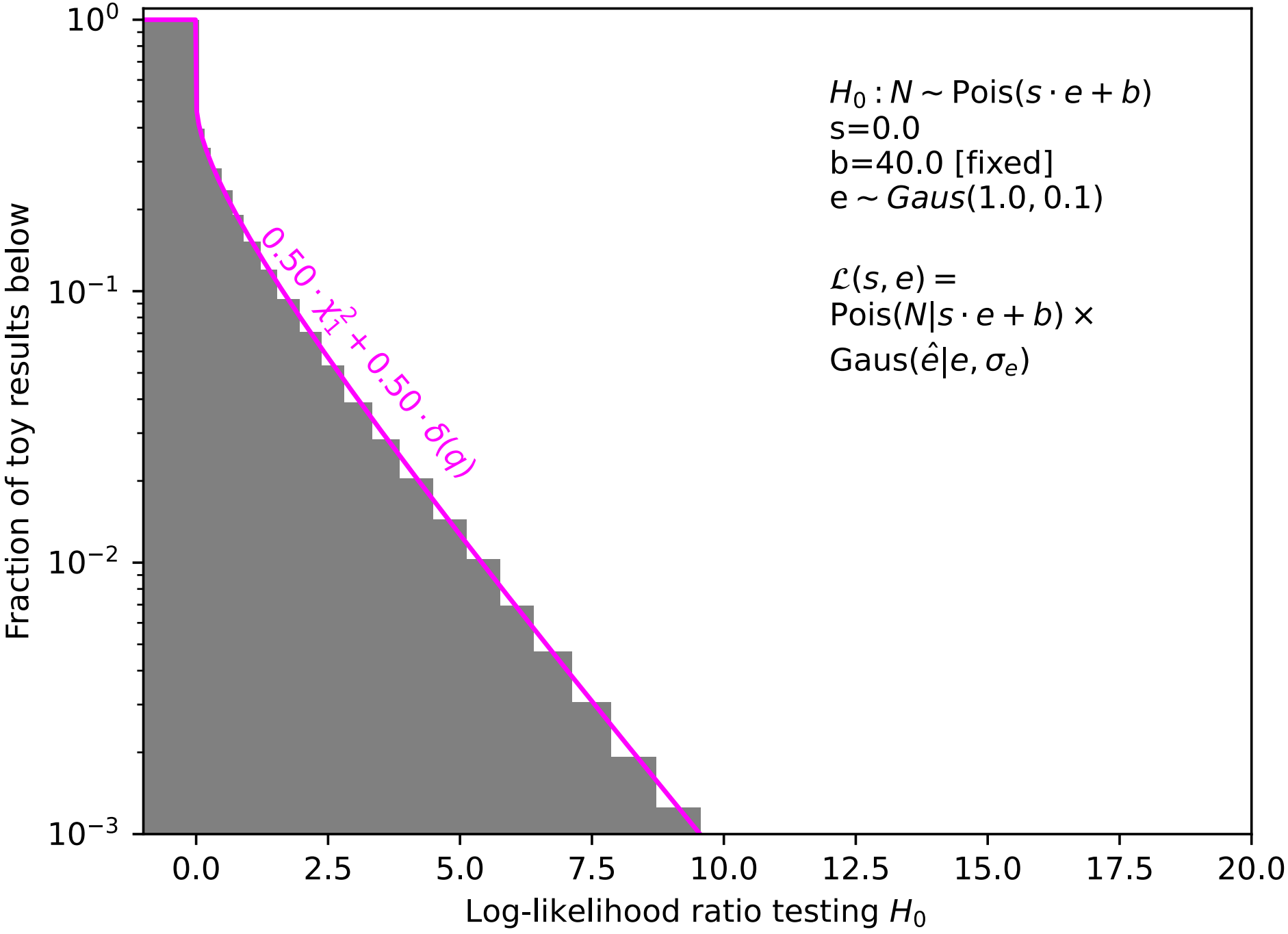
INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

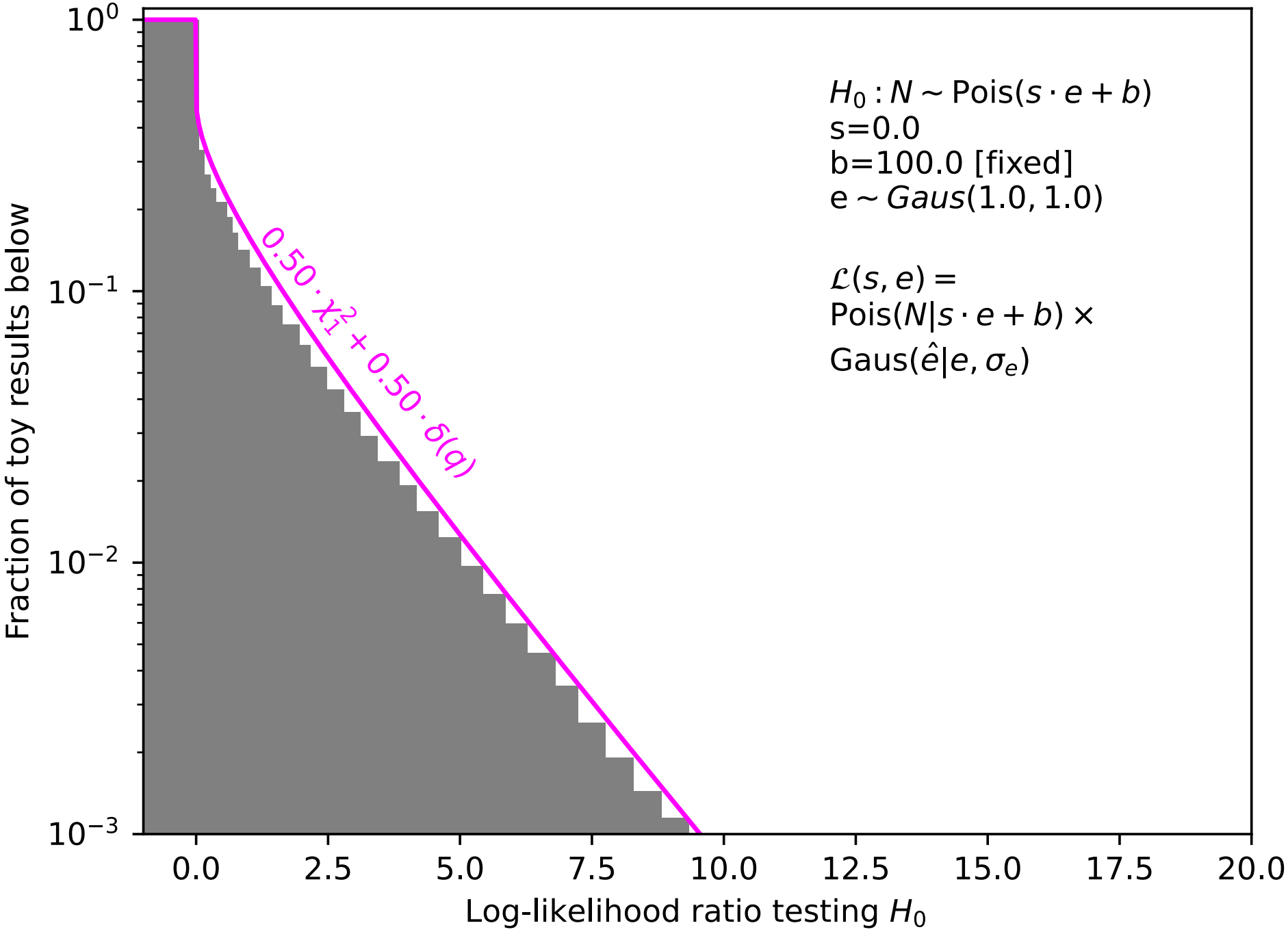
IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

$$f(q) \overset{\text{Chernoff}}{\approx} \frac{1}{2}\chi^2_{DOF=1} + \frac{1}{2}\delta(\hat{\mu})$$


10^0
10^{-1}
10^{-2}
10^{-3}
Fraction of toy results below
$0.50 \cdot \chi^2_1 + 0.50 \cdot \delta(q)$
$H_0 : N \sim \text{Pois}(s \cdot e + b)$
s=0.0
b=40.0 [fixed]
e $\sim$ *Gaus*(1.0, 0.1)
$\mathcal{L}(s, e) =$
$\text{Pois}(N|s \cdot e + b) \times$
$\text{Gaus}(\hat{e}|e, \sigma_e)$
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Log-likelihood ratio testing H_0


10^0
10^{-1}
10^{-2}
10^{-3}
Fraction of toy results below
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Log-likelihood ratio testing H_0
$0.50 \cdot \chi^2_1 + 0.50 \cdot \delta(q)$
$H_0 : N \sim \text{Pois}(s \cdot e + b)$
s=0.0
b=100.0 [fixed]
e $\sim$ *Gaus*(1.0, 1.0)
$\mathcal{L}(s, e) =$
$\text{Pois}(N|s \cdot e + b) \times$
$\text{Gaus}(\hat{e}|e, \sigma_e)$

Asymptotic Distributions

SoUP 24

What does it mean for parameters to be "identifiable"?

- If the model is degenerate for some parameter, the asymptotic approximation will not hold
- This is quite common in physics! When the signal strength is 0, the model does not depend on any other signal parameter
- This is another way of looking at the look-elsewhere effect, which we'll look at later

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

Asymptotic Distributions

SoOP 24

What does it mean for models to be "nested"?

- If the model tested is not a limit of the general hypothesis
  - Such as when testing between two disparate models
  - Or if your theory features a non-zero fixed signal you wish to test against the no-signal hypothesis
- You can always linearly add the two hypotheses' models together with a new parameter, but then you introduce Non-identifiability at the boundary!

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

https://arxiv.org/abs/1911.10237

Asymptotic Distributions

The models still need to be correct :(

- All our inference results are reliant on the true model being somewhere in our model space!
- However, we should be cognisant that this is never guaranteed
- If you have a mismodelling you are concerned about, you should test how much it can affect your results— you might well find that your method is robust to it, or you can add model uncertainties to represent this
- Another way to increase robustness is to make your model simpler— a counting experiment makes fewer assumptions on the energy spectrum than if you include the energy information

Necessary conditions for Wilks' theorem

ASYMPTOTIC: Sufficient data is observed.

INTERIOR: Only values of μ and θ which are far from the boundaries of their parameter space are admitted.

IDENTIFIABLE: Different values of the parameters specify distinct models.

NESTED: H_0 is a limiting case of H_1, e.g. with some parameter fixed to a sub-range of the entire parameter space.

CORRECT: The true model is specified either under H_0 or under H_1.

https://arxiv.org/abs/1911.10237

Asymptotic Distributions

Examples of when they may be used

- Any gaussian-distributed measurements
  - Including histograms with high bin counts
- unbinned likelihoods with significant signal-like backgrounds
- Most common extra consideration is taking care of the parameter boundaries
- The below paper presents some cases:

$$q_{\text{discovery}} = \begin{cases} q(0) \text{ if } 0 \leq \hat{\mu} \\ 0 \text{ else} \end{cases}$$

$$q_{\text{upper limit}}(\mu) = \begin{cases} q(\mu) \text{ if } \hat{\mu} \leq \mu \\ 0 \text{ else} \end{cases}$$

$$q(\mu)_{\text{unified}} = \begin{cases} -2 \cdot \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})} \text{ if } 0 \leq \hat{\mu} \\ -2 \cdot \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(0, \hat{\boldsymbol{\theta}}_{\mu=0})} \text{ else} \end{cases}$$

Note that these three can be seen as the same test statistic if you always restrict $\hat{\mu}, \mu$ to be positive!

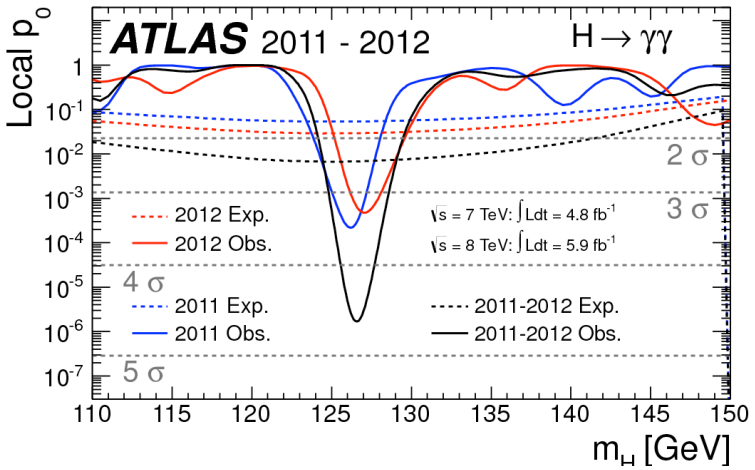
https://arxiv.org/abs/1007.1727

The Look Elsewhere Effect

SoUP 24

AKA trial factor AKA non-identifiable signal parameters

- Your probability to roll 6 on a dice increases the more dice you get to roll
- Similarly, if your experiment tests several signals, they will increase their chance to see unusual effects just by chance
- Separate between "local" significance— the probability that one single signal model tests fluctuates to some significance
- and "global" significance— the probability that *any* test fluctuates to that extent



The Look Elsewhere Effect

SoUP 24

Uncorrelated tests are simple

- A binomial process- what is the probability to get a 3-sigma deviation or more (say)
- 3 sigma (local) is 0.0027
- the probability to see a 3-sigma effect in 10 trials is 0.027, or equivalent to 2.2 sigma
- In the limit of $p->0$, you can just divide your local p-value by the number of trials

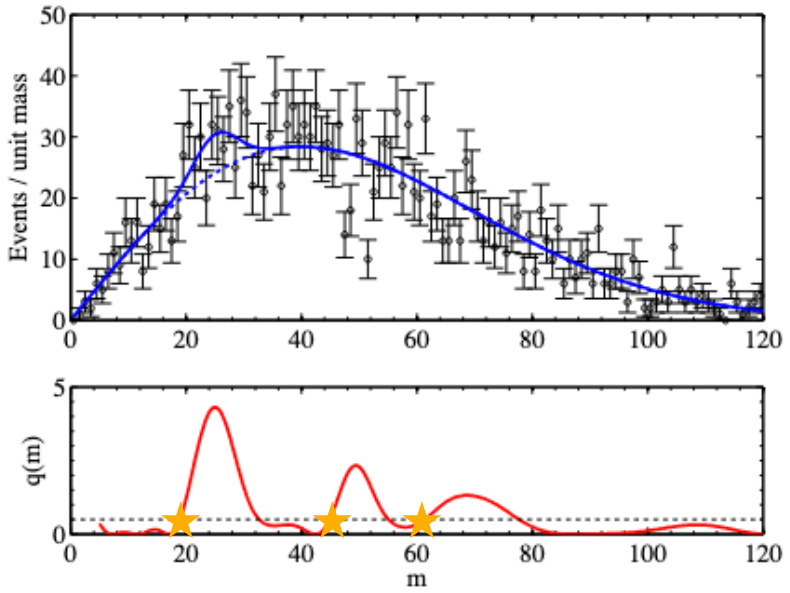
$$P(n|p,N) = \binom{N}{n} p^n \cdot (1.-p)^{N-n}$$

$$1 - \binom{N}{0} p^0 \cdot (1.-p)^N \sim_{p \ll 1} N \cdot p$$

The Look Elsewhere Effect

Correlations need Monte-Carlo or smart tricks

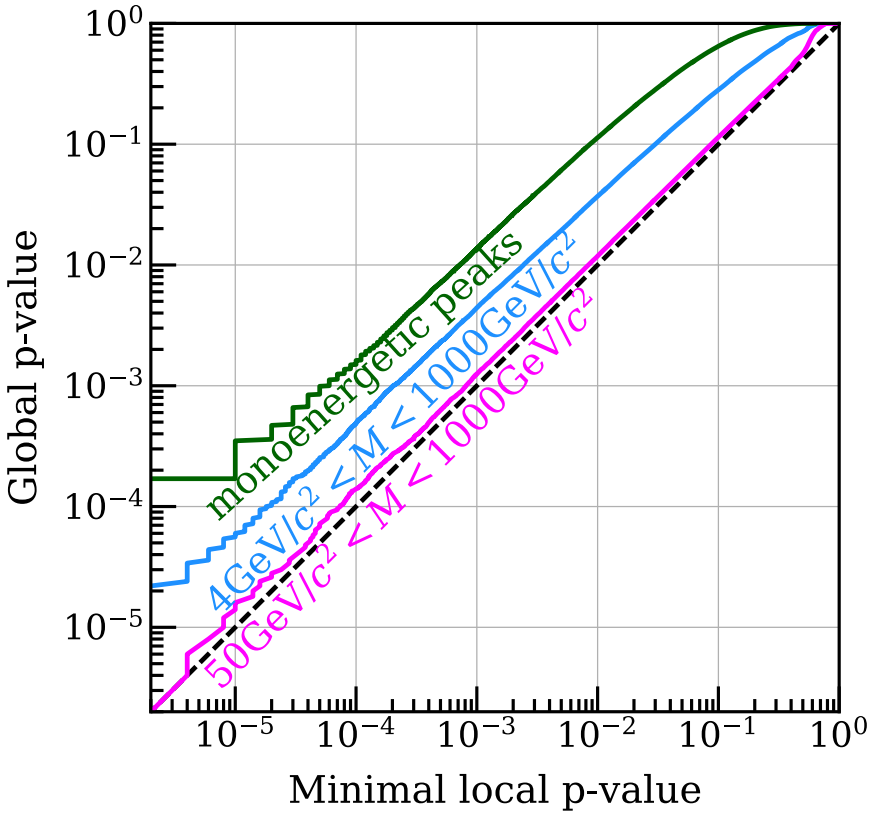
- However, in many cases the signals are not uncorrelated— for example, a peak search will be correlated with around its energy resolution
- One method is to use Toy Monte-Carlo methods — powerful, but painful if your significance is high!
  - However, if you do have a significant result, your collaboration will often be willing to expend significant computing power :)
- If your test statistic follows an *asymptotic distribution* otherwise, you may be able to use a clever method by Gross&Vitells that estimate the effective number of trials by counting how many upwards fluctuations you have ("up crossings")

