Statistics for astrophysics

J. Sitarek

2026.07.06, ISS school, Bellaterra

General plan of the lecture, assumptions, biases, disclaimer

- I will introduce basic statistics used in astroparticle physics starting from reminding some basic diploma-level statistics
- The notions, methods etc. are universal and can be applied to whatever type of experiment, but the examples are slightly biased towards VHE gamma-ray astrophysics.
- I am a “practician” of statistical methods rather then “statistics theorist”, therefore apologies for any mental shortcuts, small inaccuracies, missed assumptions etc.

Credits

- Some material is taken from a few similar presentations:
- J. Rico in 2015 MAGIC software school
- A. Babić in 2016 MAGIC software school
- X. Zinquing in ISAPP 2017

“If your experiment needs a statistician,
you need a better experiment.”

Ernest Rutherford

“If your experiment needs a statistician,
you need a better experiment.”

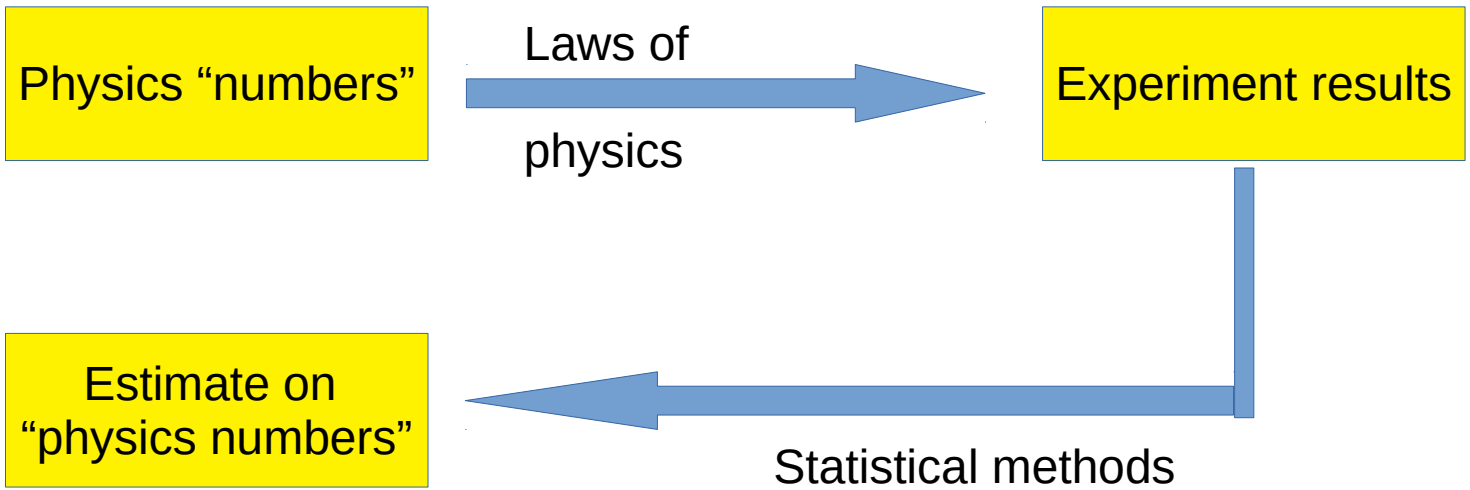
Ernest Rutherford

…

but our experiments already cost millions of EUR and are optimized as good as it gets

Nowadays it is hard to avoid having a statistical analysis treatment in astroparticle physics

Where the statistics comes into the picture



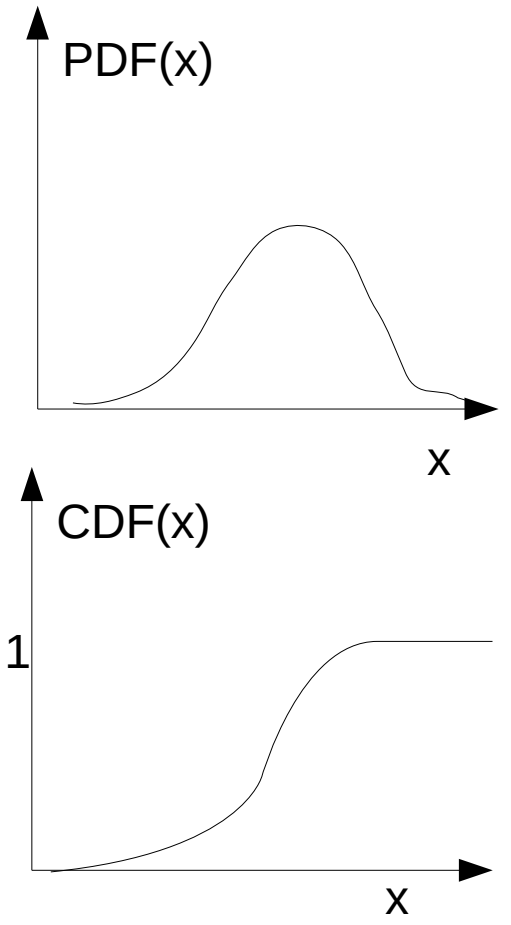
Outline

- Basic notation
- Testing hypotheses
- Likelihood ratio test
- Significance
- Upper limits
- Systematics
- Folding and unfolding
- Frequentist and Bayesian approach
- Example from multimessenger astrophysics

Basic notations

Random variable

- Variable to which we can associate a probability to each value (probability density function, PDF)
- Integral of PDF is cumulative distribution function, CDF, monotonous function going from 0 to 1

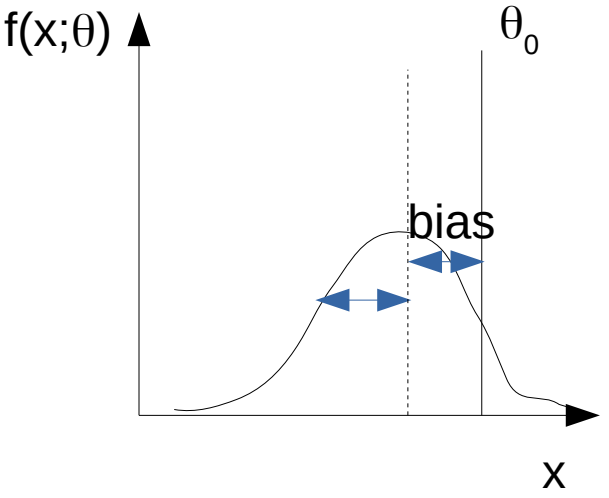


Statistic

- A function of the data alone (no model!)
- Examples:
  - Mean, median, mode, ...
  - Standard deviation, RMS
  - Max/min value, ...
- Statistic can be used to reduce/summarize the data, but also to estimate values of unknown parameters
- Statistic can be calculated from a random variable (to get another random variable)

Estimator

- The parameter value θ is estimated as x based on observational data.
- The estimator can be treated (so-called *frequentist approach*) as a random variable. Its PDF is then computed at the condition (parameter) of θ.
  $f(x \mid \theta) = f(x;\theta)$
  - x (estimate) is the variable
  - θ (physical parameter) is a model parameter
- We want the estimators to be:
  - unbiased (mean = θ_0)
  - have narrow PDF (good resolution)



Statistic and estimator example

- You want to estimate mean of a distribution by (independently!) sampling from this distribution
- Arithmetic mean of the measurements is an estimator of the mean of the distribution
- Since expected value of x_i is μ such an estimator is unbiased (for any distribution!)

$$\hat{\mu} = \frac{1}{n} \sum_{x=1}^{n} x_i$$

Systematic uncertainties/errors

- Systematic error is a kind of estimation bias, a net difference between the derived value averaged over many random iterations
- Systematic uncertainty is our estimate of how big this error is
- Even while systematic errors should be “systematic”, i.e. the same in all the measurements, in astroparticle physics it is often defined as everything that cannot be treated with statistical tools

Examples of systematics

- Examples of systematic errors:
  - Miscalibration of detector
  - Variations/imperfect knowledge of atmosphere (for gamma-ray/CR detectors)
  - Residual backgrounds that cannot be subtracted properly (e.g. nu detectors)

Likelihood

- The variable-parameter relation of estimator can be inverted into likelihood:
  $L(\theta;x) \equiv f(x;\theta)$
  it shows how “likely” is it that physical parameter has value θ if x was observed in the data
- x and θ can be two different types of quantities (e.g. events and flux), they just need to be related in a known way
- Both (or just one) x and θ can be multidimensional!