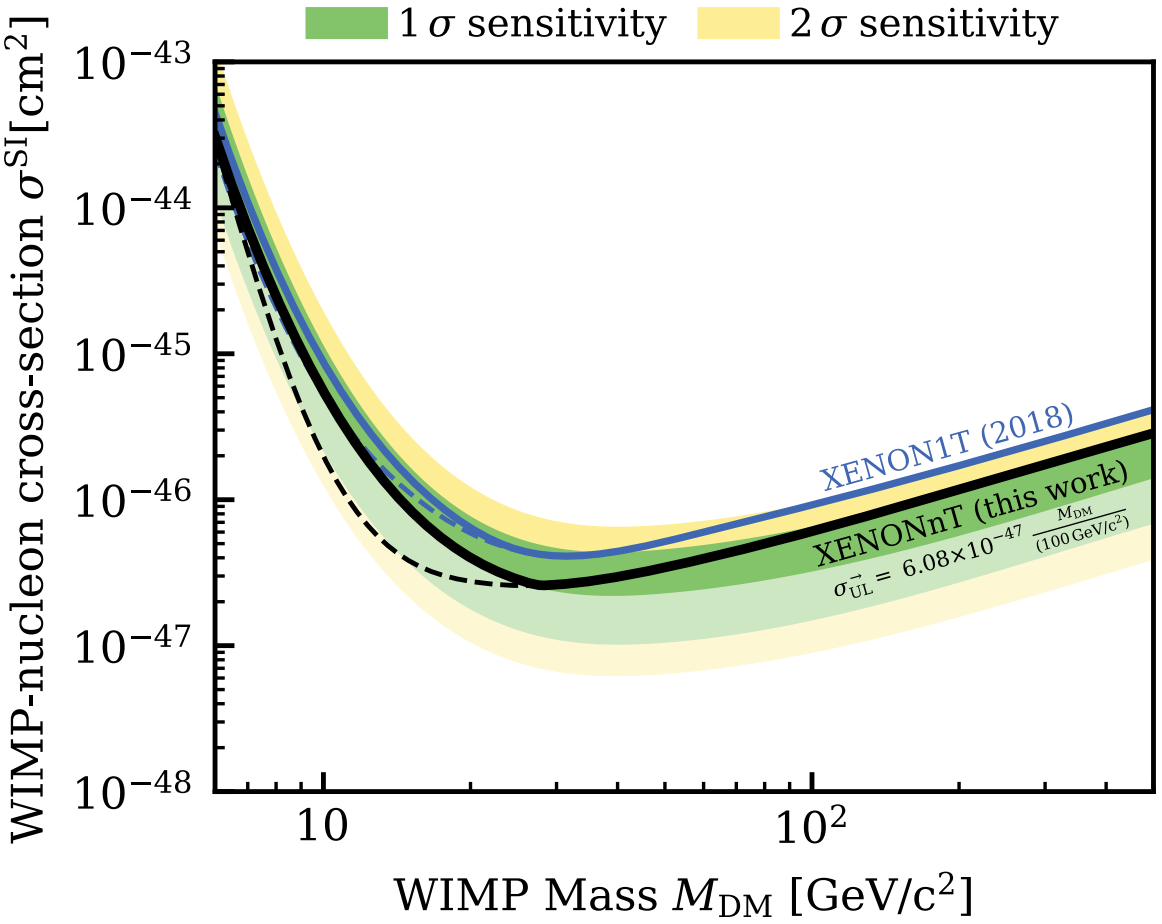


https://arxiv.org/abs/1005.1891

The Look Elsewhere Effect

The trial factor might sometimes be rather small:

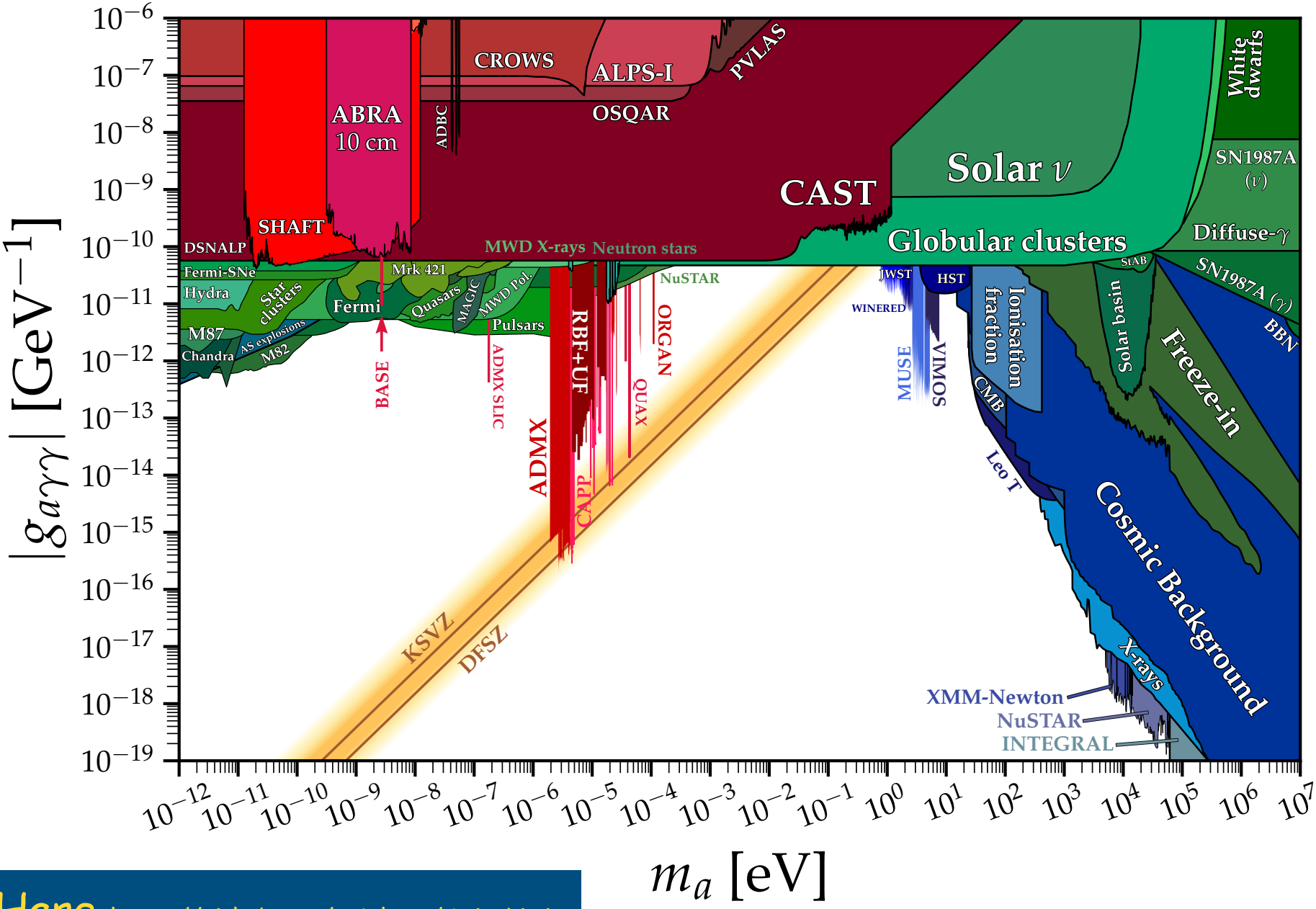




The Look Elsewhere Effect

SoUP 24

And sometimes enormous

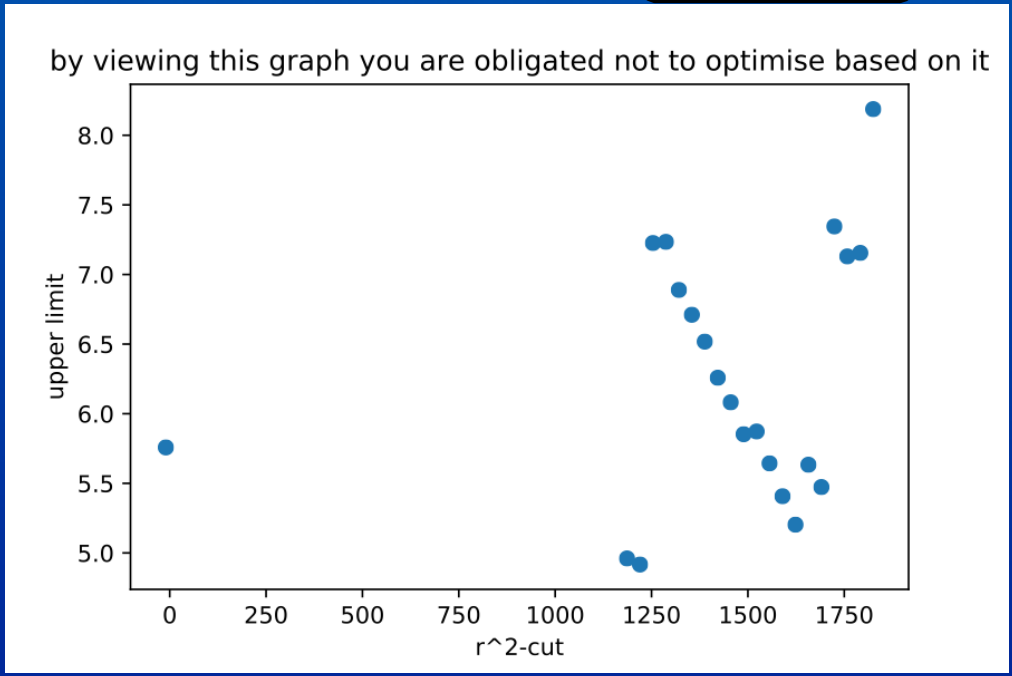


O'Hare https://github.com/cajohare/AxionLimits

Experimenter bias is a danger with few events

- With few events the effect can be drastic if you chance something in your analysis—the plot shows the 60% change in limit available to you between the best post-unblinding and the worst post-unblinding radial cut.
  - This is a necessary consequence of making your analysis sensitive to few events!
- Further, with only some hundreds of events, and many variables, every event may well be an outlier in some space

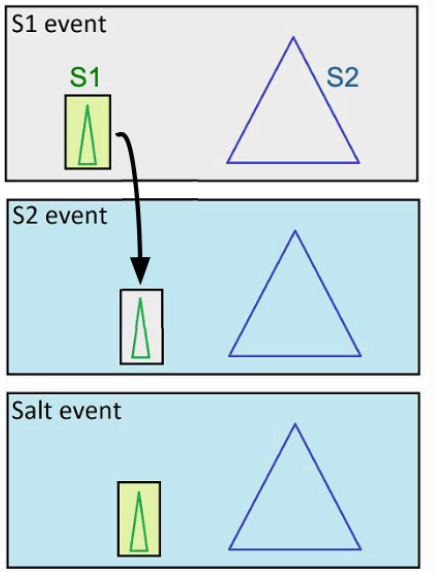
Homeopathic poison — the fewer events the greater danger



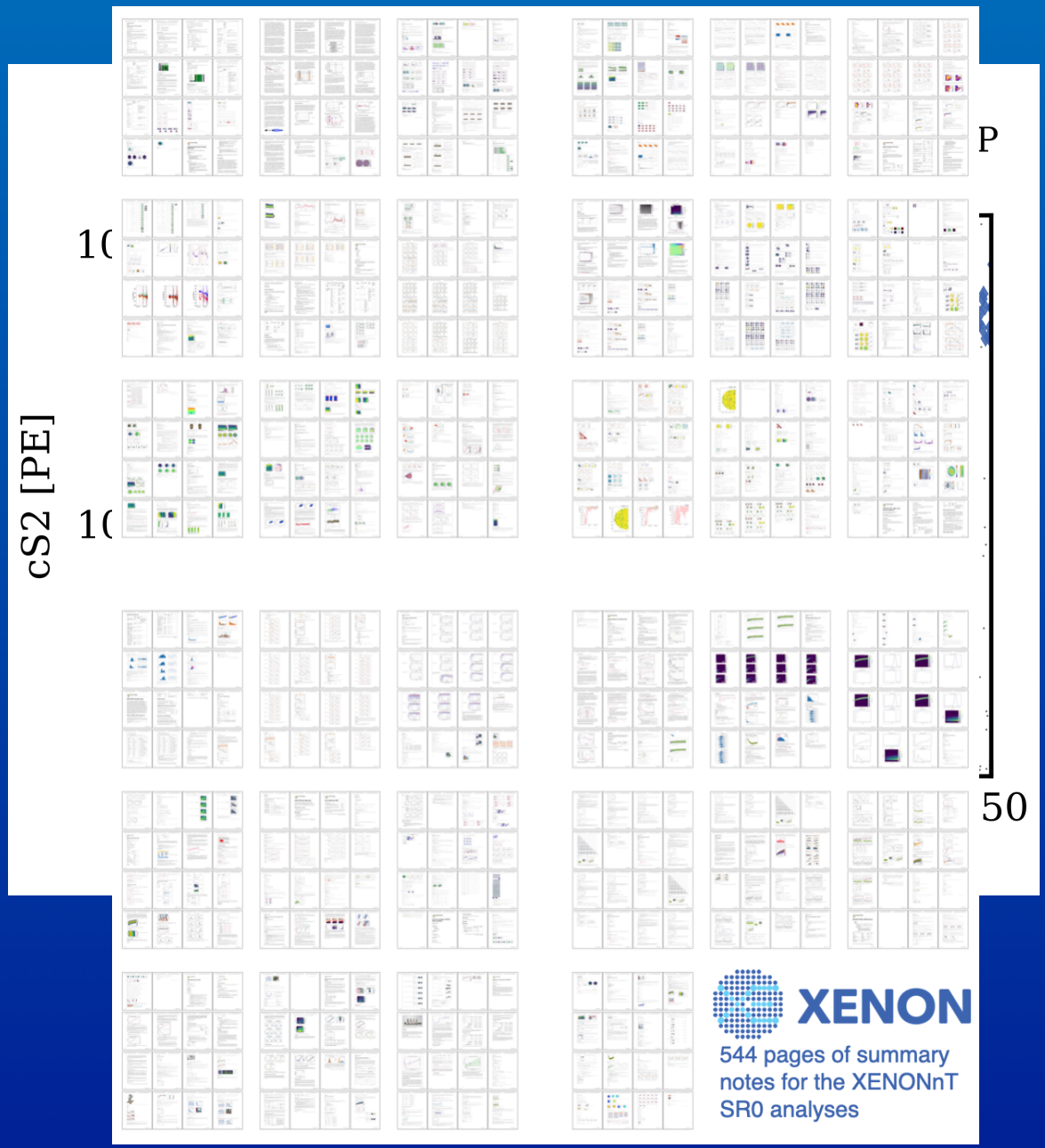
Experimenter bias is a danger with few events

- The most common experimenter bias mitigation method is “blinding”— not showing the signal-like region of parameter space until the analysis has been frozen

- LUX developed a “salting” procedure where synthetic signals were made by stitching together genuine S1 and S2 signals into full events in the data



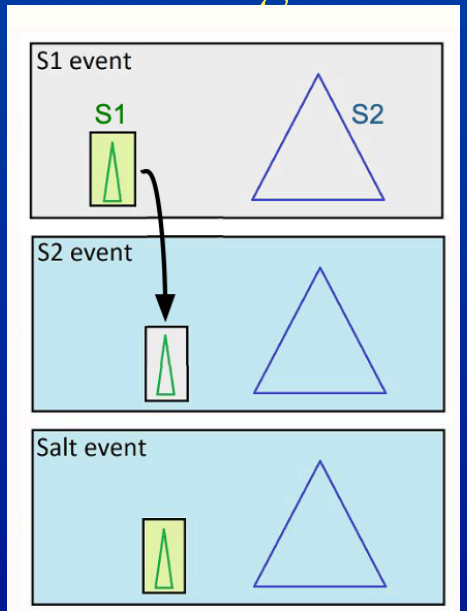
Tyler Anderson “Salting as a Bias Mitigation Technique in LZ”, presentation at LIDINE 2021

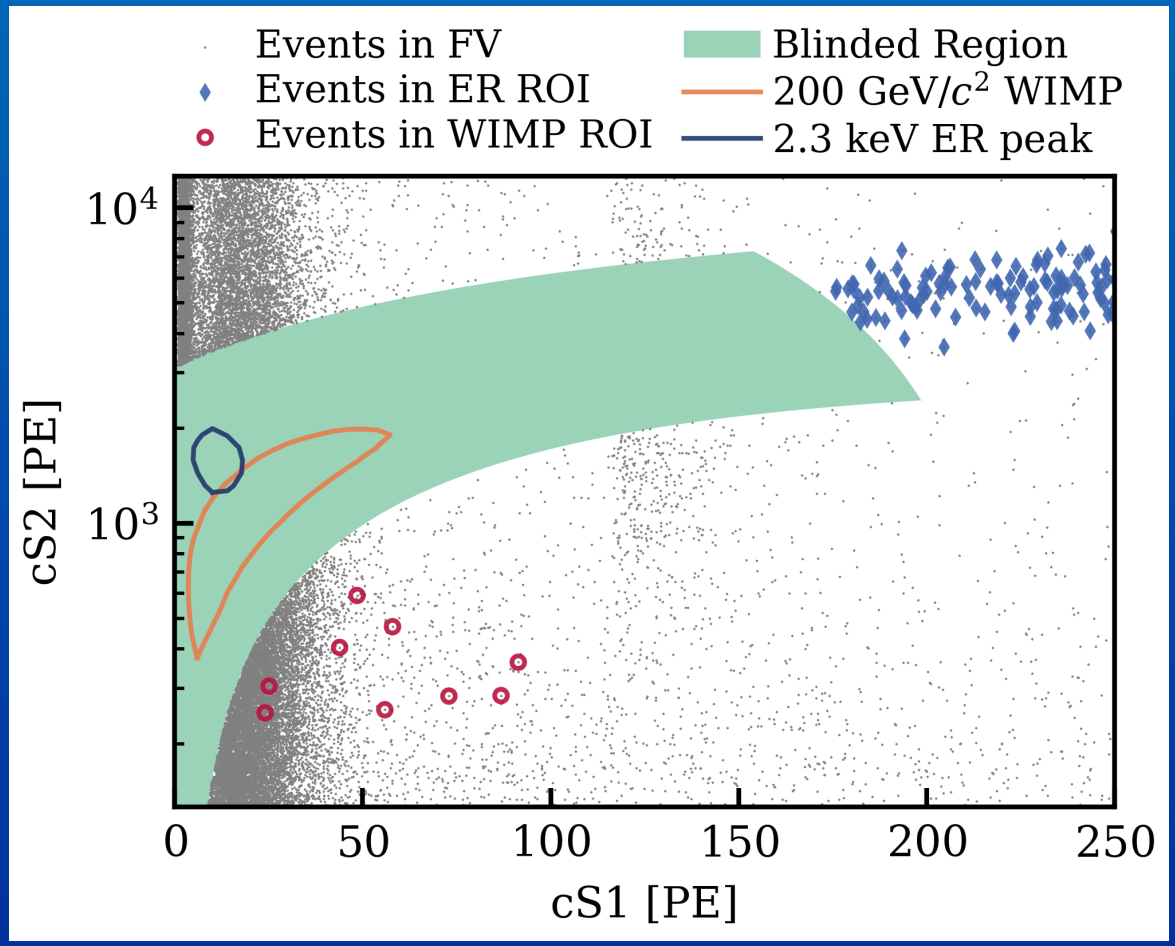


Experimenter bias is a danger with few events

- The most common experimenter bias mitigation method is "blinding"— not showing the signal-like region of parameter space until the analysis has been frozen

- LUX developed a "salting" procedure where synthetic signals were made by stitching together genuine S1 and S2 signals into full events in the data



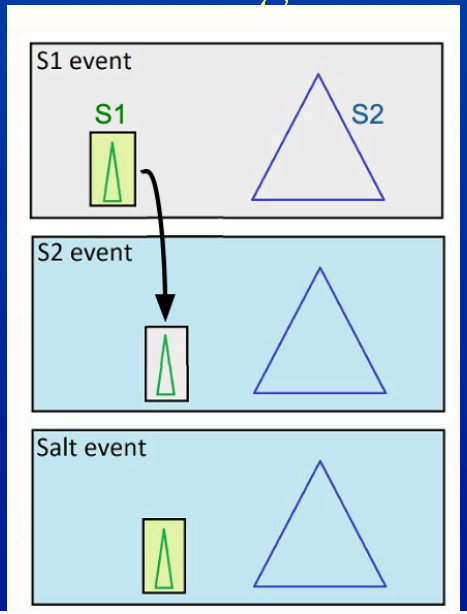


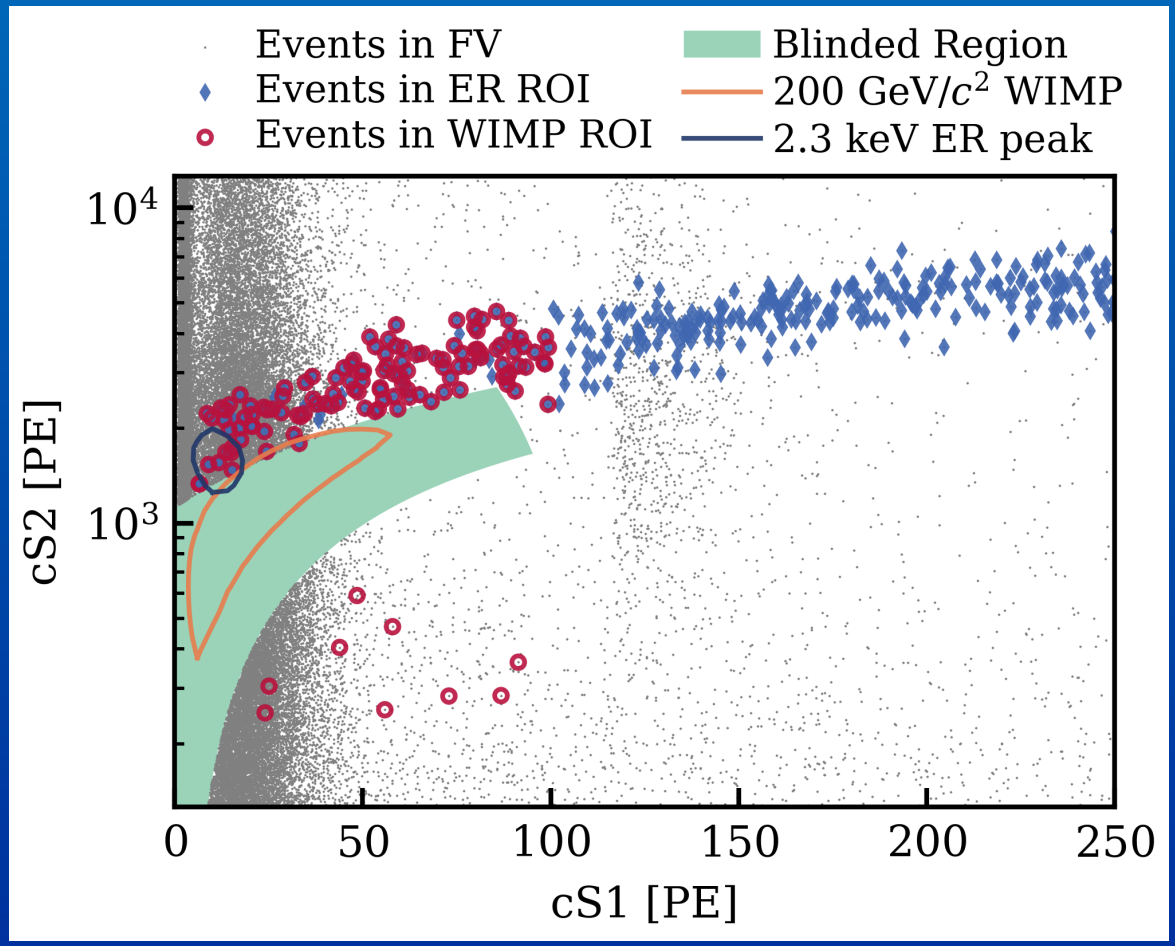
Tyler Anderson "Salting as a Bias Mitigation Technique in LZ", presentation at LIDINE 2021

Experimenter bias is a danger with few events

- The most common experimenter bias mitigation method is "blinding"— not showing the signal-like region of parameter space until the analysis has been frozen

- LUX developed a "salting" procedure where synthetic signals were made by stitching together genuine S1 and S2 signals into full events in the data



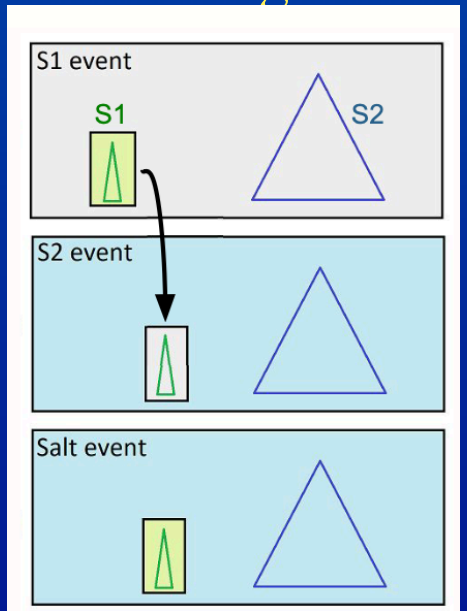


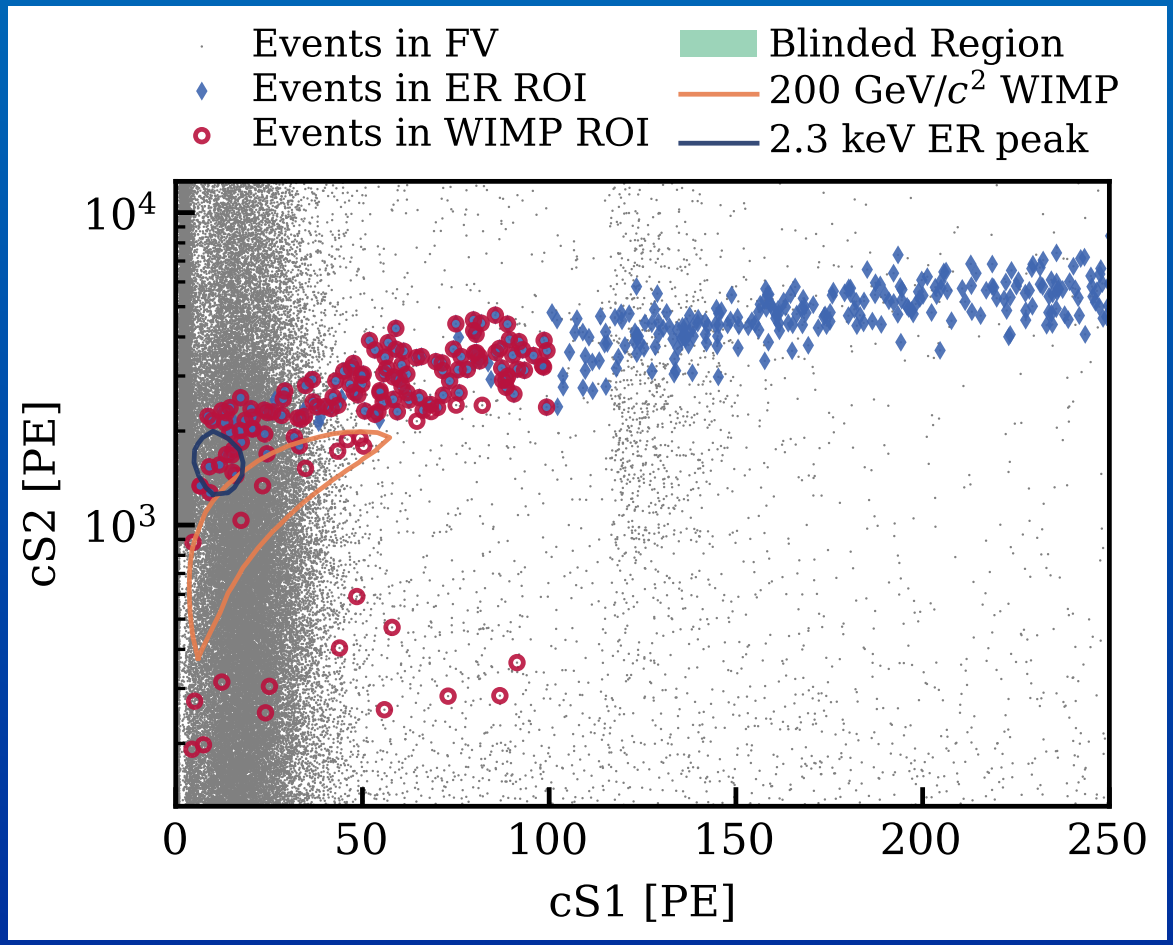
Tyler Anderson "Salting as a Bias Mitigation Technique in LZ", presentation at LIDINE 2021

Experimenter bias is a danger with few events

- The most common experimenter bias mitigation method is "blinding"— not showing the signal-like region of parameter space until the analysis has been frozen

- LUX developed a "salting" procedure where synthetic signals were made by stitching together genuine S1 and S2 signals into full events in the data





Tyler Anderson "Salting as a Bias Mitigation Technique in LZ", presentation at LIDINE 2021

For today

- Example analyses: we saw how experiments compose analytical and other models to make their full statistical model
- Profile Likelihood: we discussed minimising nuisance parameters only, if we wish to test some hypothesis
- Asymptotic distributions: How useful they are, and the common failures we encounter
- Look-Elsewhere effect: one of these effects

Hands-on session: profile likelihood, comparison with the asymptotic:

SOUP 2024 Exercise set 2: Likelihood optimisation, confidence intervals, asymptoticity checks

Welcome to the SOUP exercise set!

```
In [1]:
# Import modules with the tools we need:
import scipy.stats as sps
from scipy.optimize import minimize
import matplotlib.pyplot as plt
import numpy as np
```

0: Optimisation

scipy.optimize.minimize (or iminuit if you prefer) will be our bread-and-butter
The key trick is to remember how to set up the function to be used:

```
In [2]:
def function_scipy(x, a=0.5, b=1):
    #if 0<a, this function has a global minimum at 0,0, but if a is close to 0, this minimum will have a broad nea
    #note that only parameters inside the vector x are minimized for-- the rest are fixed parameters.
    return x[0]**2 + b*(x[1]-a*x[0]**2)**4
```

```
In [3]:
#minimize using scipy:
scipy_result = minimize(function_scipy, [3, 9.2], args=(0.5,1)) #notice that you have to provide a guess for "x"
#and that you can pass further args like a,b to the minimizer
print(scipy_result)
```

```
     fun: 1.2859447158387182e-08
hess_inv: array([[3.80897491e-01, 1.25210924e-02],
       [1.25210924e-02, 3.67946843e+02]])
     jac: array([ 9.02614194e-07, -4.83026339e-06])
 message: 'Optimization terminated successfully.'
    nfev: 102
     nit: 24
    njev: 34
  status: 0
 success: True
       x: array([ 4.43855426e-07, -1.06488795e-02])
```

Iminuit

Iminuit is heavily used in particle physics, and has some nice extra functions

```
In [4]:
def function_minuit(x,y, a=0.5, b=1):
    return function_scipy([x,y],a, b)
```

```
In [5]:
try:
    import iminuit
    nobj = iminuit.Minuit(function_minuit, x=0,y=0.2, a=0.5, b=1)
    nobj.fixed["a"] = True #here, we can set explicitly parameters to be fixed (and free them later!)
    nobj.fixed["b"] = True #here, we can set explicitly parameters to be fixed (and free them later!)
    minuit_result = nobj.migrad() #call the minimizer routine (called Migrad)
    print(minuit_result)
except:
    minuit_result = None
    print("Iminuit is not installed, install it or use scipy.optimize")
```

Migrad			
FCN = 1.475e-06		Nfcn = 161	
EDM = 1.15e-06 (Goal: 0.0002)			
Valid Minimum		No Parameters at limit	
Below EDM threshold (goal x 10)		Below call limit	
Covariance	Hesse ok	Accurate	Pos. def.	Not forced

	Name	Value	Hesse Err	Minos Err-	Minos Err+	Limit-	Limit+	Fixed
0	x	-0	1					
1	y	-0	13					
2	a	0.500	0.005					yes
3	b	1.00	0.01					yes

	x	y	a	b

Knut Dundas Morå
fysikk@dundasmora.no, he/him

School of Underground Physics at Bertinoro

Statistics and Inference

for rare event searches

What is a statistical model?
Does it describe your data?
What kinds of conclusions can we draw?

24

Yesterday

- Example analyses: we saw how experiments compose analytical and other models to make their full statistical model
- Profile Likelihood: we discussed minimising nuisance parameters only, if we wish to test some hypothesis
- Asymptotic distributions: How useful they are, and the common failures we encounter
- Look-Elsewhere effect: one of these effects

Hands-on session: profile likelihood, comparison with the asymptotic:

SOUP 2024 Exercise set 2: Likelihood optimisation, confidence intervals, asymptoticity checks

Welcome to the SOUP exercise set!

```
In [1]:
# Import modules with the tools we need:
import scipy.stats as sps
from scipy.optimize import minimize
import matplotlib.pyplot as plt
import numpy as np
```

0: Optimisation

scipy.optimize.minimize (or iminuit if you prefer) will be our bread-and-butter
The key trick is to remember how to set up the function to be used:

```
In [2]:
def function_scipy(x, a=0.5, b=1):
    #if 0<a, this function has a global minimum at 0,0, but if a is close to 0, this minimum will have a broad nea
    #note that only parameters inside the vector x are minimized for-- the rest are fixed parameters.
    return x[0]**2 + b*(x[1]-a*x[0]**2)**4
```

```
In [3]:
#minimize using scipy:
scipy_result = minimize(function_scipy, [3, 9.2], args=(0.5,1)) #notice that you have to provide a guess for "x"
#and that you can pass further args like a,b to the minimizer
print(scipy_result)
```

```
     fun: 1.2859447158387182e-08
hess_inv: array([[3.80897491e-01, 1.25210924e-02],
       [1.25210924e-02, 3.67946843e+02]])
     jac: array([ 9.02614194e-07, -4.83026339e-06])
 message: 'Optimization terminated successfully.'
    nfev: 102
     nit: 24
    njev: 34
  status: 0
 success: True
       x: array([ 4.43855426e-07, -1.06488795e-02])
```

Iminuit

Iminuit is heavily used in particle physics, and has some nice extra functions

```
In [4]:
def function_minuit(x,y, a=0.5, b=1):
    return function_scipy([x,y],a, b)
```

```
In [5]:
try:
    import iminuit
    nobj = iminuit.Minuit(function_minuit, x=0,y=0.2, a=0.5, b=1)
    nobj.fixed["a"] = True #here, we can set explicitly parameters to be fixed (and free them later!)
    nobj.fixed["b"] = True #here, we can set explicitly parameters to be fixed (and free them later!)
    minuit_result = nobj.migrad() #call the minimizer routine (called Migrad)
    print(minuit_result)
except:
    minuit_result = None
    print("Iminuit is not installed, install it or use scipy.optimize")
```

Migrad			
FCN = 1.475e-06		Nfcn = 161	
EDM = 1.15e-06 (Goal: 0.0002)			
Valid Minimum		No Parameters at limit	
Below EDM threshold (goal x 10)		Below call limit	
Covariance	Hesse ok	Accurate	Pos. def.	Not forced

	Name	Value	Hesse Err	Minos Err-	Minos Err+	Limit-	Limit+	Fixed
0	x	-0	1					
1	y	-0	13					
2	a	0.500	0.005					yes
3	b	1.00	0.01					yes

	x	y	a	b

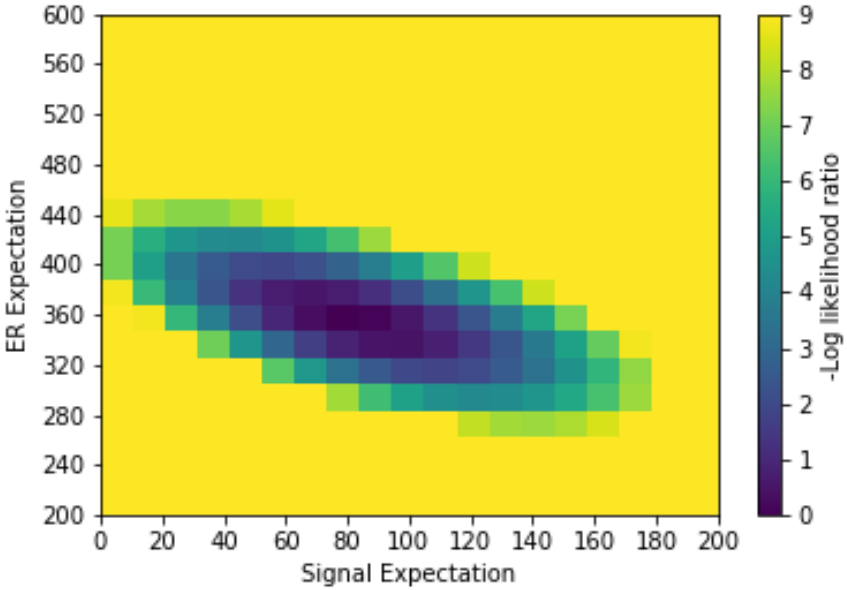
For today

- Frequentist confidence intervals
- The profile construction
- A couple of tools