Likelihood of independent measurements

- Likelihood of a set of **independent** measurements *behaves like* probability it is the multiplication of likelihoods of individual measurements:

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i, \theta)$$

Confidence intervals

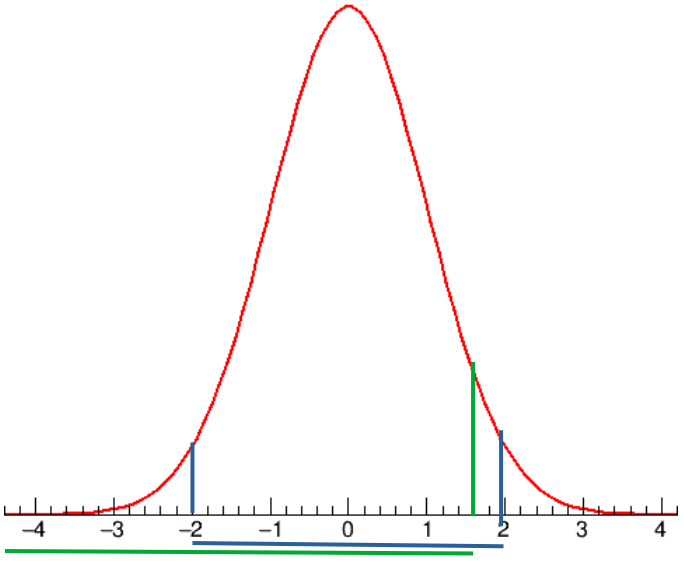
- The derived values of the parameters are expressed as confidence intervals: we state that within a given probability the values should be in a specific range. E.g.:
  - $x_0 \pm \Delta x$ $(x_0 - \Delta x, x_0 + \Delta x)$
  - $x_0 + \Delta x_+ - \Delta x_-$
  - $<x_0$ (upper limit)
  - $>x_0$ (lower limit)
- One should remember that uncertainty of the derived parameter is also a random variable and it has its own uncertainty. The “uncertainty of uncertainty” Is usually relatively high, therefore:
  - Only 1-2 significant digits are given of the number and uncertainty
  - Even if the derived confidence intervals are asymmetric, but the asymmetry is small, it might be still fine to report it as symmetric

Confidence level

- The probability with which we define confidence intervals is called the confidence level (e.g. with C.L. 95% we should be “wrong” 5% of times)
- The higher the confidence level, the confidence interval increases as well.
- Confidence level is often expressed in units of Gaussian sigma: 1σ = 68%, 2σ = 95%, 3σ = 99,7% (if you have two-sided limits)

Interval vs limit

- (upper/lower) limits are one sided, thus we integrate only in one side of the tail
- Always specify if the limits are one or two sided.



Hypothesis

- Hypothesis is a statement about the distribution from which data are coming
- Simple hypothesis: e.g. “ON-source data are coming from the same distribution as the background”
- Complex hypothesis: e.g. “data are coming from a Gaussian with mean = 0 and unknown variance”
- Null hypothesis: the default that our statement is valid
- Alternative hypothesis: typically opposite

Testing hypotheses

“Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth.” (Sherlock Holmes) – Artur Conan Doyle

"Once you ~~eliminate the impossible~~, show that H_0 has probability less than requested whatever remains, ~~no matter how improbable, must be~~ is *likely* the truth."

Test Statistics (TS)

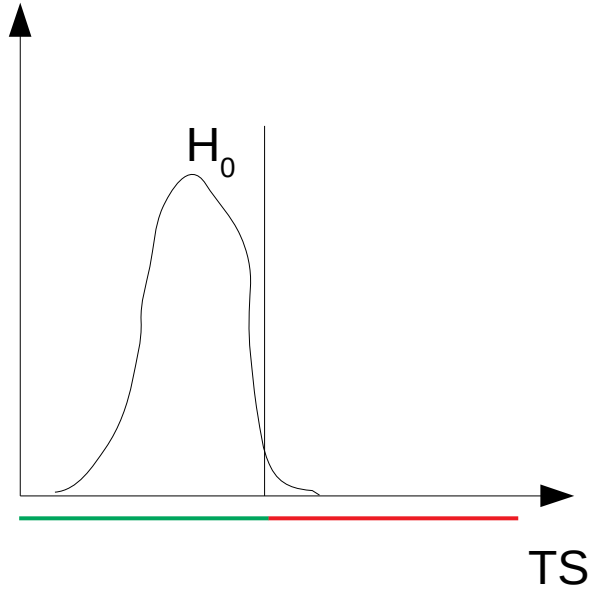
- It is a type of statistic that is used for testing hypothesis
- Different types of TS can be introduced:
  - likelihood (ratio)
  - chi^2
  - Maximum difference of CDFs
  - …
- To test the hypothesis with TS:
  - we need to know PDF of the TS for H_0
  - TS PDF for alternative hypothesis must differ (the more the better) from H_0

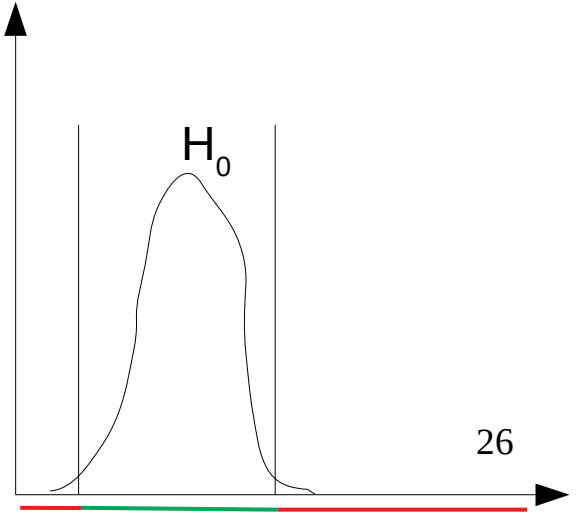
Rejecting H_0 vs proving H_1

- Consider:
  - H_0: the data are consistent with particular distribution
  - H_1: the data are coming from different (unknown) distribution
- Describing mathematically H_1 would be difficult (impossible?) if we do not know underlying distribution, it is much easier to see if we can reject H_0.

Testing hypothesis

- Knowing distribution of TS for null hypothesis we can select the ranges of low-probable values that we use for rejecting the hypothesis (the range will depend on particular TS)
- The probability α that we reject a true hypothesis comes directly from integration of the TS PDF over those ranges

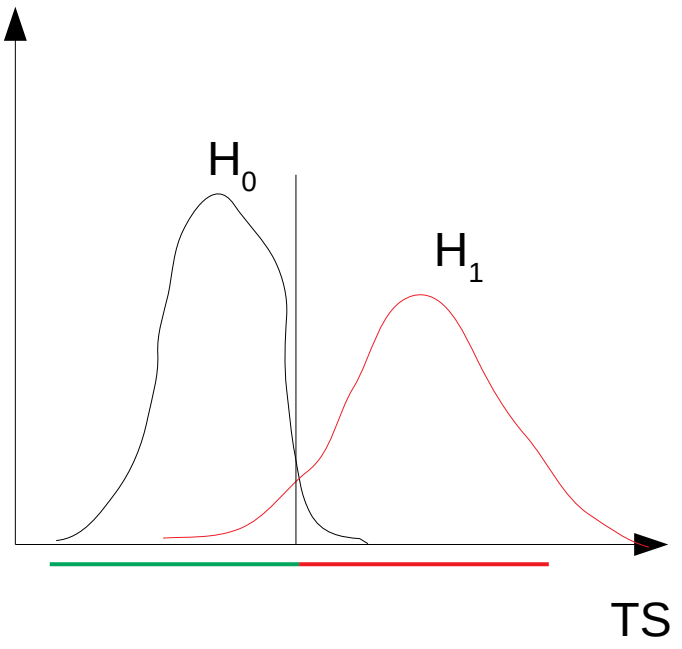




Testing hypothesis

	H_0 accepted	H_0 rejected
H_0 true	OK	Type I error (α)
H_0 false	Type II error (β)	OK

- α comes from the definition of the test
- The probability of how often we would miss rejection of H_0 depends on the details of the tests – we try to minimize β i.e. maximize the power (1-β) of the test



Likelihood ratio test

Most powerful test

- *Will I have to spend the rest of my scientific life trying many test statistics to see which one is the best?* **NO!**
- Neyman-Pearson Lemma states that the maximum power of the test of H_0 against H_1 is obtained by computing ratio of likelihoods:
  $\lambda(x)=f(x|H_0) / f(x|H_1)$ with $\lambda(x) > \lambda_{cut}$
  this is called **likelihood ratio test**
- This is a very general lemma and works with (nearly) any kind of hypothesis
- But depending on the problem $\lambda(x)$ can be computed analytically, with Monte Carlo simulations, or can be very difficult to compute (multidimensional problems)

Parametrized hypothesis

- The hypothesis sometimes can be parametrized with a number of parameters
- The number of parameters can be different between H_0 and H_1 – some of them might be in both
- Example: “Energy dependence of the source emission is of power-law type” (2 parameters), “… of log parabola” (3 parameters), “… of power-law with a cutoff”
- The likelihood ratio can be then written as:
- $\lambda(\theta_0;x)=L(\theta_0;x) / L(\theta_1;x)$
- Where θ_0 and θ_1 are parameters that maximizes the likelihood of H_0 and H_1