Hands-on session: confidence intervals, profile construction

Profiling Likelihoods & nuisance parameters

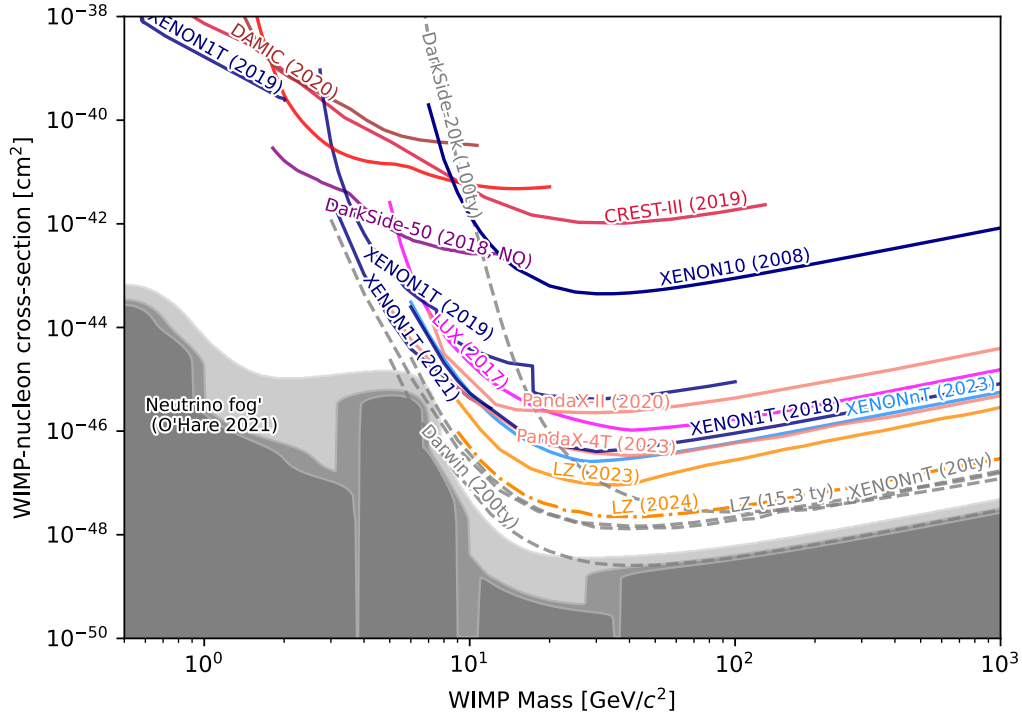
- We seldom have completely specified hypotheses
- Our background and signal models have uncertainties, parameterised by nuisance parameters (θ)— you'll see some examples in the next slides.
- The global best fit we denote with $\hat{s}, \hat{\theta}$
- However, we also want to test other s — for example s=0 for discovery significance or a range of s for confidence intervals.
- In these cases, we set the other nuisance parameters to their conditional best-fit $\hat{\hat{\theta}}$.

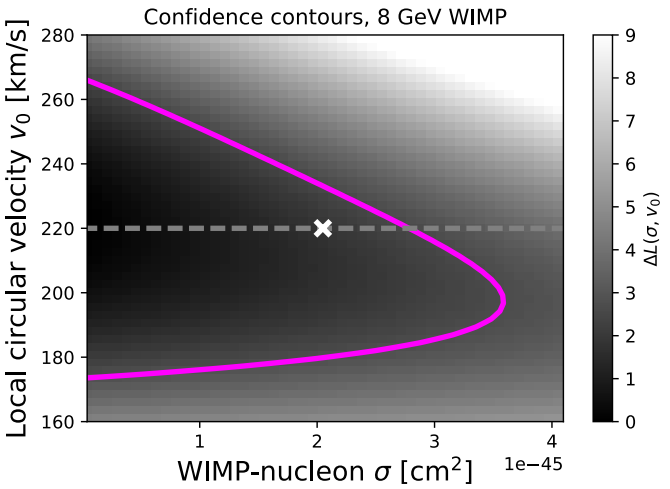


$$\delta \log \mathscr{L}(s, \hat{\hat{\boldsymbol{\theta}}}) / \delta \theta_j = 0;$$

Confidence Intervals

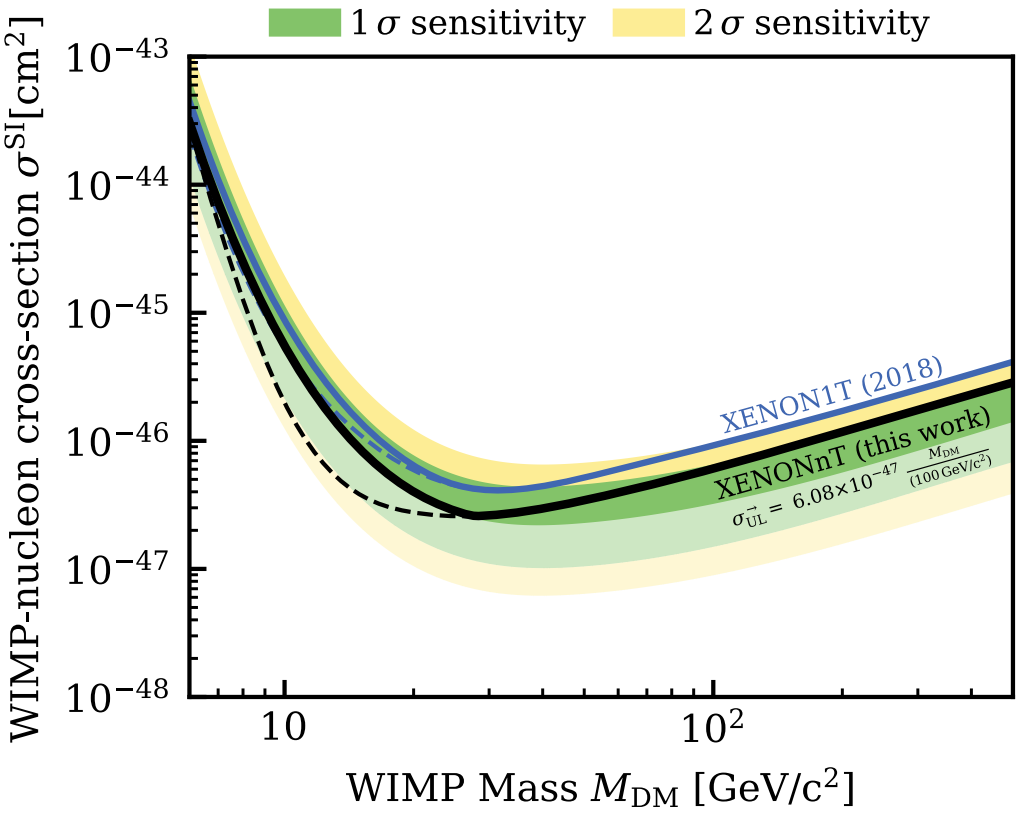
- More important than any point estimate is being able to quantify our uncertainty about our measurement
- Within the frequentist paradigm, we do this with confidence intervals, that are required to have certain properties exactly analogous to hypothesis tests— a certain probability of false rejection.
- Don't need to be one-dimensional — but it can become tricky to work with high dimensions





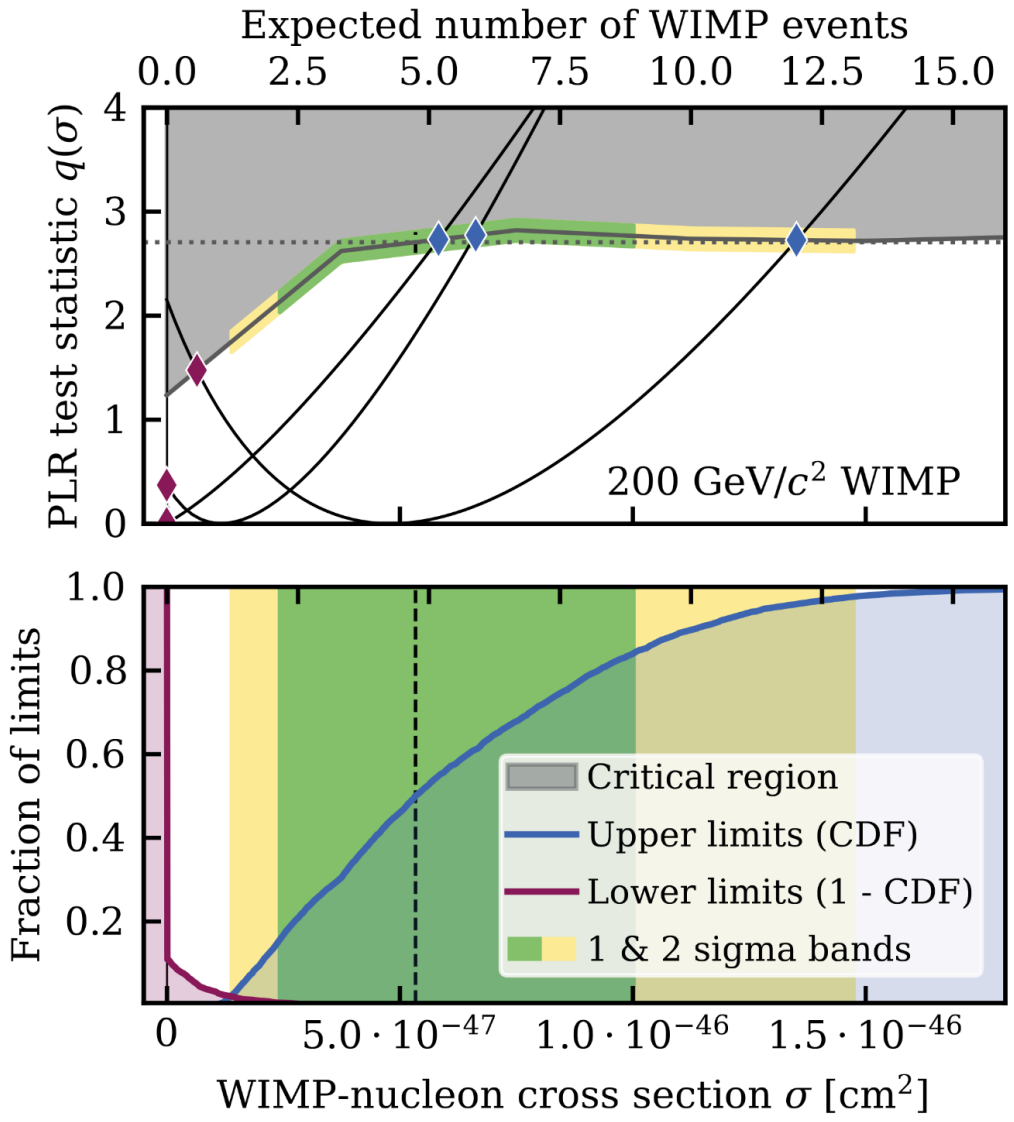
Confidence Intervals

- You might be asked "what is the uncertainty of your upper limit"
- The answer is that the upper limit *is the uncertainty*
- So far, almost all direct detection searches have measured dark matter as $0^{+\text{Upper Limit}}_{-0}$ or, if they were "unlucky" $\epsilon^{+\text{Upper Limit}}_{-\epsilon}$
- The "Brazil Band" shown with limits show the expected result—so it can serve as a proxy of significance at the most



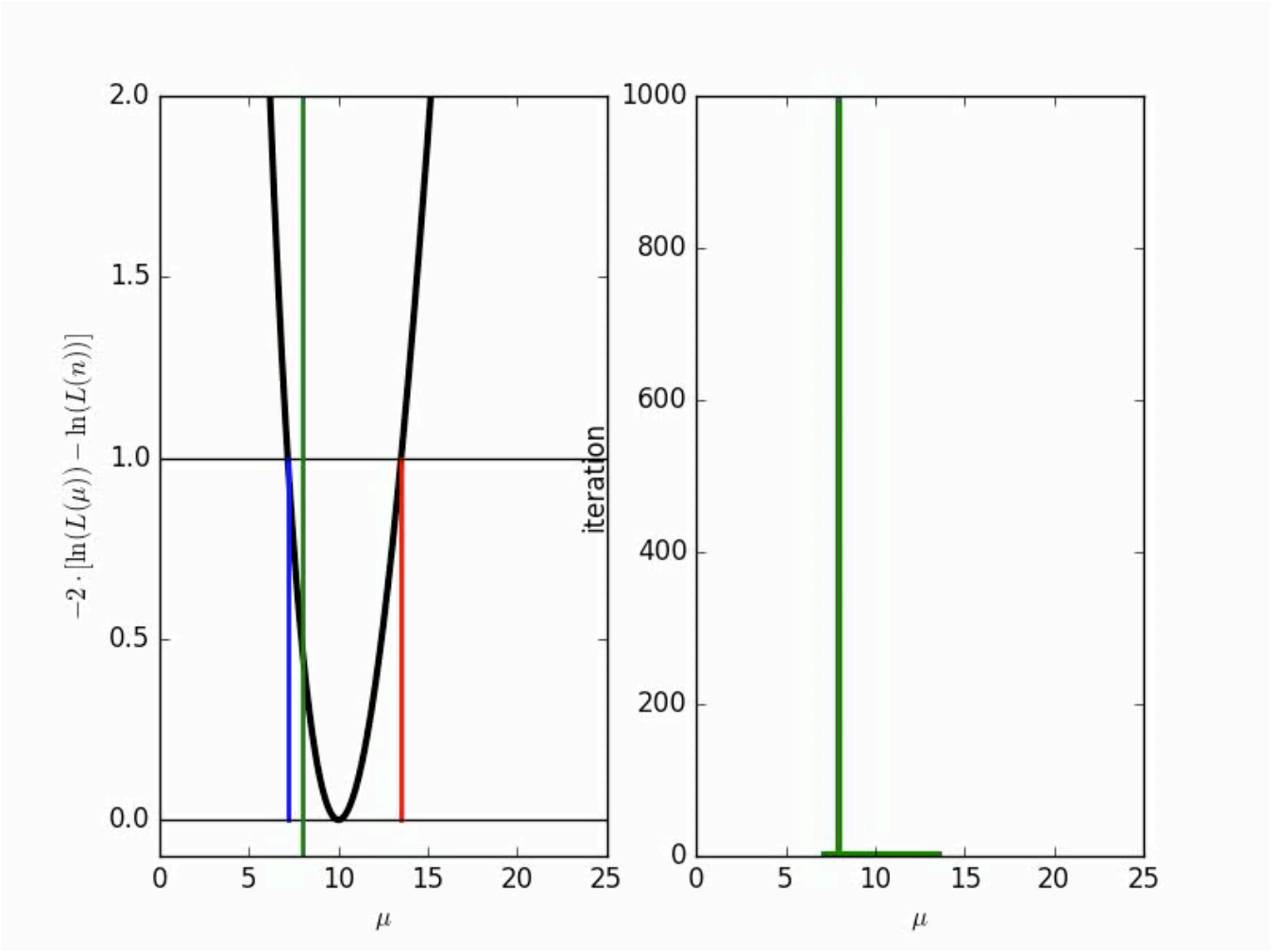
Confidence Intervals

- You might be asked "what is the uncertainty of your upper limit"
- The answer is that the upper limit *is the uncertainty*
- So far, almost all direct detection searches have measured dark matter as $0^{+\text{Upper Limit}}_{-0}$ or, if they were "unlucky" $\epsilon^{+\text{Upper Limit}}_{-\epsilon}$
- The "Brazil Band" shown with limits show the expected result—so it can serve as a proxy of significance at the most



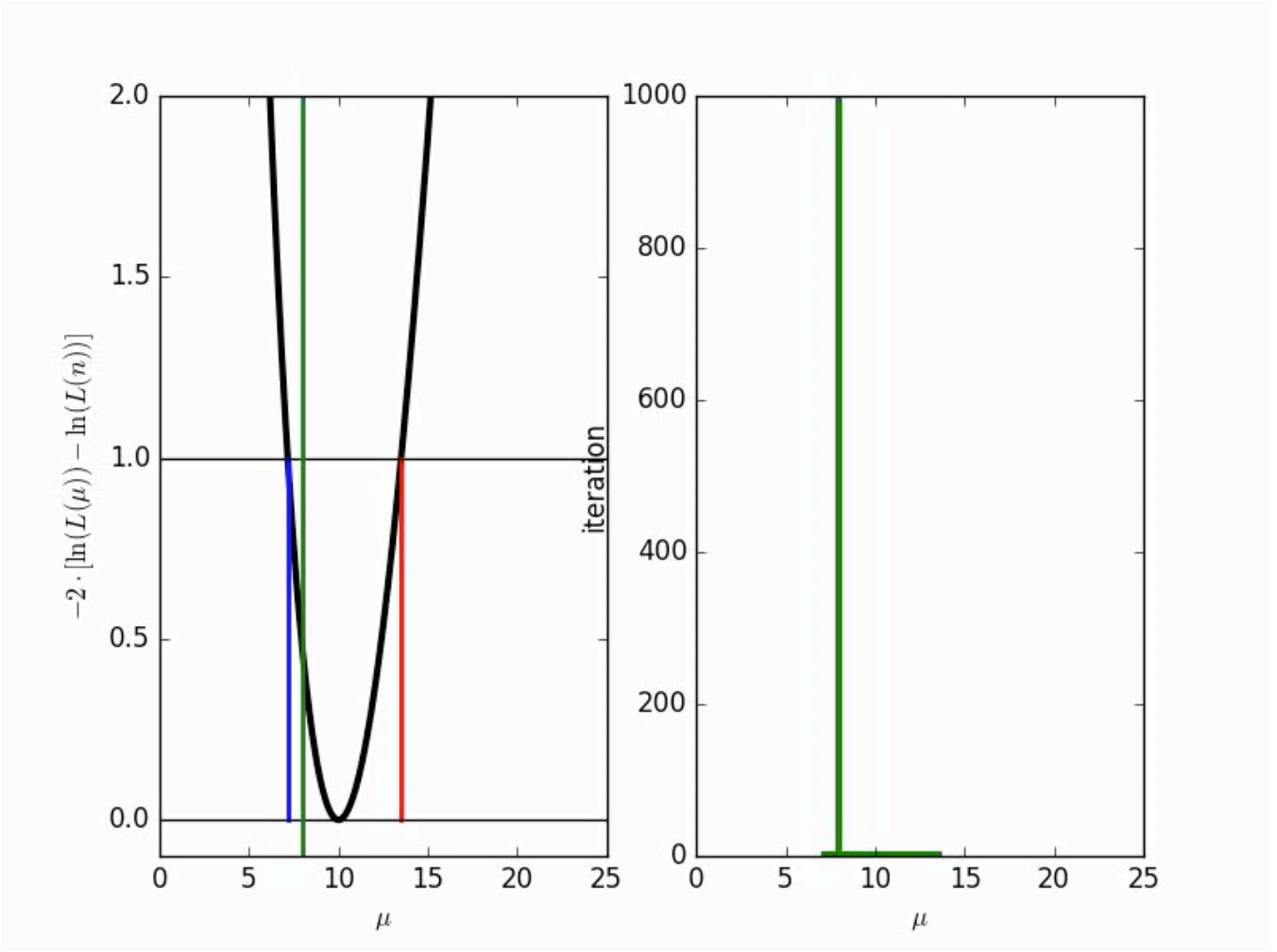
What is "Coverage"?

- The principle feature of frequentist confidence intervals is *coverage*—in the long run, the fraction confidence intervals reported by experiments that contain the true value should approach the stated confidence level (CL).



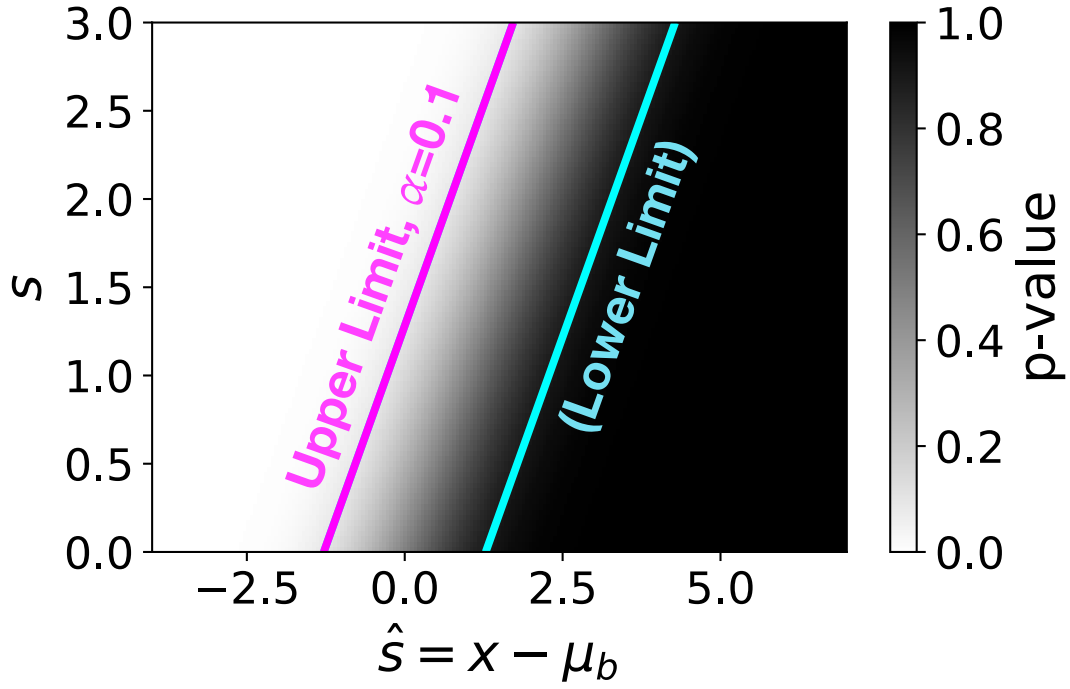
What is "Coverage"?

- The principle feature of frequentist confidence intervals is *coverage*—in the long run, the fraction confidence intervals reported by experiments that contain the true value should approach the stated confidence level (CL).

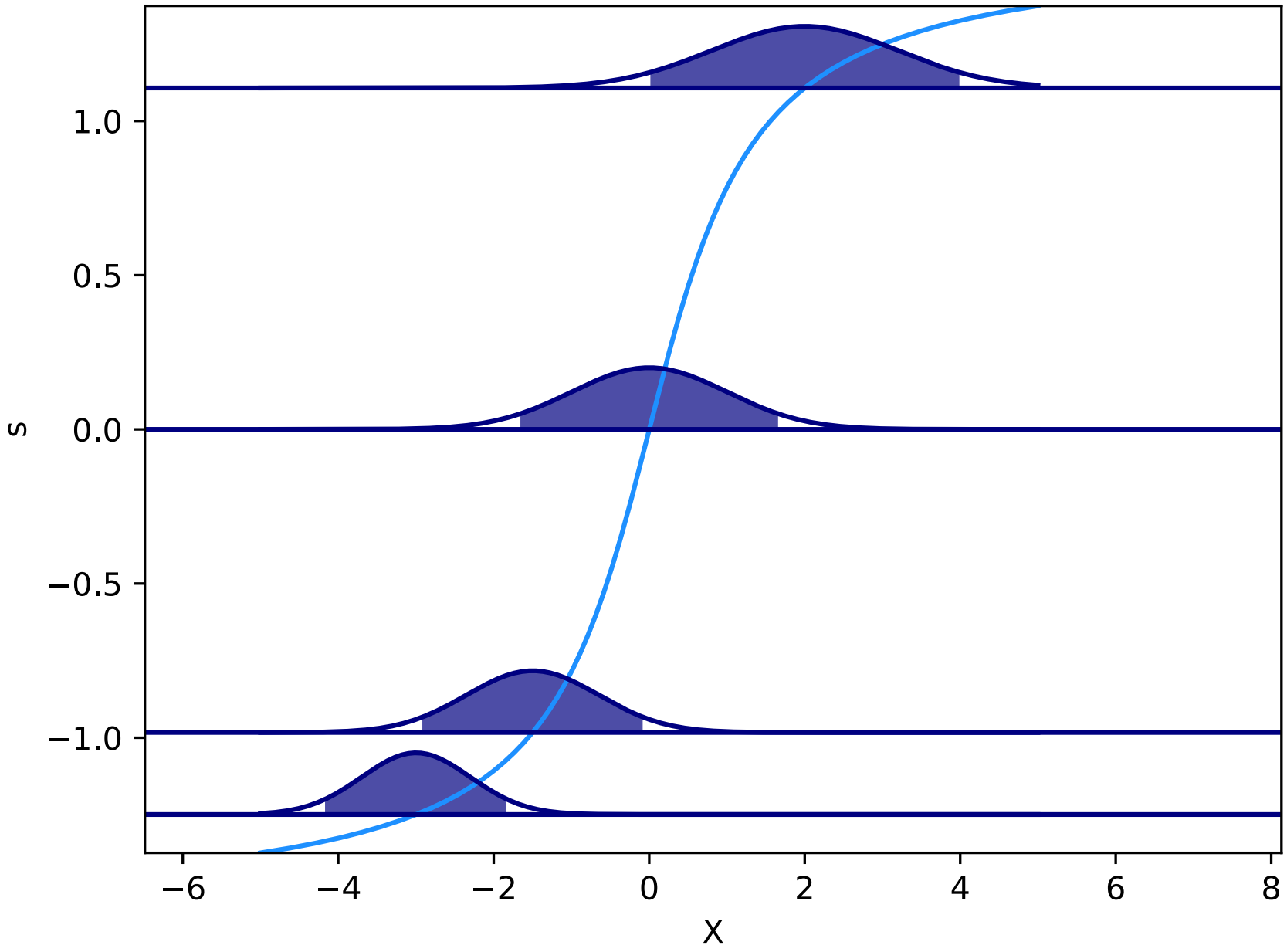


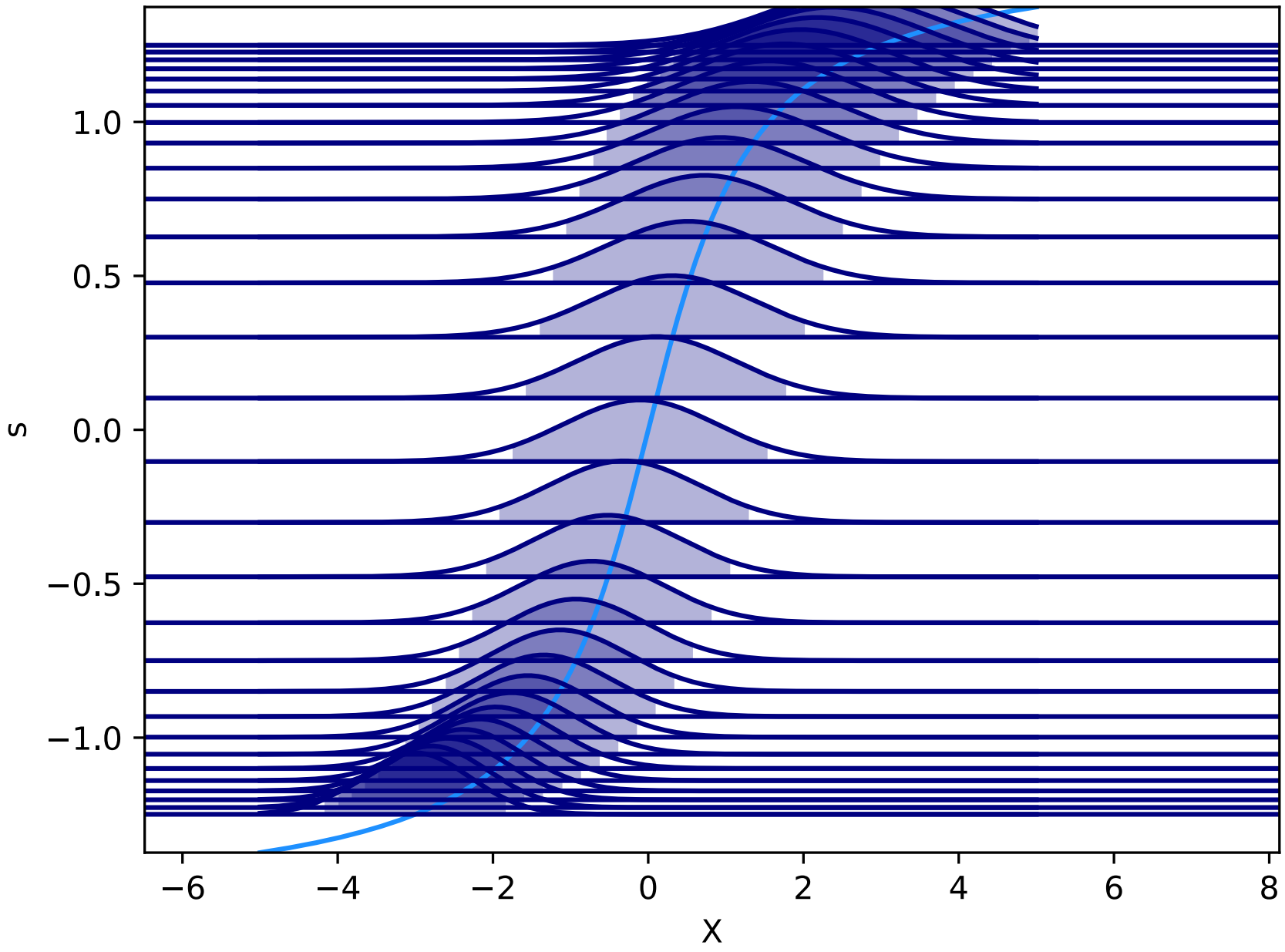
The Neyman construction

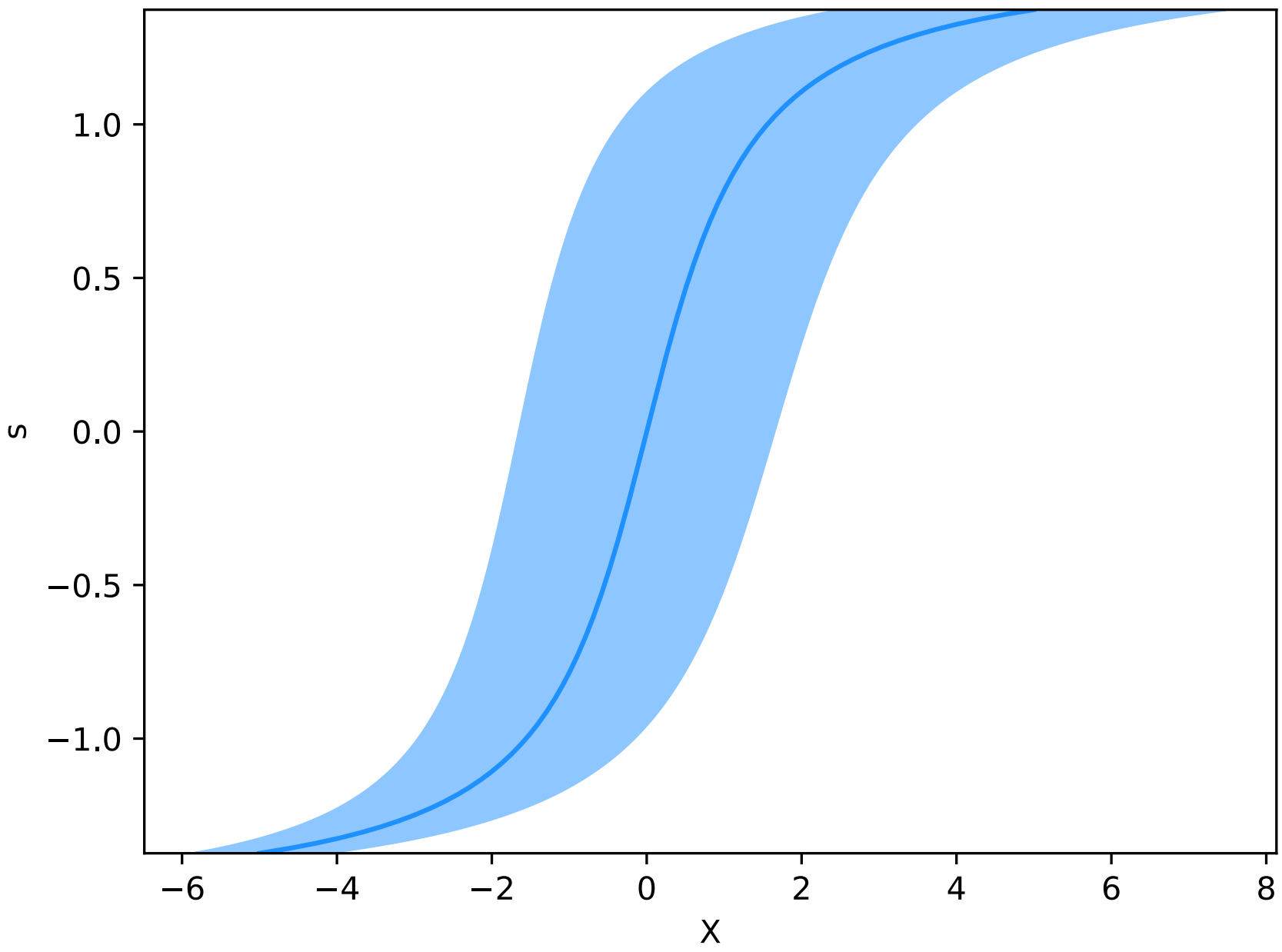
- For *each signal value s:*
  - Find the distribution of the test statistic given s, $f(X \mid s)$
  - Find limits, $x_{\mathrm{dn}}(s), x_{\mathrm{up}}(s)$ between which X occurs CL of the time
  - Invert them, then the confidence interval for S is $x_{\mathrm{up}}^{-1}(X), x_{\mathrm{dn}}^{-1}(X)$