Special case: nested hypothesis

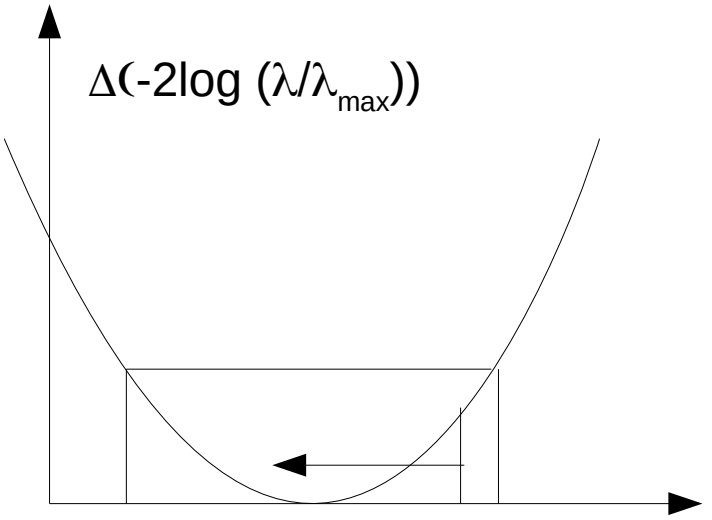
- If H_0 is nested in H_1, i.e. H_1 is more general than H_0 then $L(\theta_0;x) <= L(\theta_1;x)$ (you can always “fit” closer to the points if you have additional freedom)
- According to Wilk’s theorem:
  $-2 \ln \lambda(\theta_0;x)$ follows a χ^2 distribution with m degrees of freedom, where m is the number of extra parameters between H_1 and H_0
- **This provides a very simple analytical solution!**

Back to example with energy spectra

- The data can be described
  - with power-law with best likelihood $L_0 = 0.01$ for $A_0=1.2$, $p_0=-2.1$
  - with logparabola with best likelihood $L_1 = 0.02$ for $A_1=1.3$, $p_1=-2.2$, $q=0.5$
- $\lambda=-2 \ln(0.01/0.02) = 1.39$
- χ^2 with 1 d.o.f. is a square of a Gaussian, so logparabola-like curvature is sqrt(1.39)=1.18 sigma on top of the powerlaw hypothesis
- Even while the likelihood is twice better having the extra parameter that possibly overfits the data we cannot say with high probability that the second model is better

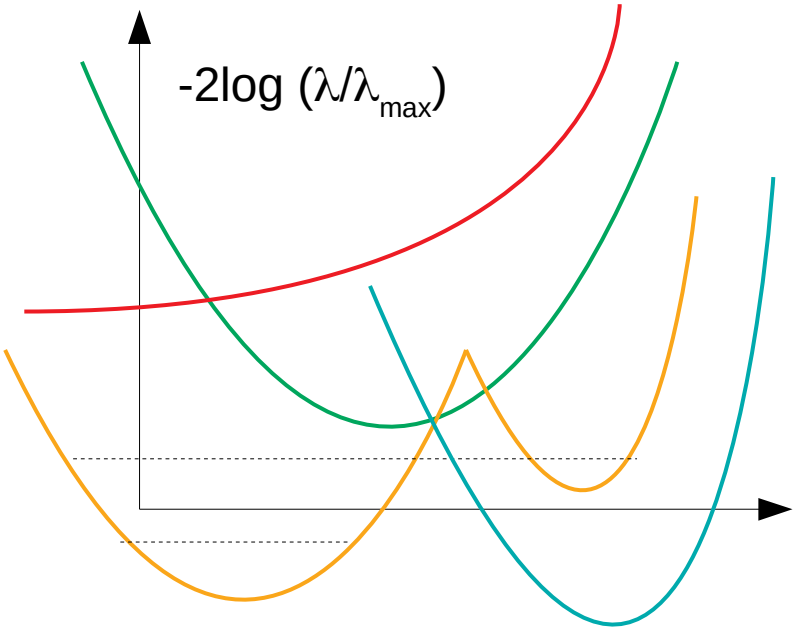
Confidence intervals in LRT

- As mentioned before LRT for nested hypotheses test statistic that is distributed as chi^2 with a given number of d.o.f.
- E.g. putting an interval with C.L. 95% with 1 d.o.f. we need to find a step of 3.84 in TS
- If we want to put a 95% upper/lower limit, it corresponds to 90% confidence interval (5% of tail on each side), i.e. step of TS by 2.71



LRT TS profile shapes

- In closeby neibourhood of the minimum LRT TS should behave parabolic
- Some asymmetry is not a problem (can be a natural consequence of the model better constrained from high/low values)
- For complicated models having multiple minima is possible as well – extra care must be taken to make sure that we are working with global minimum not with local one!
- If no minimum is reached it is a problem (can happen for bounded parameters)



>1 d.o.f.

- For 1 d.o.f. chi^2 distribution is the same as a square of Gaussian, so sqrt(TS) can be interpreted as number of “sigmas” (e.g. TS=25 is 5σ)
- But if you have more parameters this is no longer a case
- E.g. you try to detect a source with unknown normalization and spectral index and get TS of 25 when adding this source to the model, the actual significance if only ~4.6σ

Not-nested hypotheses

- If we want to compare not nested hypotheses (e.g. log-parabola vs powerlaw with cutoff) we can still compute likelihood ratio, but we cannot use Wilk's theorem, so we do not know to what to compare it
- One possibility is Akaike Information Criterion (Akaike 1974):
- The principle is that we want to compare how much information is lost by a particular parametrization of the data
- AIC = $2k - 2 \ln L(\theta;x)$ with k being the number of free parameters
- Relative likelihood of the model: $e^{(AICmin - AIC)/2}$ shows how much better is one model w.r.t. another

Restrictions of AIC

- AIC only shows relative goodness of models:
  - The smaller AIC the better, but absolute scale is meaningless
  - even if one is significantly better than the other it might be that neither describes well the data
- There are many variations of AIC with different penalty factors more appropriate for different cases (low statistics, priors on the model probabilities,...)

Goodness of a model

- What if we have only one hypothesis and want to check if it is consistent with the data?
- We use a test statistics TS which distribution we know how to compute from our hypothesis

$$p = \int_{TS_{obs}}^{\infty} f\left(TS|H_0\right) dTS$$

- TS_{obs} is the value observed in the data
- p shows how (un-)likely it is to get a large TS outlier

Example: χ^2

- Typical example of the quality of fit measure is χ^2 (that follows χ^2 distribution)
- Very low values of probability (e.g. 0.000123) are very unlikely and most likely show that the model is too simple to describe the data
- Value very close to 1 (e.g. p = 0.999877) are also very improbable, but show that the model is “too good”. Normally this means that either the errors are overestimated or there are some neglected correlations

Trials (the look-elsewere effect)

- A way of correcting for how many times you test a single hypothesis with different data.
- If you have N measurements with a p-values of (rejecting the) null hypothesis p_i you can compute the global p-value of that hypothesis using binomial distribution:

$$p = 1 - (1 - p_{min})^N$$
$$p_{min} = \min_i p_i$$

Typical cases of trials

- Observations of a variable source for N days and trying to detect it (at least in one of the days)
- Skymap with an uncertain position of the source
- Unknown energy spectrum of the source – a few different sets of cuts used to try to detect a source

- In the last two cases the trials **are not independent**, however treating them as independent is a conservative (and much simpler) approach

Nuisance parameters

- In order to describe the hypothesis we might need to use additional parameters that are not part of the new theory (nuisance parameters)
- In order not to test the theory for every possible value of those nuisance parameters we can use **profile likelihood ratio**:
  $\lambda_p(\theta_0 \mid x)=f(x \mid \theta_{0,} \nu_{max}(\theta_0)) / f(x \mid \theta_1, \nu_1)$
  $\nu_{max}(\theta)$ maximizes f for θ, while θ_1, ν_1 maximize f globally
- λ_p can be then compared with a χ^2 distribution following Wilk's theorem.
- Such likelihood is often considered together with a second likelihood component dependent mostly on the nuisance parameters

Examples of nuisance parameters

- Detection of source on top of the background:
  - the background count can be considered a nuisance parameter
  - can be constrained with a component describing Poissonian statistics in off-source region
- Systematic uncertainties
  - e.g. systematic error on the efficiency of detection can be considered as a nuisance parameter
  - Gaussian “penalty” factor can be added to the likelihood to describe that high systematic errors are unlikely

Significance

Significance in event counting

- A typical problem is to estimate a significance of an excess of events in ON region over the background estimated from OFF region
- The likelihood can be written as:

$$L(g, b; N_{on}, N_{off}) = \frac{(g+b)^{N_{on}}}{N_{on}!} e^{-(g+b)} \times \frac{(\tau b)^{N_{off}}}{N_{off}!} e^{-(\tau b)}$$

N_{on} – number of events measured in ON region

N_{off} – number of events measured in OFF region

g – estimate of the number of signal events

b – estimate of the number of background events (in ON region)

τ – (assumed to be perfectly known) ratio between the exposures of OFF and ON, it can relate to the difference in observation time, difference in the solid angle of ON and OFF regions, ...