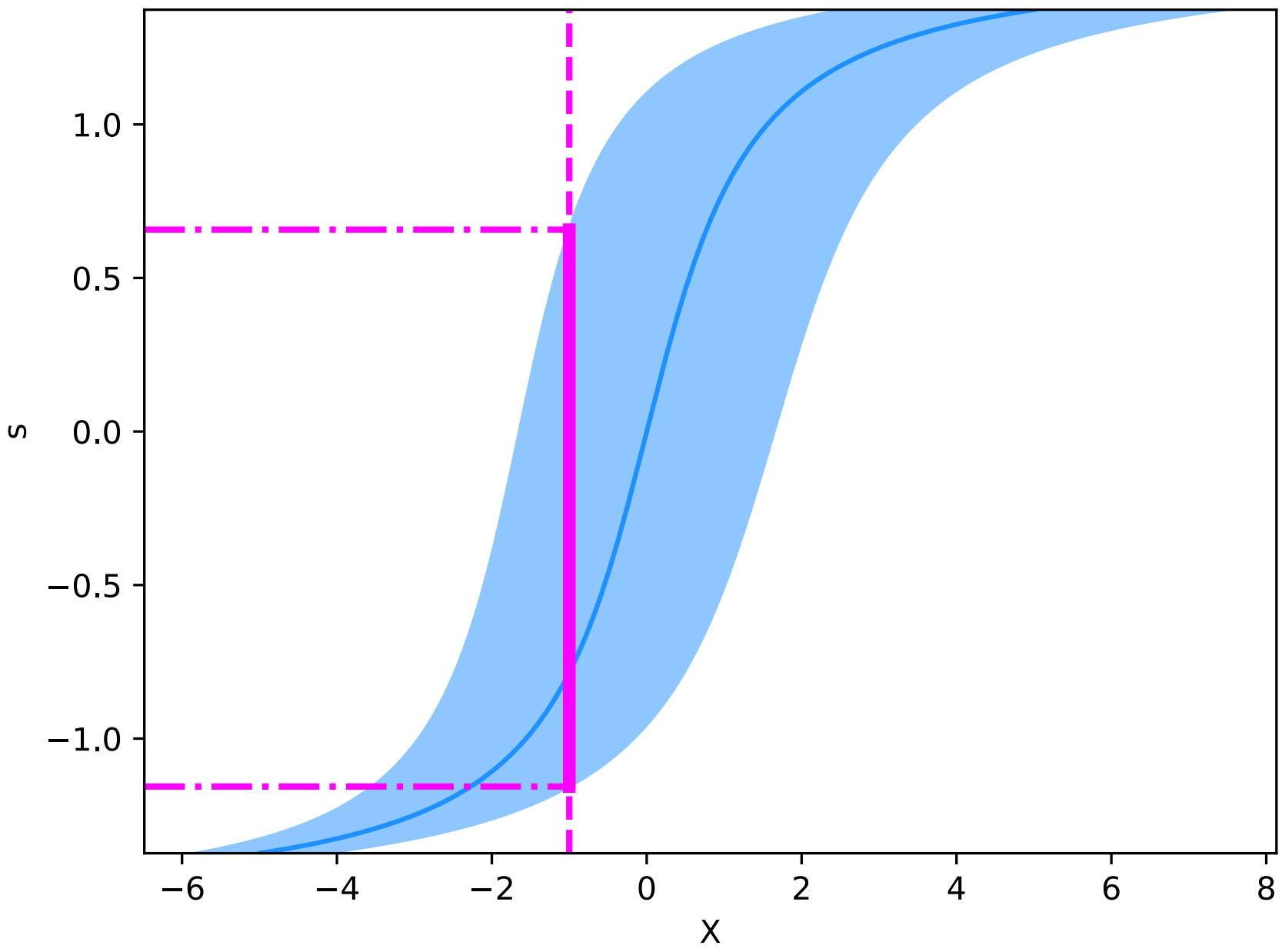


Neyman construction for a signal s
x ~ Gaussian with μ = s+μb σ=1,


1.0
0.5
0.0
−0.5
−1.0
s
−6
−4
−2
0
2
4
6
8
X


1.0
0.5
0.0
−0.5
−1.0
s
−6
−4
−2
0
2
4
6
8
X


1.0
0.5
0.0
−0.5
−1.0
s
−6
−4
−2
0
2
4
6
8
X

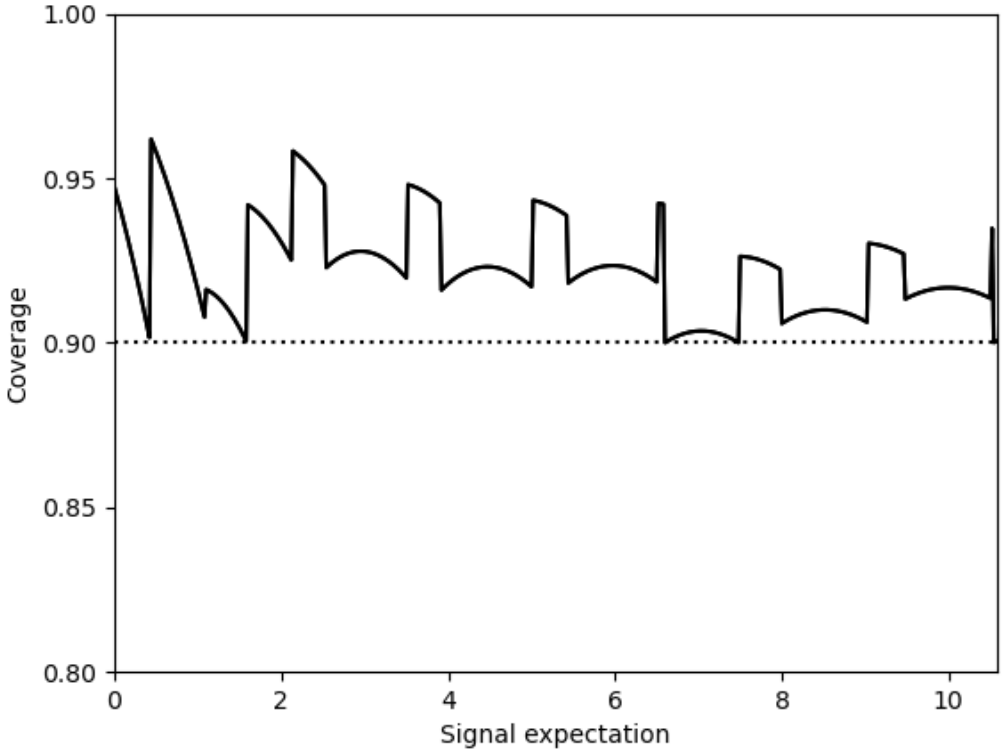

1.0
0.5
0.0
−0.5
−1.0
s
−6
−4
−2
0
2
4
6
8
X

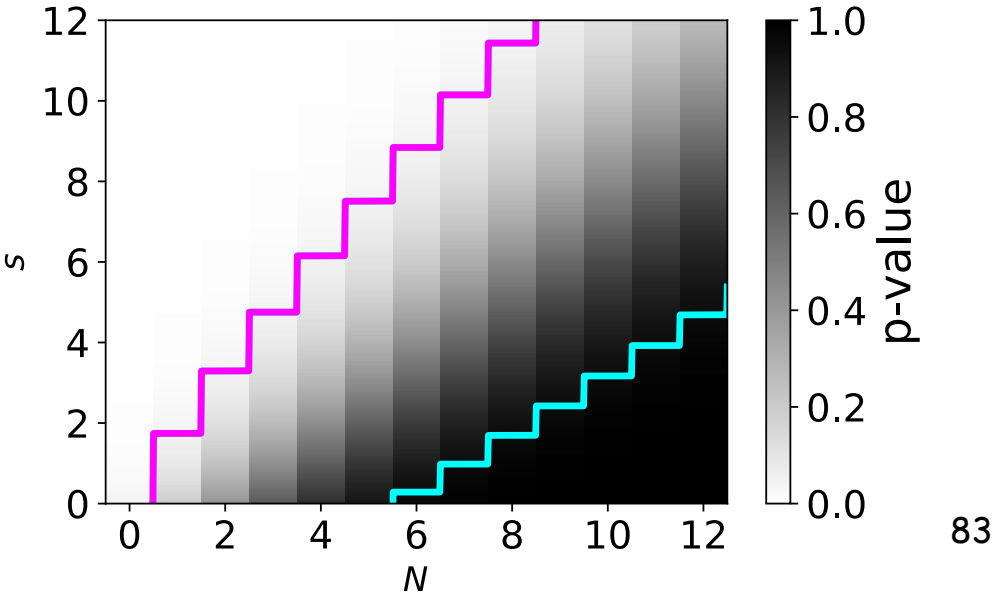
Coverage

- *N*ote that since coverage is concerned with the integrated probability rather than the density, confidence intervals *can* be transformed using any monotonic transformation and remain correct, unlike estimators, for example.

Coverage with discreteness

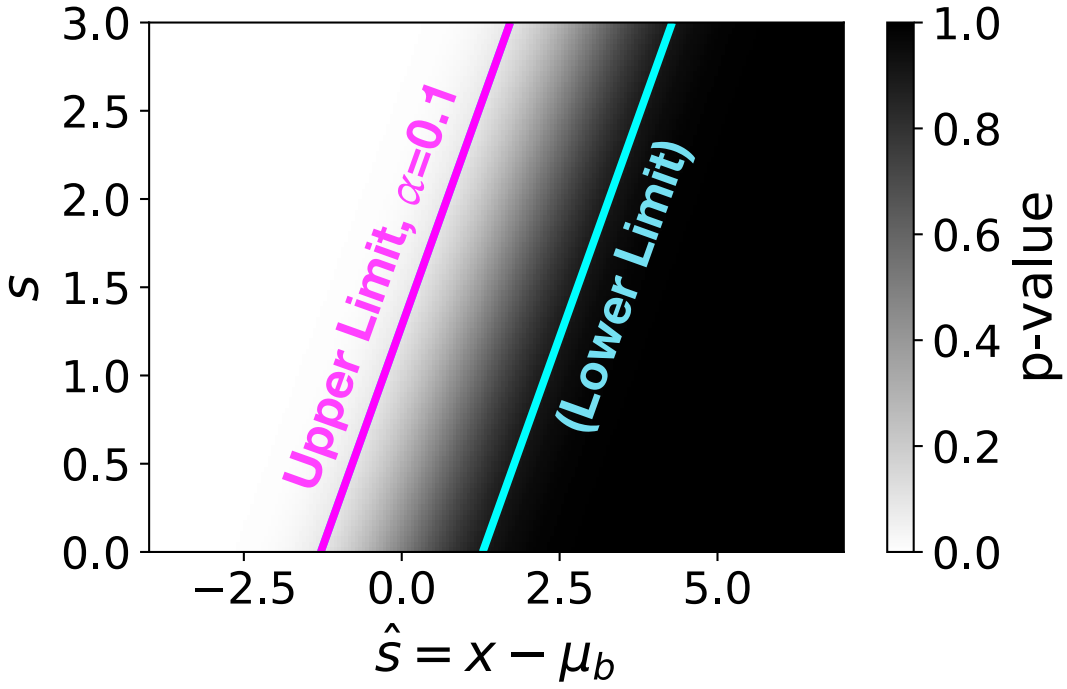
- Ideally, we have exact coverage, but in particular discrete distributions may not allow this
  - In which case we try to always err on the side of *overcoverage*
- It can often be irksome to get the "step" exactly right when doing confidence intervals for discrete problems— do a small toy test if you're unsure





The Neyman construction

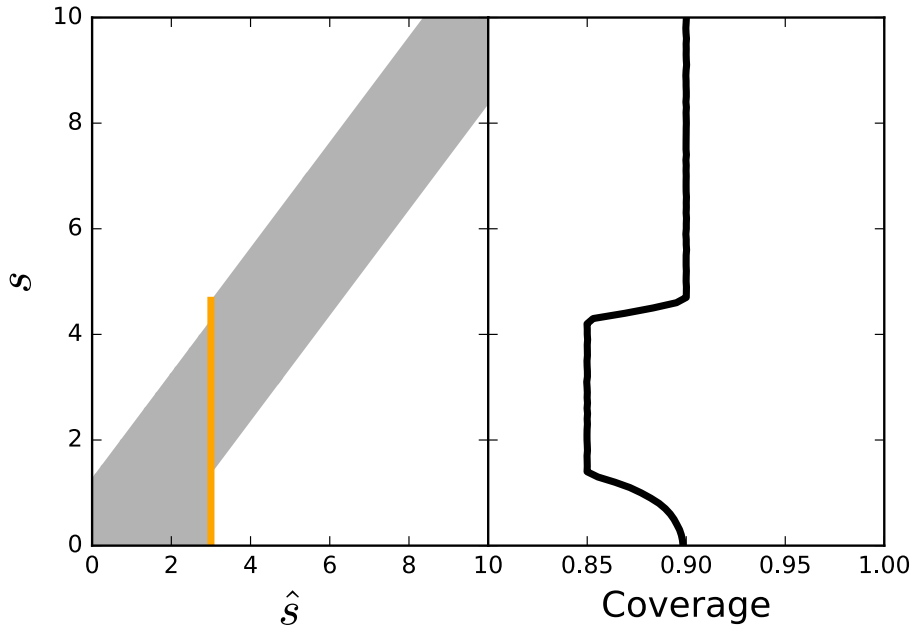
- Note— what we discussed so far does not completely determine the Neyman construction!
- I chose to have an equal probability for X to be below and above the bounds
- But if you want to set only upper limits, for example, you'll choose to shift the bounds to get the right coverage

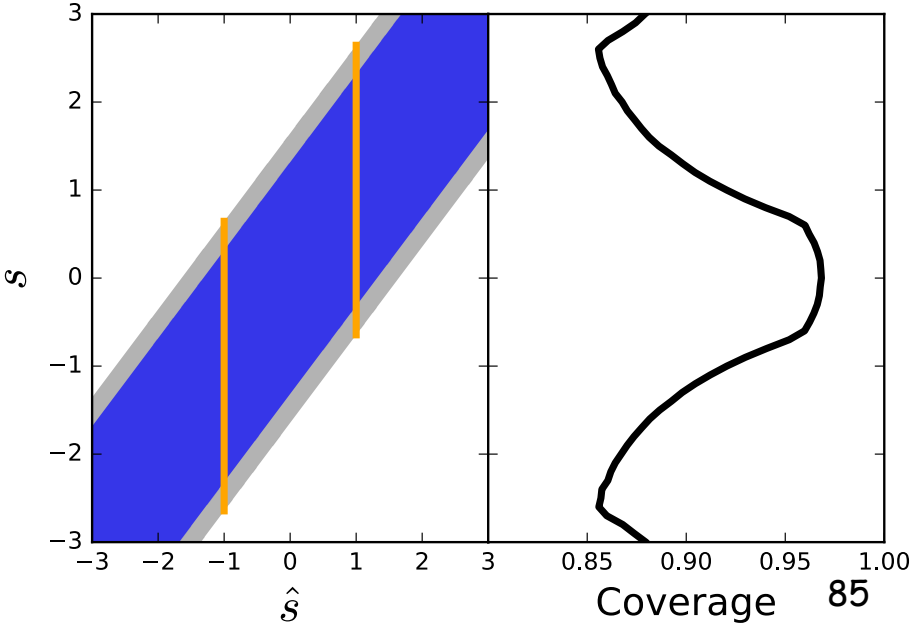


Neyman construction for a signal s
x ~ Gaussian with μ = s+μb σ=1,

Flip-flopping

- If an experiment uses a cut on their *result* to decide between a one- or two-sided interval, the decision line causes the complete method to undercover
- (In a similar vein, if an experiment re-runs its observation if it believes it is very unlikely according to H0 will also undercover)





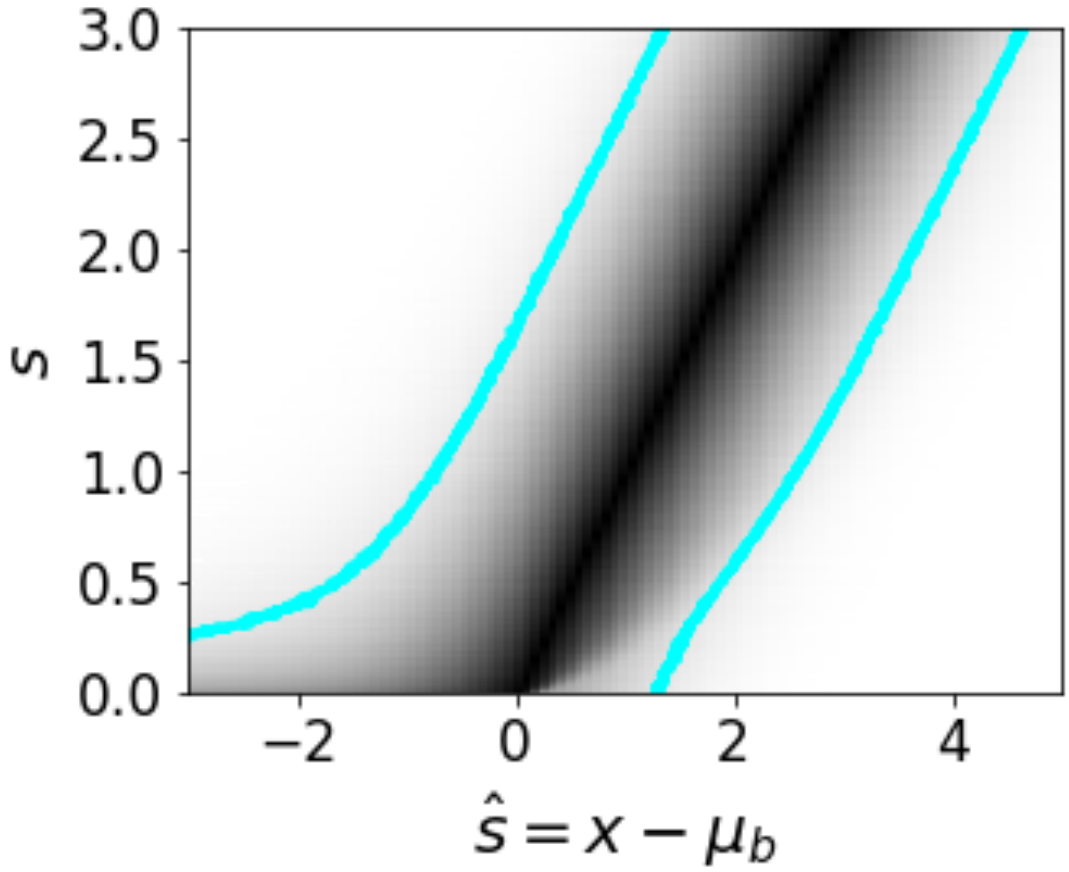
Unified Confidence Intervals

SoUP 24

AKA Feldman-Cousins intervals

$$R(\theta) = 2 \cdot \log\left[\mathcal{L}(\hat{s})/\mathcal{L}(s)\right]$$

- Since a single Neyman construction guarantees coverage, one that yields a smooth data-determined interpolation between upper limits and two-sided intervals may be just the ticket!
- FC uses the likelihood ratio as an ordering parameter, R, to decide what observations to include first in the Neyman construction
- Note: As defined, FC allows no nuisance parameters