Significance in event counting

- Null hypothesis g=0 (that we want to reject to show that there is a source)
- b is nuisance parameter
- Profile likelihood ratio:

$$\lambda_p(g=0; N_{on}; N_{off}) = \frac{L(0, b_0; N_{on}, N_{off})}{L(g', b_{g'}; N_{on}, N_{off})}$$

- Since there is only one extra parameter (g) χ^2 distribution with $n_{d.o.f.}=1$ is a square of Gaussian
- Significance can be calculated as: $S = \sqrt{-2 \ln \lambda_p(g=0; N_{on}, N_{off})}$

Significance in event counting

- The problem can be solved analytically
- Famous Eq. 17 of Li&Ma 1983:

$$S = \sqrt{-2 \ln \lambda} = \sqrt{2} \left\{ N_{\text{on}} \ln \left[\frac{1+\alpha}{\alpha} \left(\frac{N_{\text{on}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] + N_{\text{off}} \ln \left[(1+\alpha) \left(\frac{N_{\text{off}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] \right\}^{1/2}$$

- Note that S is measuring deviation from pure background, so it is positive no matter if there is an excess or deficit
- For convenience a sign can be added to S following the sign of the excess

Upper limits

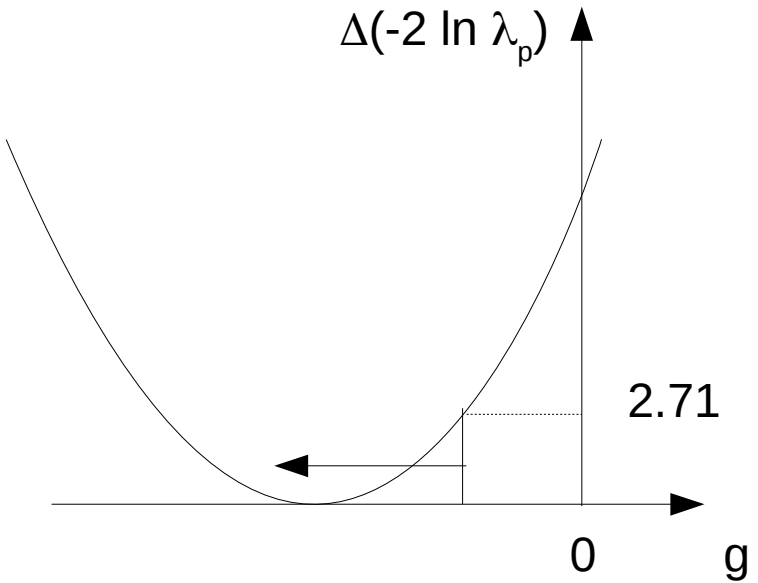
Upper limit on the number of events

- The same likelihood can be used to place upper limit on the number of events
- E.g. for 95% limits find *g* from the equation:

$$-2\ln\lambda_p\left(g;N_{on},N_{off}\right)=2.71$$

Negative upper limit?

- The considered likelihood does not have bounds on the parameters
- Imagine that your data show a strong negative fluctuation (e.g. -3σ)
- If you want to place 95% C.L. limit the $g_{u.l.}$ value would still be negative
- Statistics-wise: it is perfectly fine, such strong downfluctuations are rare, and thus you are allowed to be “wrong” with your limit in the 1-95% = 5% cases
- Physics-wise it does not make any sense – such a limit would exclude any theory (with this 5% chance of error)



Negative upper limit?

- There are different ways to solve the problem with negative upper limits:
  - Set the excess to 0 if it is negative: this avoids the problem (and helps against various systematics!) but is a conservative approach so the limits are “worse”
  - Rolke et al. 2005 proposes some solutions, in particular increasing the excess in steps of 1 until the limit is positive: still the limits can be unrealistically strong

Systematics

Description of systematics

- By definition we do now know the systematics of the measurement
- Sometimes we try to describe systematics in “statistical” way, treating the systematics as random variable.
- We do not know the mean of such variable (we assume 0 or 1 depending if it is additive of multiplicative) – **if we knew it, we would correct for it**.
- We might have some estimate on how big those systematics can be, but normally we do not know what kind of distribution such a systematics random variable would have.

Systematics

- One way of including those effects is adding nuisance parameters describing the systematic effect with some generic distribution (Gaussian, flat, …)
- Our knowledge of systematics is represented as parameters of such a distribution

Systematics in upper limits

- Rolke et al. 2005 and Lundberg et al. 2010 provides prescriptions how to include such Gaussian systematics with event statistics with Poisson and Gaussian distributions
- Depending on specific case either there is an analytical solution, or it can be easily find numerically.

Likelihood with Gaussian systematics on efficiency

$$L(g,b,\epsilon;N_{on},N_{off})=\frac{(\epsilon g+b)^{N_{on}}}{N_{on}!}e^{-(g+b)}\,x\,\frac{(\tau b)^{N_{off}}}{N_{off}!}e^{-(\tau b)}\,x\,\frac{1}{\sigma_\epsilon\sqrt{2\pi}}e^{\frac{-(\epsilon-\epsilon_0)^2}{2\sigma_\epsilon^2}}$$

- ε_0 and σ_ε are assumed to be known!
- LRT is done the same way:

$$\Delta TS(g)=-2\ln\left(L(g,b_g,\epsilon_g;N_{on},N_{off})/L(g',b',\epsilon';N_{on},N_{off})\right)$$

with g', b', ε' maximizing the likelihood globally

Example effect of systematics

- 10 000 events in ON, 10 000 +- 100 events in OFF, 95% one-sided C.L.:
  - No systematics: excess < 233
  - 30% Gaussian sigma systematics on efficiency:
    excess < 268 (15% worse)
- 99% one-sided C.L.:
  - No systematics: excess < 329
  - 30% Gaussian sigma systematics on efficiency:
    excess < 460 (40% worse)
- High C.L. limits need to include also very unlikely case of very large systematics, hence end up with very bad limits.

Combination of data

- The problem: we have a few datasets trying to measure the same quantity, how to combine them?
- The data can be observations from the same instrument, but performed in different conditions, observations of different objects, trying to measure a single, more fundamental quantity, or even observations from different instruments
- Assuming that likelihood method is used we have two possibilities:
  - Sum up the data
  - Multiply the likelihoods

Summing up the data

- Poissonian statistics of individual observations can be summed up

$$N_{ON} = \sum_{i=1}^{n} N_{ON,i} \qquad N_{OFF} = \sum_{i=1}^{n} N_{OFF,i}$$

for τ parameter what is summed up is N_{OFF}/τ

$$\tau = \frac{N_{OFF}}{\sum_{i=1}^{n} \frac{N_{OFF,i}}{\tau_i}}$$

Summing up the data

- The signal and background parameters can be also summed up

$$g = \sum_{i=1}^{n} g_i \qquad b = \sum_{i=1}^{n} b_i$$

- The total exposure needs to be summed

$$A_{eff} T_{eff} = \sum_{i=1}^{n} A_{eff,i} T_{eff,i}$$

Multiplying likelihoods

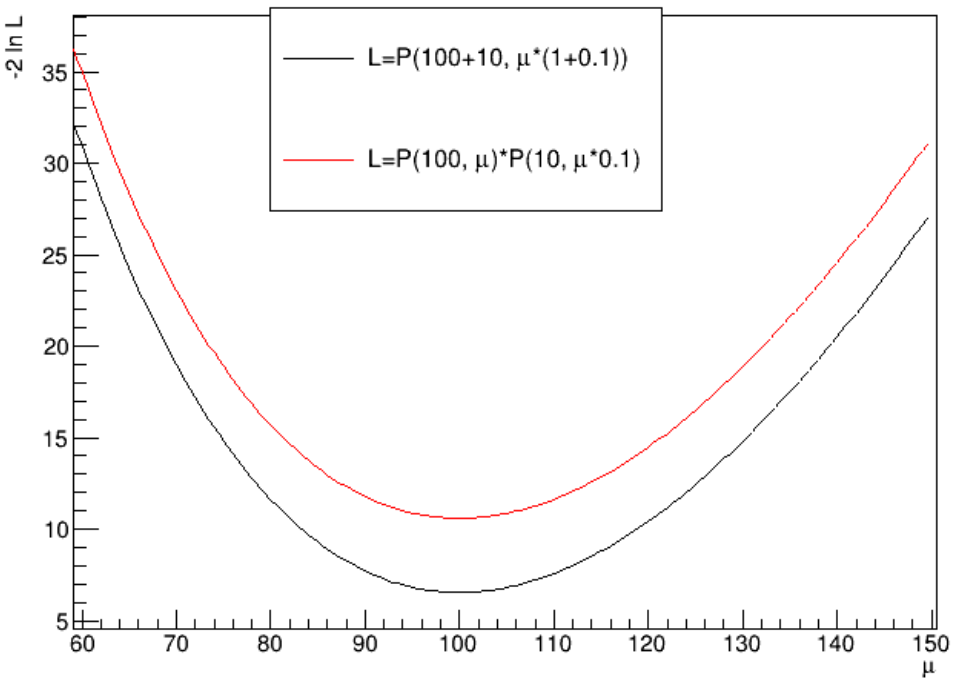
- The likelihood is a multiplication of likelihoods describing individual observations

$$L\left(\Phi, b_1, .., b_n; N_{on,1}, N_{off,1}, \ldots N_{on,n}, N_{off,n}\right) =$$
$$\prod_{i=1}^{n} \left(\frac{\left(g_i(\Phi) + b_i\right)^{N_{on,i}}}{N_{on,i}!} e^{-\left(g_i(\Phi)+b_i\right)} x \frac{\left(\tau_i b_i\right)^{N_{off,i}}}{N_{off,i}!} e^{-\left(\tau_i b_i\right)} \right)$$

- All the parts are connected with a common “flux” parameter Φ

Which one to use?