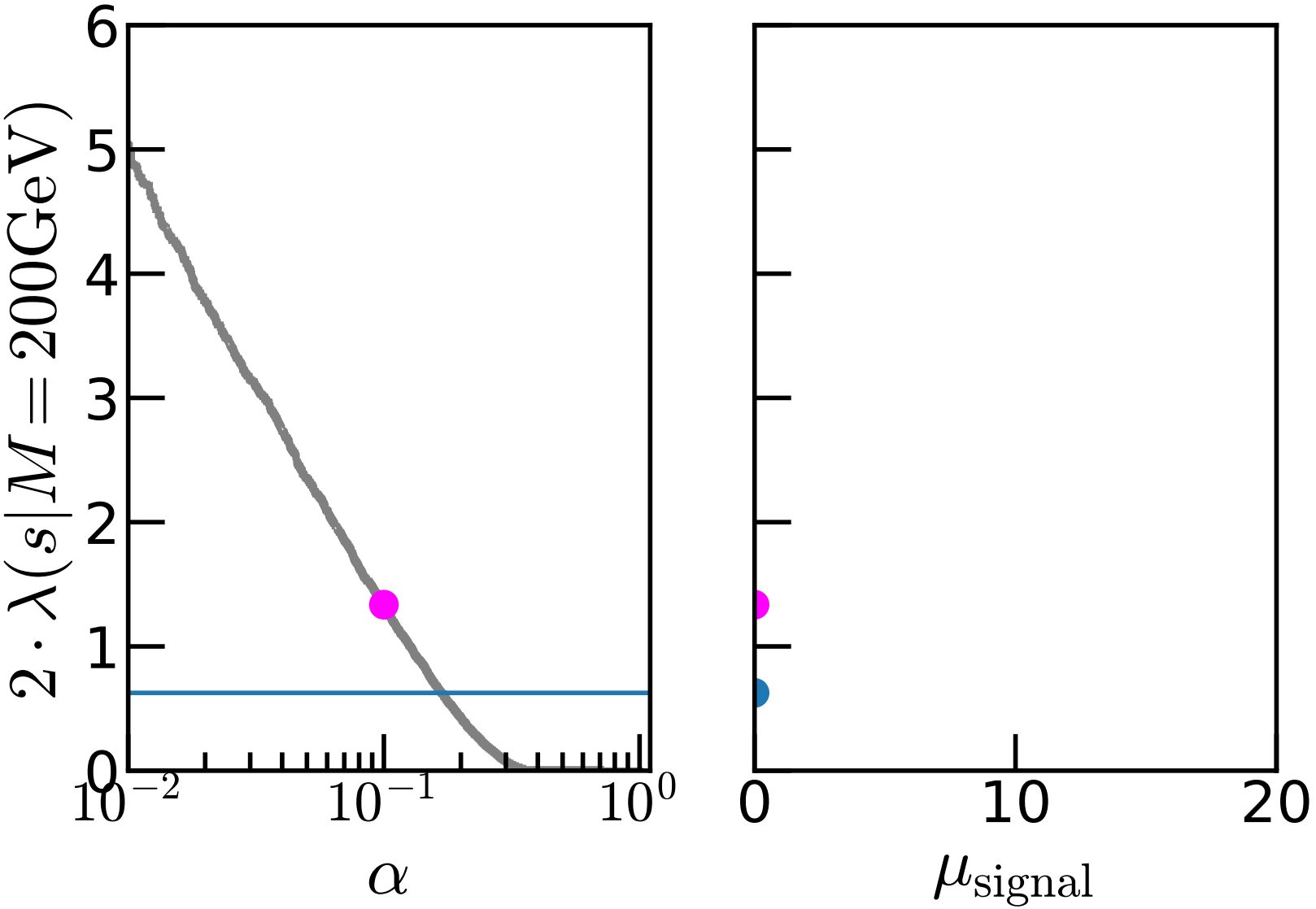


The Profile Construction

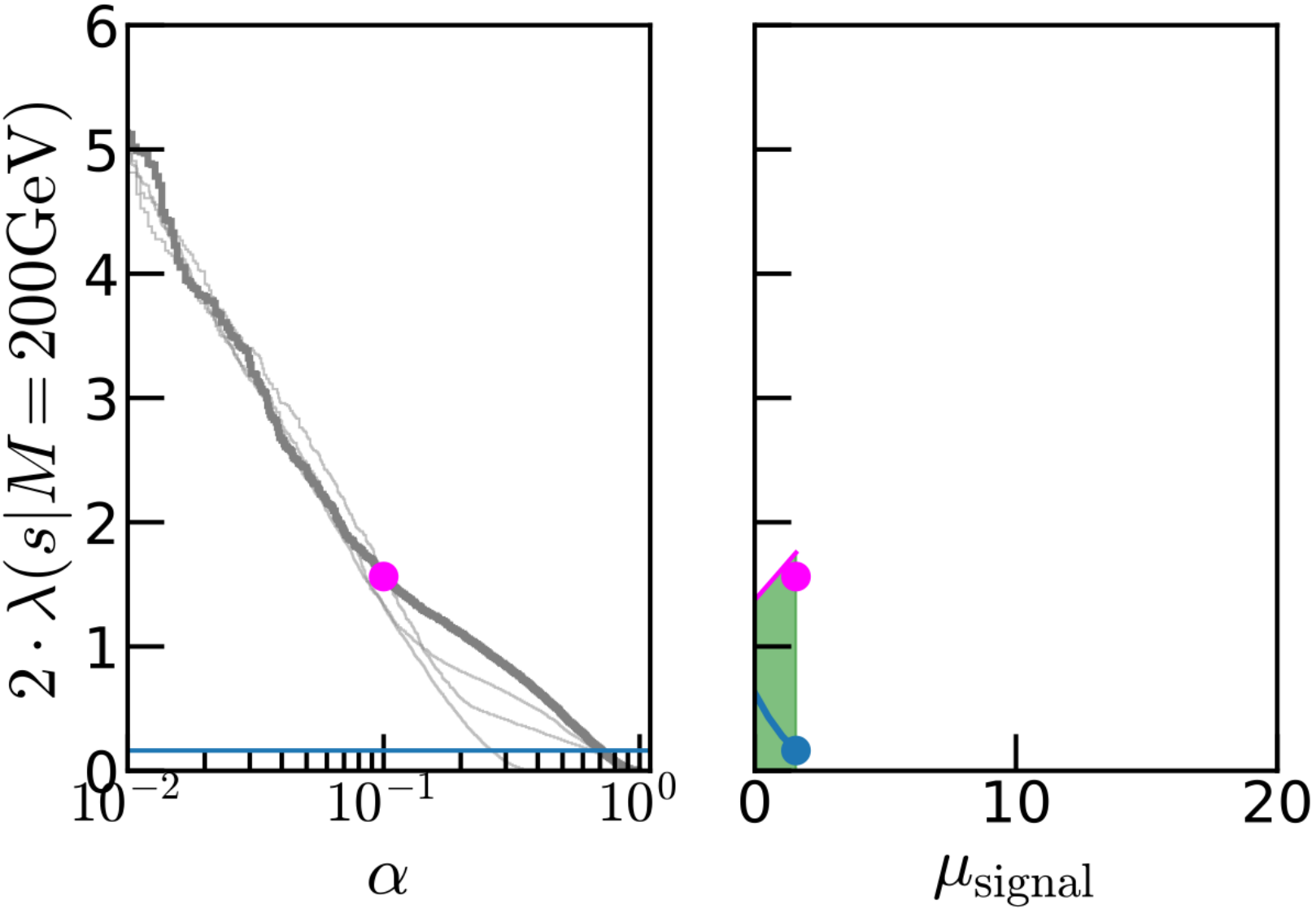
Unified intervals with nuisance parameters

- The unified construction may be extended to higher dimensions, but only at a cost— you have to know/define the distribution for your entire model space
- The Profile Construction uses the profiled likelihood ratio instead— this means *coverage is no longer guaranteed*
- Critically, this is not just a failure of asymptoticity— you'll get this problem even with toyMC methods.
- In particular, the method may fail if nuisance parameters are close to parameter boundaries or if they are strongly correlated with the signal

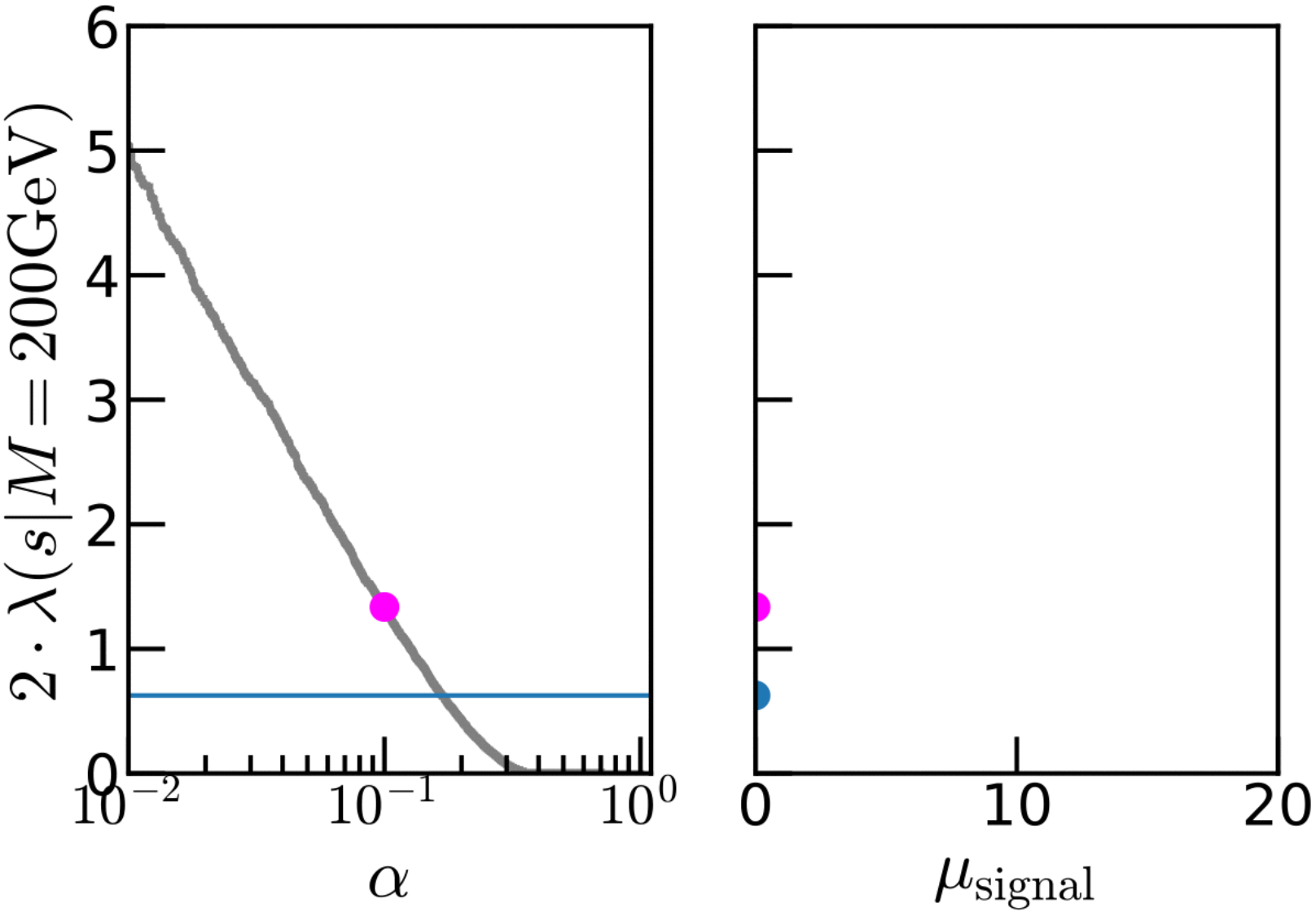
Power Construction



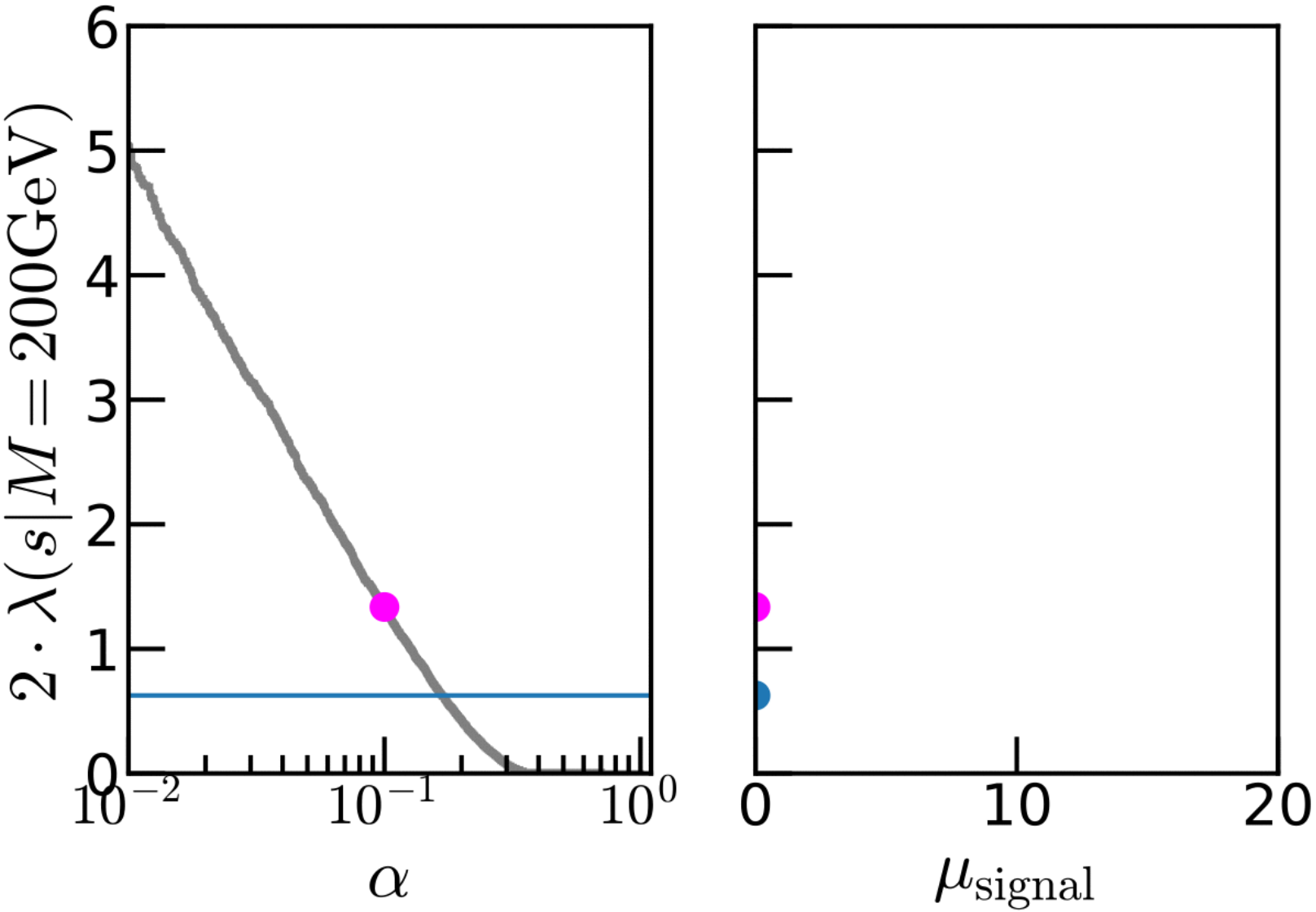
Constructing Confidence Intervals



Constructing Confidence Intervals

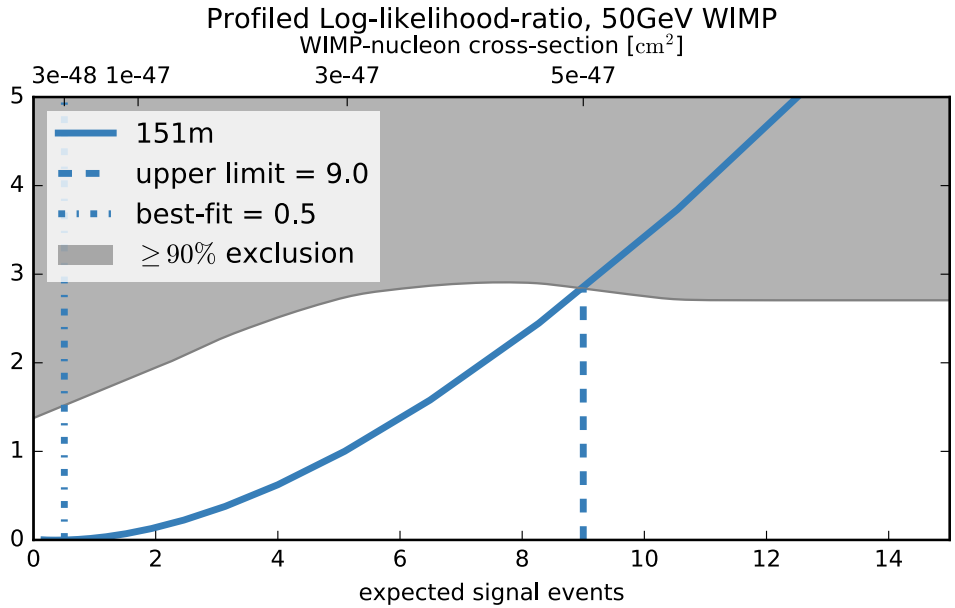


Constructing Confidence Intervals



The profile construction is hypothesis tests!

- The threshold we compute with the profile construction is the area where if our test statistic (the profile likelihood) exceeds it, we reject that hypothesis
- The confidence interval is the union of all signal models we cannot exclude

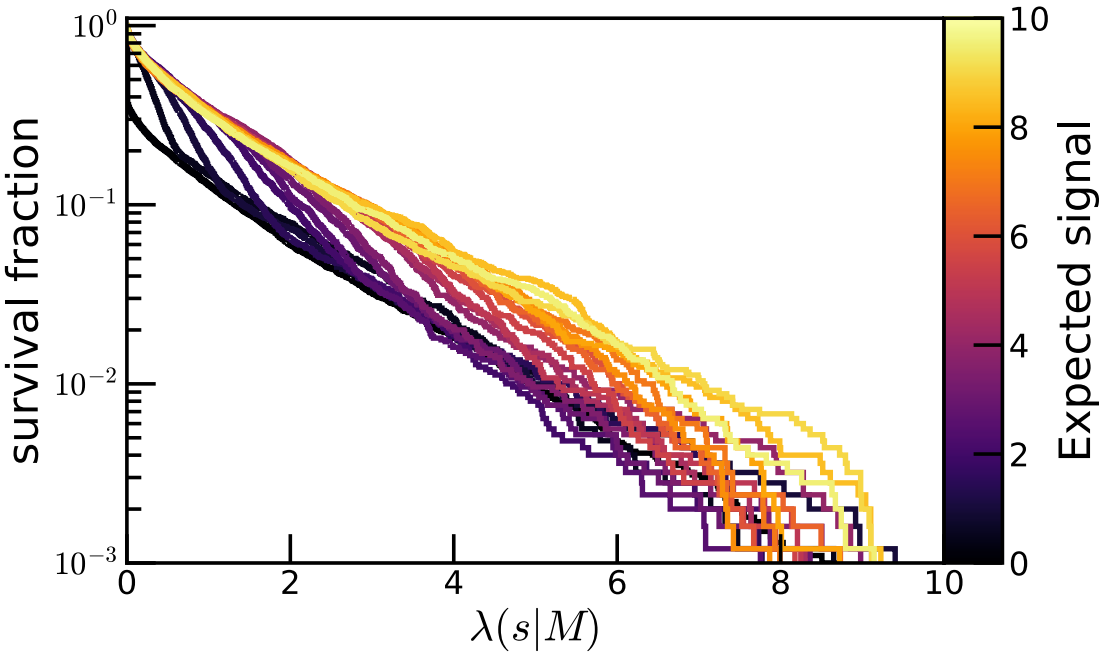


The Profile Construction

What to do with nuisance parameters?

$$\mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b) = \mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) \times \mathscr{L}_{\text{cal}}(\vec{\theta}_b) \times \mathscr{L}_{\text{anc}}(\vec{\theta}_b)$$

- the toy simulations needed to compute the test statistic distribution include randomising all measured parameters
- including calibration measurements and ancillary measurement terms
- Simulations done for a set of signal strengths, and usually also for a range of signal shapes (varying mass or similar parameters)
- However, no firm procedure exists for nuisance parameters
- Fix them to best-fit values and be certain you are not using the true values?
- Randomise them according to uncertainty risks double-counting uncertainties