- Sometimes it does not matter.
- Example of two measurements with Poissonian PDF, either summed statistics or multiplied probability result in just shifted likelihood – the minimum and steepness is the same
- If you are combining very similar data just summing up statistics might be sufficient



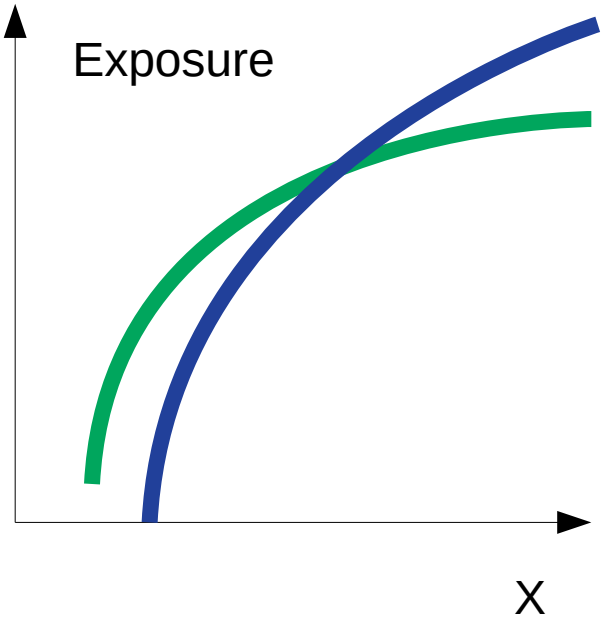
Which one to use?

- If you are combining much different ON/OFF data, the multiplication of likelihoods is a better option

- Example:
  A observed 40 ON events over 10 events of background
  B observed 10060 events over 10000 events of background (poor background rejection)
  if you sum up: 10100 events over 10010 events of background is not significant: the excess of A drown in the background of B

Joint likelihood and systematics

- Imagine combining two datasets measuring at different X values, with with somewhat unknown exposure, especially at low values.
- If joint likelihood is used each dataset is described separatelly, including the “blue” one that have a high uncertainty at the lowest ‘x’ values (statistical is ok, but systematic can be a problem)
- If summing up of statistics is used, at the low ‘x’ the “green” curve will dominate and the result will be safer systematics-wise.



Folding and unfolding

Folding and unfolding

- Consider a distribution X of unknown parameter, and instrument response A causing observed distribution Y

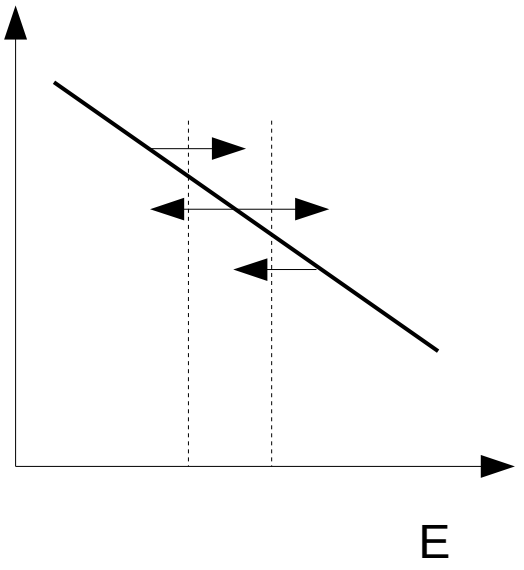
$$Y = A\,X$$

- X and Y can be the same kind of quantity (e.g. true energy and estimated energy, true vs estimated particle charge, true vs estimated direction), or different quantities (true energy and estimated “signal” in the detector)
- Unfolding is a mathematical operation of inverting this relation

$$X = A^{-1}\,Y$$

Example – energy migration

- Typical example is an energy spectrum:
  - for every measured event we assign an energy (estimated)
  - We can bin the energies and comparing then to the IRF calculate the spectra
  - But due to finite energy resolution events migrate from their “true” energy bin to the estimated energy bin
- This causes a number of effects:
  - All features will be smeared out with the energy resolution
  - Since the spectra are usually of power-law (with curvature, cut-off etc.) nature typically there is a net migration (bias) of events from lower to higher energies

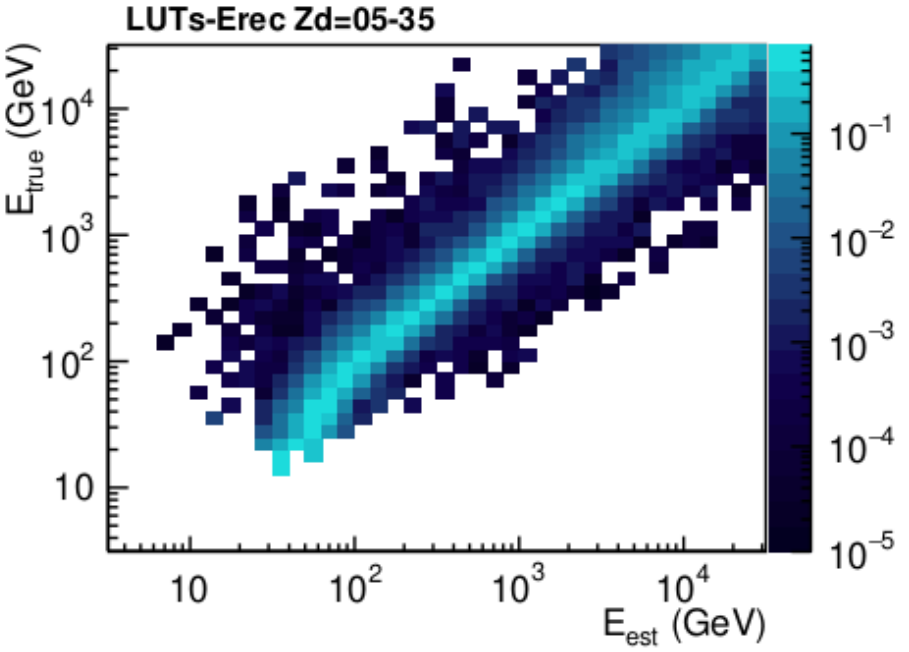


Binned or continuous

- The description of the migration of a given energy E into various estimated energies E’ is a property of the instrument (one of the IRFs), and can be described as a migration function m(E,E’) normalized such that:
  $\int_0^\infty m(E,E')\,dE'=1$
- When dealing with binned distribution a discreet (binned) version can be used:
  $\Sigma_{j=0}^{n} M(E_i,E'_j)=1$
- Binned version, migration matrix $M(E_i,E'_j)$, depends on the spectral shape (true distribution), because there is some migration within the bin as well.
- Migration matrix is usually assumed to be perfectly known.

Example of a migration matrix

- For a resonably estimated parameter we expect a main part of the migration matrix at the diagonal
- Vertical features show degeneracies



MAGIC telescopes,
Ishio & Paneque 2025

Point spread function

- Instrument PSF can be also considered as a similar smearing.
- In this case it is 2D ==> 2D:
  PSF(α, δ, α', δ')

Folding and unfolding

- Two possibilities to deal with event migration:
  - “folding” – include the migration in the likelihood
  - “unfolding” – migration matrix is being inverted (with extra conditions!) and applied to measured distribution

Folding

$$L\left(\Phi, b_{1,}..,b_n; N_{on,1}, N_{off,1}, ... N_{on,n}, N_{off,n}\right) =$$

$$\prod_{i=1}^{n}\left(\frac{\left(g_i(\Phi)+b_i\right)^{N_{on,i}}}{N_{on,i}!}e^{-\left(g_i(\Phi)+b_i\right)} x \frac{\left(\tau_i b_i\right)^{N_{off,i}}}{N_{off,i}!}e^{-\left(\tau_i b_i\right)}\right)$$

$$g_i(\Phi) = \int_{i\,bin} dE'_i \int_o^{\infty} dE\, m\left(E, E'_i\right)\Phi(E)$$

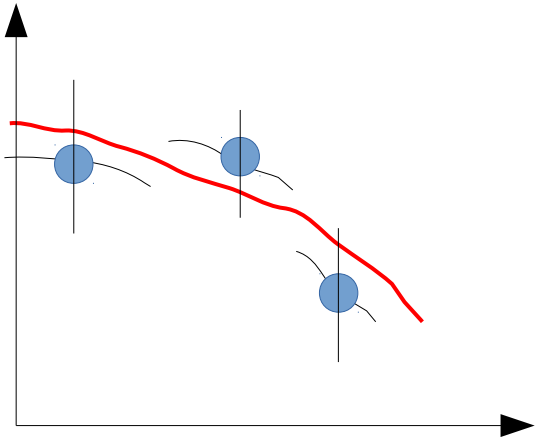
- Likelihood has the same structure and it is multiplied over all estimated energy bins
- The expected signal is calculated by integrating flux folded with migration matrix.
- Dependence on Φ is practically realized by assuming a specific shade described by a few “regular” parameters.

Folding

- If the shape of the function can be constrained folding (forward unfolding) is the best way.
- As long as all the entry parameters (migration matrix, event statistics, function shape) are defined it provides unique solution (except of details of minimalization)
- **The results is the function parameters, not the individual spectral points**

Points with forward folding?

- After the “global” fit is done one can make series of fits using only one estimated energy bin at the time and fixing the spectral shape and only leaving free the normalization
- This way (gammapy is using this approach) individual points are obtained, and they are not correlated
- But: points are computed at the assumption of specific spectral shape so e.g. refitting them to a different spectral shape would not be statistically correct



Unfolding – the problem

- In principle unfolding is a matrix inversion that for square matrices has except of det A=0 unique and analytical solution
- The problem is that:
  - we do not know exact values of number of events in estimated bins (fluctuations)
  - for not extremely small matrix calculation of A^{-1} requires huge number of operations and it is not stable against small differences of A, while m is only known with limited precision (and its matrix form depends on the true spectrum)
- Each true energy bin is a unknown, each estimated energy bin is an equation, we can have cases:
  - Variables < equations
  - Variables = equations
  - Variables > equations

Regularization

- Direct inversion minimizes the variances of obtained result, but will produce artifacts
- Regularization allows for controlled increase of the variance imposing additional conditions (e.g. smoothness of the solution)
- Less regularization: better agreement with the input data, but more “jumpy” solution
- More regularization: smoother solution looking like our model, but diverging more and more from the data

Unfolding

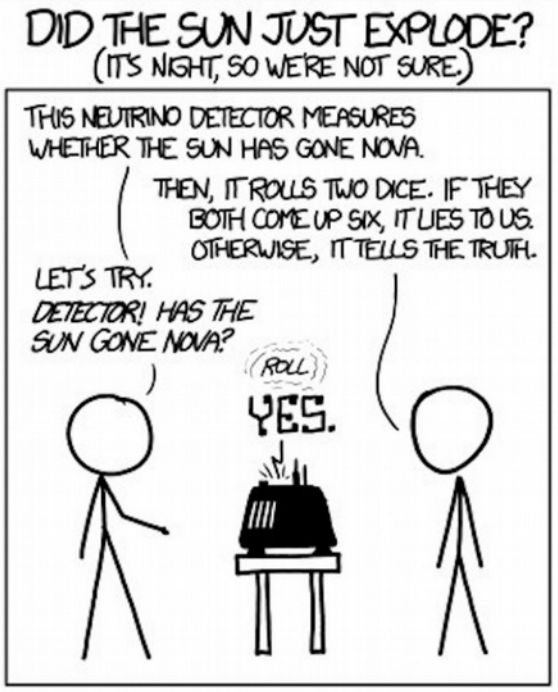
- There are multiple algorithms how to perform unfolding – they correspond to different selections of the regularization function.
- In addition there are different prescriptions to select how much regularization is needed/acceptable.
- The final result will depend on all those
- Bottom line: if your theory can be described by a set of parameters, just do forward folding and do not worry about individual points.

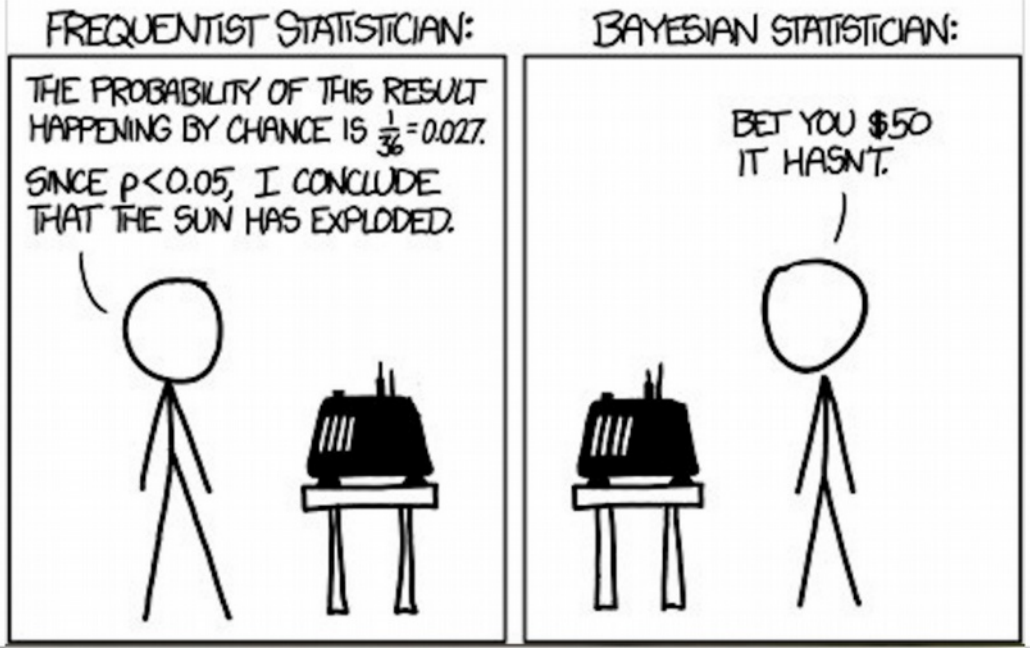
Bayesian vs frequentist

Bayesian and frequentist

- There are two main “schools” of statistics: frequentist and Bayesian
- The two approaches not only provide different answers, but they answer slightly different questions
  - Frequentist: if you make the experiment many times how big is the chance that something happens
  - Bayesian: considering our current knowledge about a parameter and new data, what is the new knowledge

Taken from xkcd





Frequentist

- Frequentist probability is defined as a limit of frequency of particular result

$$P(x) = \lim_{n_t \to \infty} n_x / n_t$$

- The experiment happened only once, but we make a thought experiment of repeating it multiple times and calculate the expected outcome (either analytically or numerically)

Bayesian

- Usage of Bayes theorem for statistical inference

$$P(\theta|d) = \frac{P(d|\theta)\,P(\theta)}{P(d)}$$

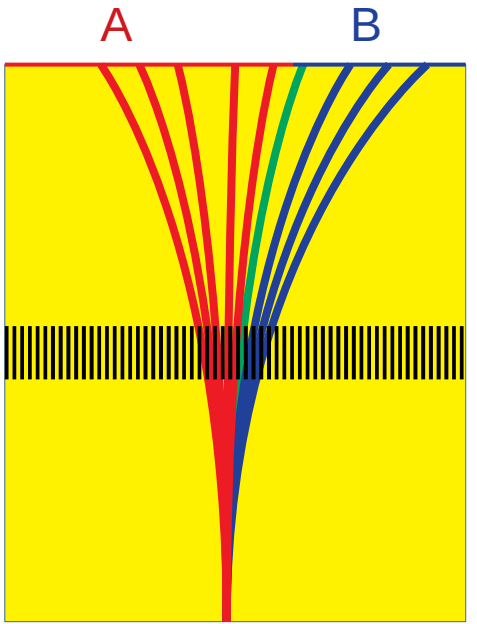
- P(θ) – prior (what we know before the experiment)
- P(θ|d) – posterior (what we know after the experiment)
- P(d|θ) – likelihood function (how likely it is that given parameters produce data that we see)
- P(d) – normalization (marginal likelihood)

Priors

- Priors represent our current knowledge of the parameter
- If we know nothing, commonly a flat prior $P(\theta)$=const is assumed
- **But this is also a selection, and the results will depend on it.**
- E.g. flat distribution in x is not the same as flat distribution in ln x
- Notably for Gaussian likelihood if $P(\theta)$=const, $P(\theta|d)$ is maximized at the same value as $P(d|\theta)$ – the likelihood

The Bayesian Billiard Game

- Carol rolls a ball down the table, and marks where it lands. Once this mark is in place, Carol begins rolling new balls down the table.
- If the ball lands to the left of the mark, Alice gets a point; if it lands to the right of the mark, Bob gets a point.
- The first person to reach six points wins the game.
- Now say that Alice is leading with 5 points and Bob has 3 points.
- What can be said about the chances of Bob to win the game (if you cannot see the billiard table)?



Example from Xia Ziquing presentation

Frequentist (naive) Approach to The Billiard Game

- Five balls out of eight fell on Alice's side of the marker
- Maximum likelihood estimate of p that any given roll lands in Alice's side: $\hat{p} = 5/8$
- The next three rolls must fall on Bob side, which would have probability:

$$P(B) = (1 - \hat{p})^3 = 5/8 = 5.3\%$$

Bayesian Approach to The Billiard Game

- D: data (5 points for A, 3 for B)
- P(p) = 1: flat prior on probability of ball falling on Alice's slide
- P(p|D) = P(D|p) * P(p) / P(D) – posterior on probability p assuming the data
- $P(D|p) = p^5(1-p)^3$: likelihood of data assuming p
- $P(D) = \int_0^1 P(D|p)\, dp = 1/504$: normalization of probability
- $P(p|D) = 504\, p^5(1-p)^3$

Bayesian Approach to The Billiard Game

- $P(p|D) = 504\ p^5(1-p)^3$
- $P(B|p) = (1-p)^3$: probability of winning of Bob for a given p:
- $P(B|D)\ \int_0^1 504\ p^5(1-p)^3 * (1-p)^3 * 1\ dp = 9.1\%$

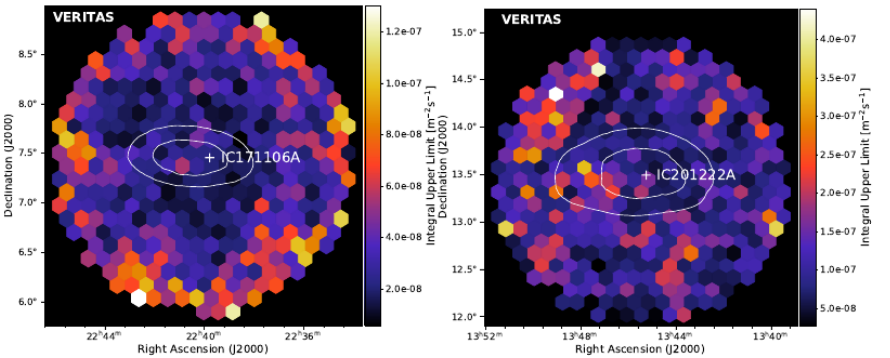
Different results?