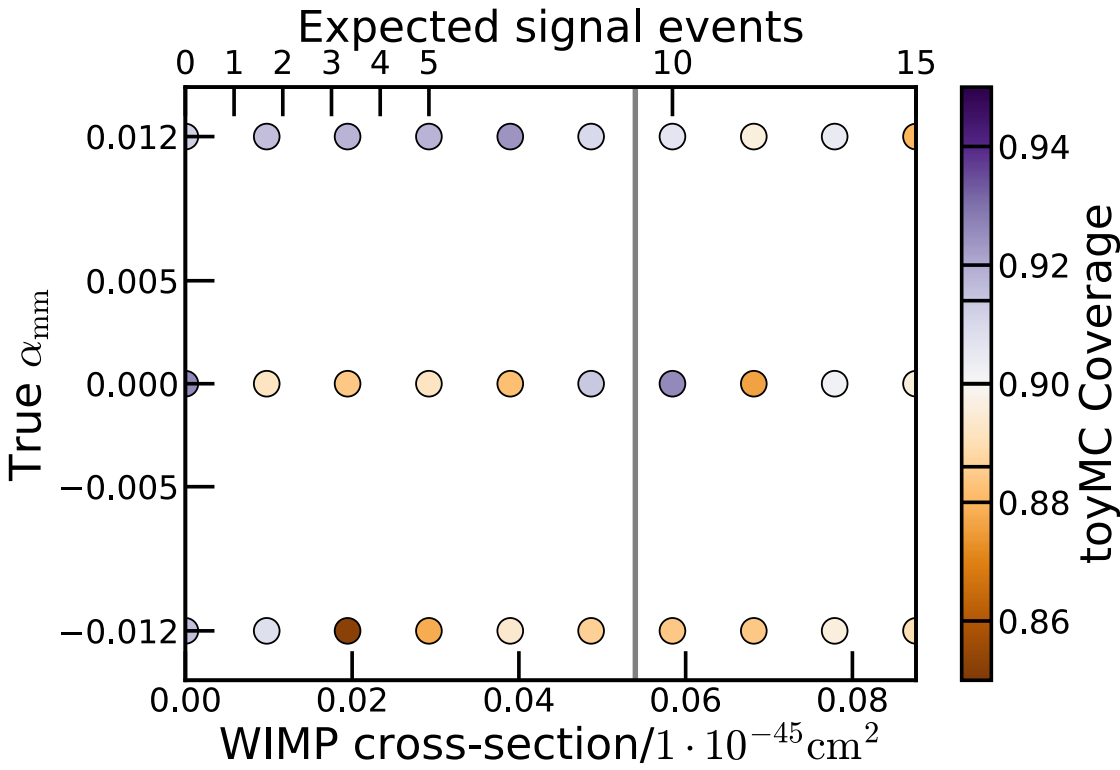


The Profile Construction

What to do with nuisance parameters?

$$\mathscr{L}(s, \vec{\theta}_s, \vec{\theta}_b) = \mathscr{L}_{\text{sci}}(s, \vec{\theta}_s, \vec{\theta}_b) \times \mathscr{L}_{\text{cal}}(\vec{\theta}_b) \times \mathscr{L}_{\text{anc}}(\vec{\theta}_b)$$

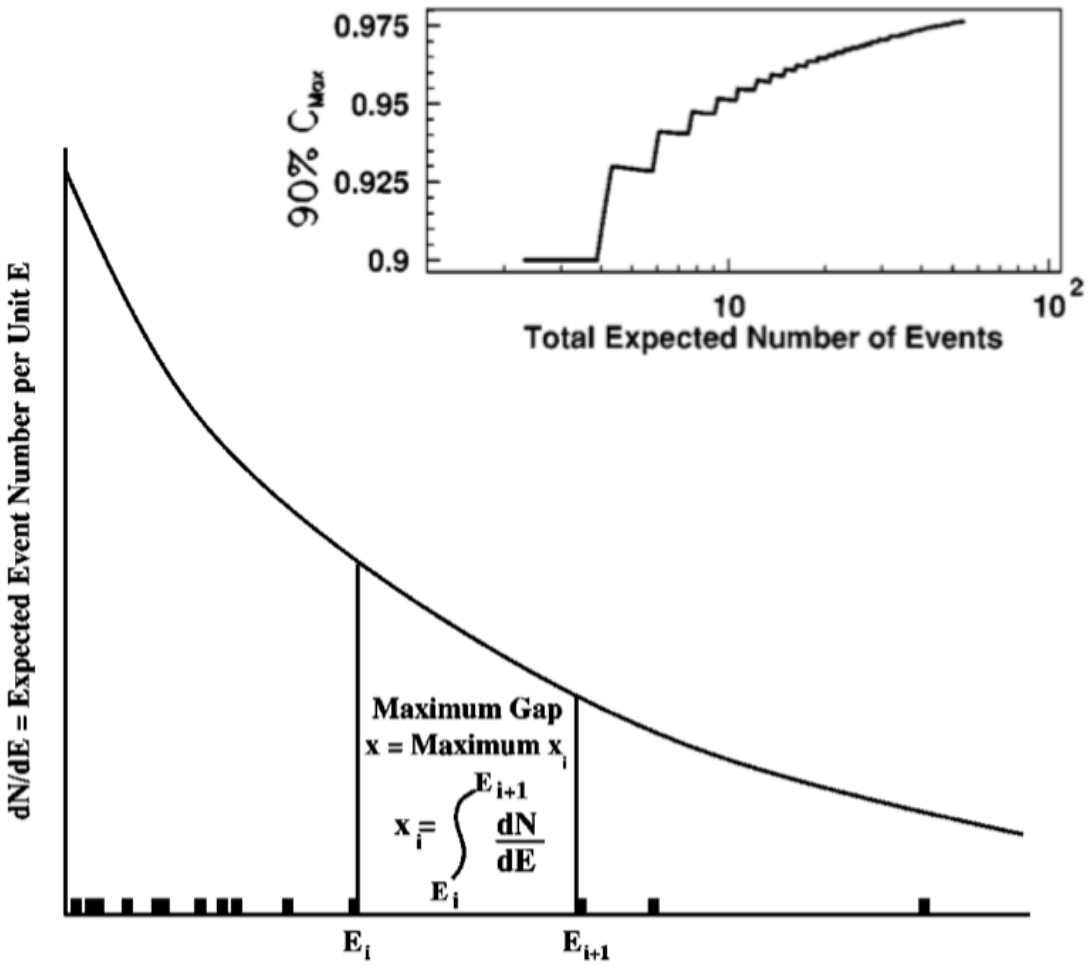
- **XENON used the best-fit values of nuisance parameters**
- **In the latest XENON WIMP search, the robustness of this construction to mis-measuring nuisance parameters was also estimated (right)—changing the value to +-0.012 yielded a percentage point change in coverage**
- **For comparison, the best-fit value was -0.004**



E. Aprile et. al (XENON). XENON1T dark matter data analysis: Signal and background models and statistical inference. Phys. Rev. D, 99(11):112009, 2019. doi: 10.1103/Phys- RevD.99.112009.

Max Gap and Optimum Interval

- If the signal distribution is known along some variable, the maximum gap/ optimal interval method can incorporate this, even in the presence of an unknown background
- Find the space between observed events containing the largest signal expectation, and find the largest signal compatible with this largest “gap”.
- The method can be extended as “optimum interval” where you search for the largest interval containing 0,1,2 etc events
- threshold for the best interval test statistic found via toyMC methods

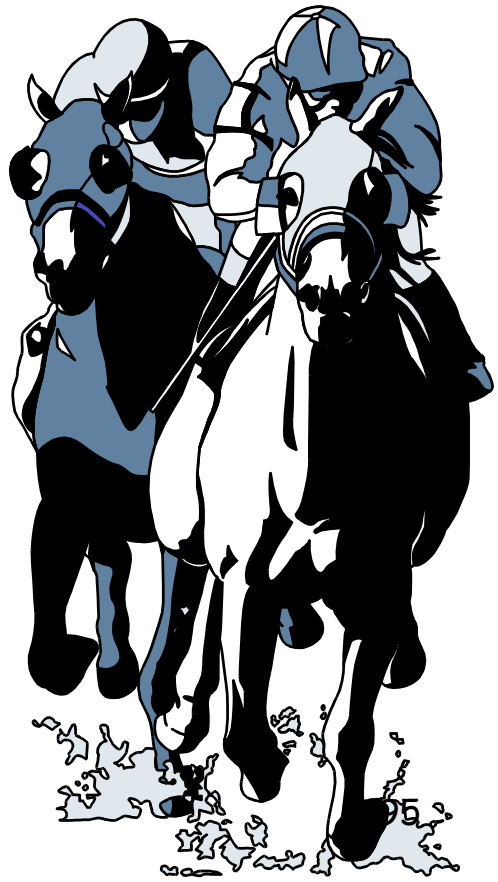


S. Yellin. Finding an upper limit in the presence of an unknown background. Physical Review D, 66(3), Aug 2002. ISSN 1089-4918. doi: 10.1103/physrevd.66.032005.

Bayesian Interpretation

- Probability may also be consistently interpreted as degree of belief— in essence, given everything you know, how much would you bet that something will happen
- However, you have to introduce a *prior*— the degree of belief you held in A before doing the experiment
- This allows the assignment of probability to in principle non-reproducible events
  - in particular the event "this physics parameter is X", so a Bayesian analysis will return a probability for your hypotheses, which would not make sense in a frequentist interpretation

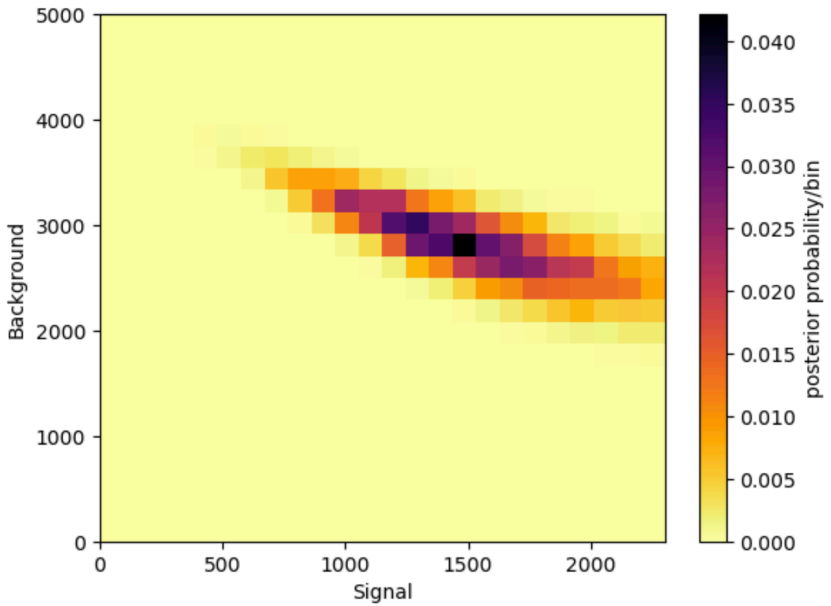
$$P(A \mid D) = \frac{P(D \mid A)P(A)}{P(D)}$$



Credible Intervals

$$p(s,\boldsymbol{\theta}) = \frac{\mathscr{L}(s,\boldsymbol{\theta}) \cdot \pi(s,\boldsymbol{\theta})}{\int \mathscr{L}(s,\boldsymbol{\theta}) \cdot \pi(s,\boldsymbol{\theta}) \mathrm{d}s \mathrm{d}\theta}$$

- Bayesian credible intervals express where the experimenter has a certain total amount of posterior belief
- Computed by finding a region that integrates the posterior to the required value
  - Very often via toyMC methods
- Bayesian and frequentist results *will* converge in the limit of infinite data, but in the meantime, it is not guaranteed
- No analogue to coverage (and it does not require to have it)

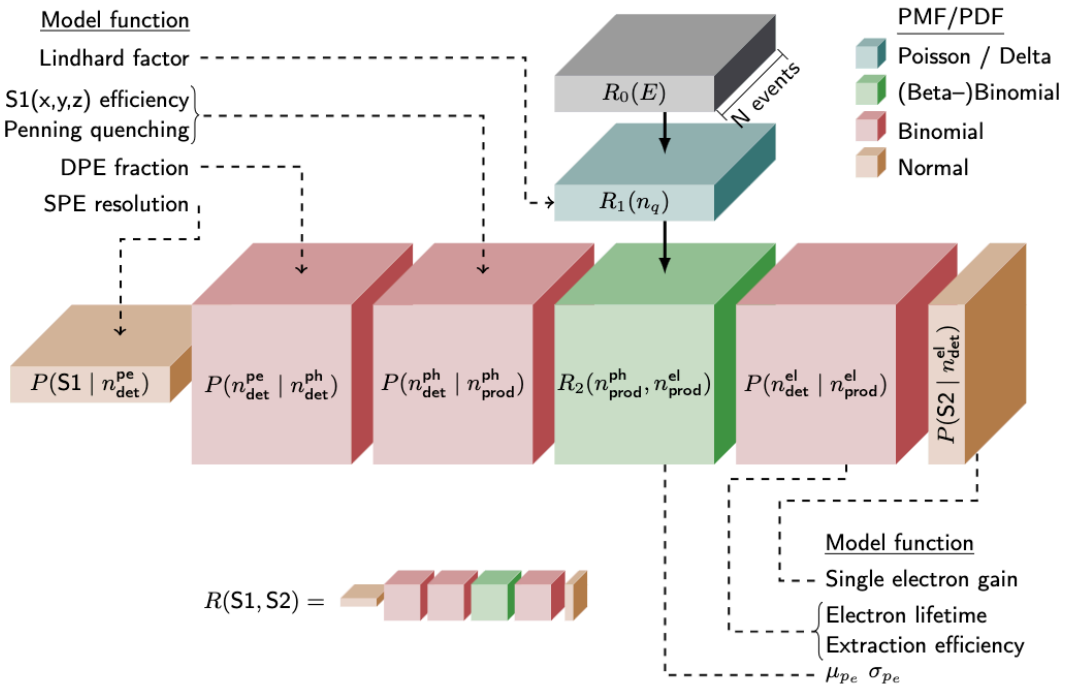


emcee is a good place to start to make a posterior point cloud: https://emcee.readthedocs.io/en/stable/

High-dimensional explicit likelihood

https://github.com/FlamTeam/flamedisx

- A recent alternative proposed to the toyMC-based detector model, is to use the explicit full likelihood expression using efficient matrix multiplication+summation
  - An explicit model avoids troubles in template-morphing,
- This method inherently requires your model to be 6-dimensional in the current formulation, which gives added power, in particular for higher exposures, at the expense of difficulty in model validation
- Computation time is higher for a single best-fit with the full flamedisx model—not a problem for the best-fit but a challenge for toyMC estimates of test statistic distributions



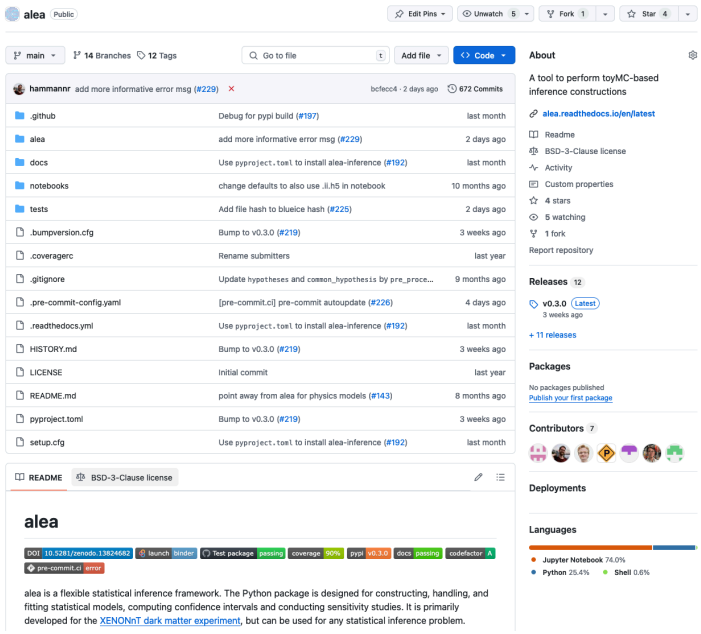
Aalbers, Pelssers, Antochi, Tan and Conrad. "Finding Dark Matter Faster with Explicit Profile Likelihoods" Phys.Rev.D 102 (2020) 7, 072010

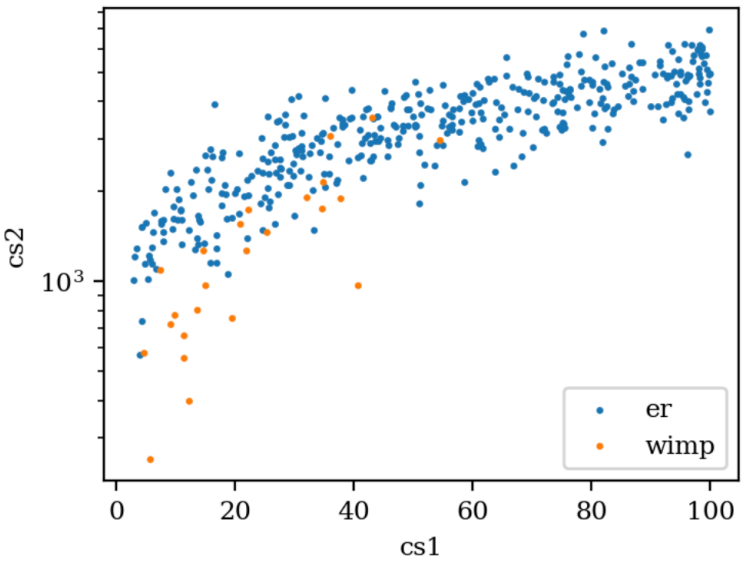
alea— current XENON tool

SoOP 24

https://github.com/XENONnT/alea/

- Provides runners, submitters for toyMC profile construction
- And a flexible framework to add your own likelihood—define a functions for the likelihood and to generate data in a common form
- Includes a full simplified lXe TPC-style likelihood!
- Based on XENON1T and XENONnT SR0 WIMP analyses, with improved maintainable code





Most WIMP events have lower values of `cs1` and `cs2` compared to the ER events. We can use this to discriminate WIMP signal events from ER background events.

For today

- Frequentist confidence intervals
- The profile construction
- A couple of tools

Hands-on session: confidence intervals, profile construction