- If you made (very naive!) simulations with probability of falling on Alice's side 5/8 you get back the frequentist result
- If you made (slightly tedious) simulations of randomizing the probability p with flat distribution and only selecting the games in which the score after the 8 rolls is 5-3 you get back the Bayesian result
- If you simulate with a different prior (the first roll has a non-flat probability) you get yet another result.

Example from multimessenger astrophysics

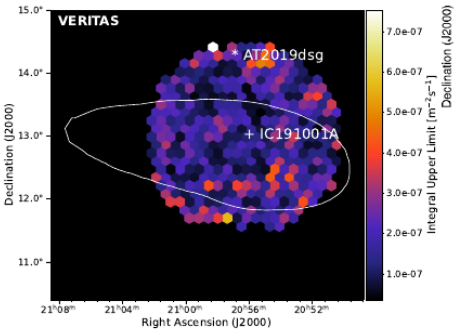
The problem

- Instruments measuring GW, nu, CR, … events normally have bad angular resolution (~ 1-10 deg)
- Cherenkov telescopes have good angular resolution (~0.1 deg) and can be used to try to pin point the location of GW, nu, CR, … events.
- If there is a detection it is relatively easy (just need to estimate chance probability that two things flared independently), but if there is no significant signal...
- In different places in the sky we see fluctuations of ON and OFF, that affect strongly the value of the derived upper limit.
- How to place a proper upper limit?
- There are already a few approaches tried (not yet for IACTs!), but the specifics of the instrument matter

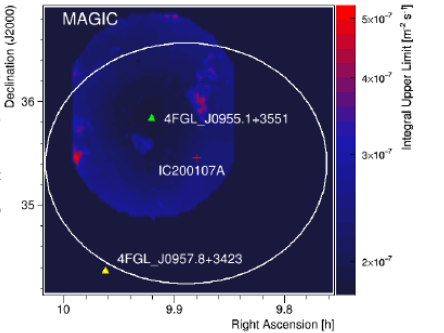


(a) IceCube-171106A (Section 6.1)

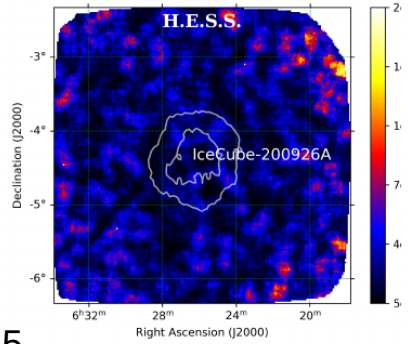
(b) IceCube-201222A (Section 6.11)



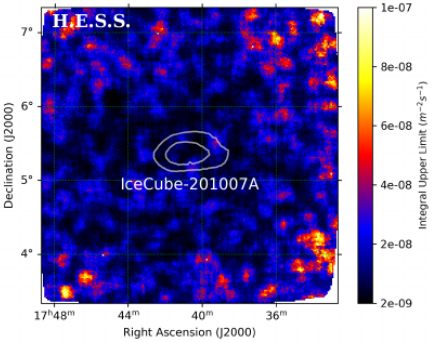
(c) IceCube-191001A (Section 6.6)



(d) IceCube-200107A (Section 6.7)



(e) IceCube-200926A (Section 6.8)



(f) IceCube-201007A (Section 6.9)

Abhir et al 2025

IC vs GW

- Multiplication of likelihood of particular event statistics with a weight following a GW alert probability map
- Unbinned likelihood (counting event by event!)

$$\mathrm{TS} = 2\ln\left(\frac{L(\hat{n}_\mathrm{s}, \hat{\gamma})}{L(n_\mathrm{s}=0)}\right); \quad w = \frac{P_\mathrm{GW}(x_\mathrm{s})}{\Omega_\mathrm{pixel}}$$

$$\Lambda = 2\ln\left(\frac{L(\hat{n}_\mathrm{s}, \hat{\gamma}) \cdot w}{L(n_\mathrm{s}=0)}\right) = \mathrm{TS} + 2\ln(w)$$

Hussain et al. (ICRC 2019)

$$L = \frac{e^{-(n_\mathrm{s}+n_\mathrm{b})}(n_\mathrm{s}+n_\mathrm{b})^N}{N!} \prod_{i=1}^{N} \frac{n_\mathrm{s} S_i(x_i, E_i; \gamma) + n_\mathrm{b} B_i(x_i, E_i)}{n_\mathrm{s}+n_\mathrm{b}}$$

Fermi-LAT vs GW

- Introducing priors on the position following the Bayesian approach
- Unbinned likelihood analysis
- Only considering the region around the investigated position of the source!

$$P(\alpha, \boldsymbol{\delta}, F|D) \propto P(D|\alpha, \boldsymbol{\delta}, F)\, \pi(\alpha)\, \pi(\boldsymbol{\delta})\, \pi(F),$$

$$P(D|\alpha, \boldsymbol{\delta}_h, F) \propto \prod_{D_{\mathrm{RoI}}} m_i(\alpha, \boldsymbol{\delta}_h, F) e^{-m_i(\alpha, \boldsymbol{\delta}_h, F)}$$

Vianello et al. 2017

Methods to apply to IACT data

- Agnostic approach
- Frequentist approach
- Bayesian approach

Method 1 - agnostic

1) Calculate the upper limit in all the bins of the skymap (with Rolke et al. 2005, or gammapy-like wstat)

2) Take the least constraining one within a ROI determined by the alert probability map

Method 1 - agnostic

conservative, but:

- For large ROI there will be thousands of bins, so random 3-4 sigma excesses are to be expected with a very poor U.L.
  (e.g., 10 deg^2 ==> 320 trials ==> fluctuations up to ~2.7σ ==> ~2.6 times worse limit)
- If part of the ROI is at the edge of the IACT acceptance the limits there will be also very poor

- Also remember that if we cover e.g. 90% of the probability map, and derive 95% C.L. in this region, in reality it is only ~85% C.L. (that is common for all the methods)

Method 2 – frequentist

- Modify your test statistics to include prior on position:

$$\mathrm{TS}(x) = 2\ln(L_{src}(x, \hat{\mu}, \hat{b})/L_0(x, \hat{b_0})) + 2\ln(p_{GW}(x))$$

- Find the highest TS in the map of your data
- Generate MCs injecting signals of different strength until 95% of the realization have larger TS than the one obtained from the data
- Hussein et al. (2019) – like but binned

Method 2 – frequentist

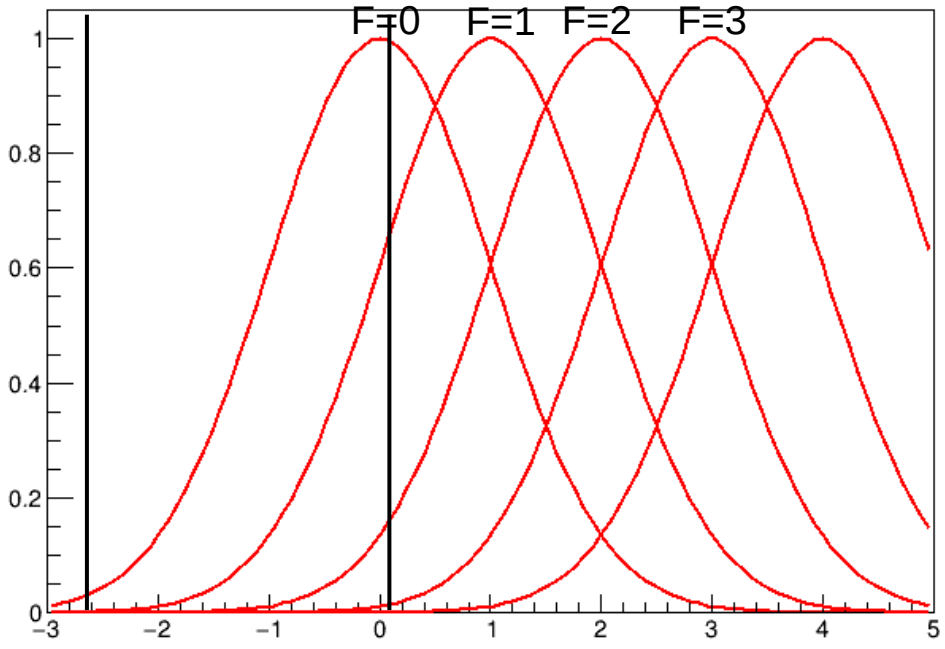
- The method exploits alert probability map
- It also gives what we want for upper limit (in frequentist meaning)

but:

- The two likelihoods are not describing the same thing (one describes only the data other data + GW), so their ratio is not making a LRT therefore one cannot prove with Neyman-Pearson lemma that this is the most powerful method
- The method needs MC which is quite slow

Method 2 – frequentist

- If there is a “significant” downfluctuation in the data (TS is below 1-C.L.) even with a null flux we are still rejecting null flux at >C.L.
- Nominally the upper limit would be negative!
- One can avoid this by clipping TS value in the data to e.g. median of the value from the null flux, but not profiting from negative fluctuations will weaken the power of the method



Method 3 – Bayesian

- We start from the general Bayes theorem:

  p(Model|Data) = p(Data|Model) * p(Model) / p(Data)

- p(Data|Model) we can write using general likelihood terms:
- p(Model) are the priors: probability skymap and flat prior in flux
- p(Data) is obtained from normalization to 1

$$p(x, f) = A \prod_{x'} (P(N_{ON}(x'), \mathrm{PSF}(x, x')\mu + b(x')) \times P(N_{OFF}(x'), b(x')\beta) \times p_{GW}(x')),$$

$$P(k, \lambda) = \lambda^k e^{-\lambda}/k!$$

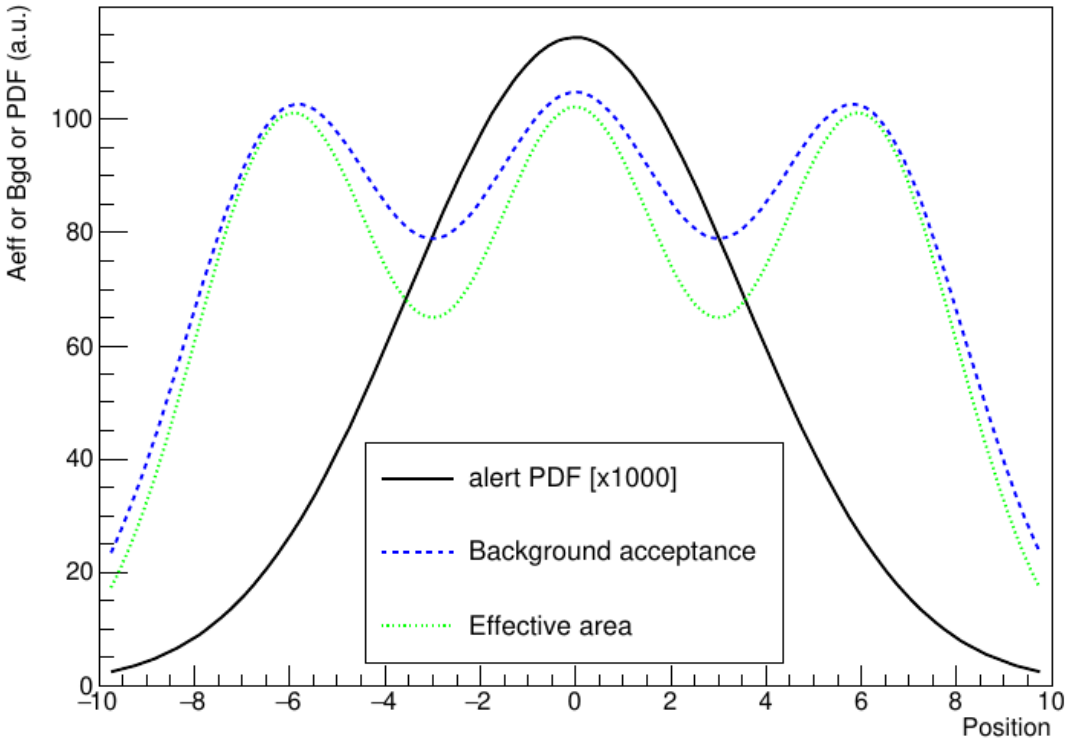
Method 3 – Bayesian

- Parameters:
  - Position of the source: we have a prior, but if there is not detection we are not interested in posterior
  - Flux: the only natural prior is that it is not negative (we can use flat > 0, or some model)
  - Background **in every bin**: nuisance parameter: no prior and we are not interested in its value
- Marginalizing over hundreds-thousands of background nuisance parameters is very cumbersome, but we can analytically derive the best estimate of the background in every bin at the assumption of signal μ.
- Having p(x,f) we can marginalize over ‘x’ to get $p_f(f)$, and integrate it to have a given C.L. limit, C.I. over detection etc.
- **This method has a possible advantage of exploiting the PSF shape!**
- Similar to Vianello et al. (2017), but binned and without simplifications on ROI

Toy MCs

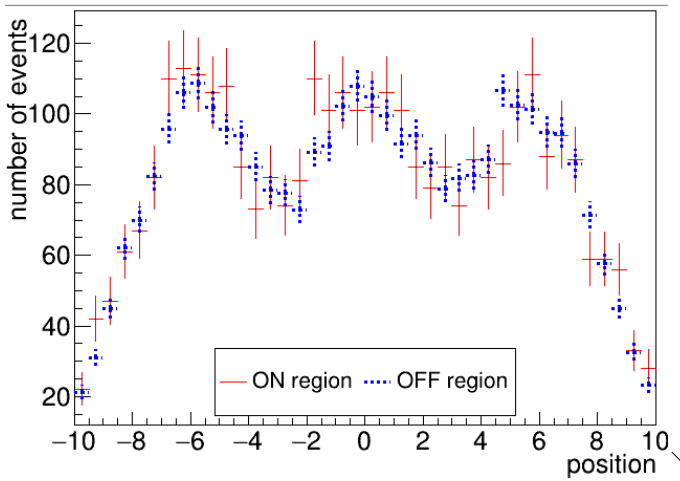
Toy MC case

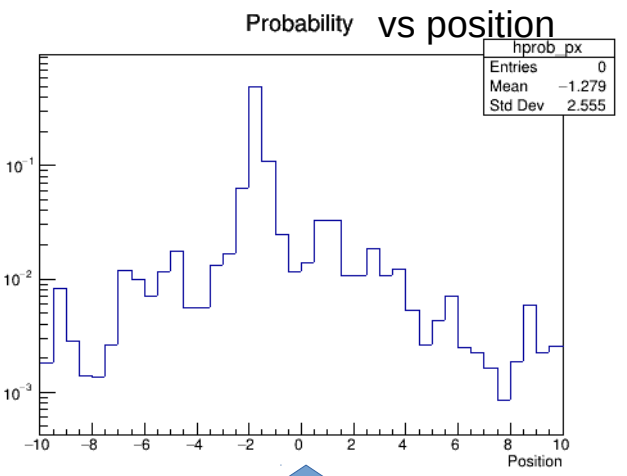
- Simple 1D case
- 40 bins
- Specific shapes of Aeff, background and alert probability map
- Flux in arbitrary units (1 = about 100 events)
- Perfect PSF, and no energy binning (i.e. also no energy migration)

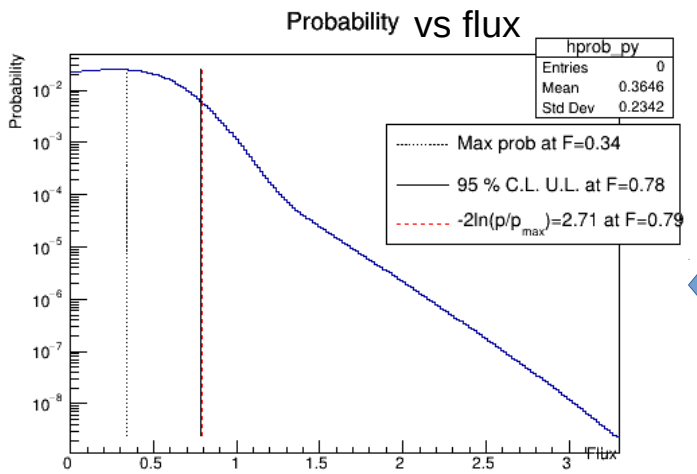


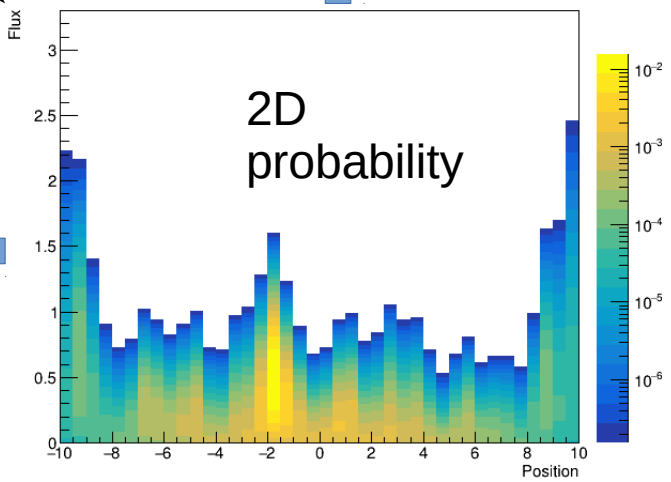
All the plots should be considered preliminary as they are from a paper in JHEAP review

Bayesian method

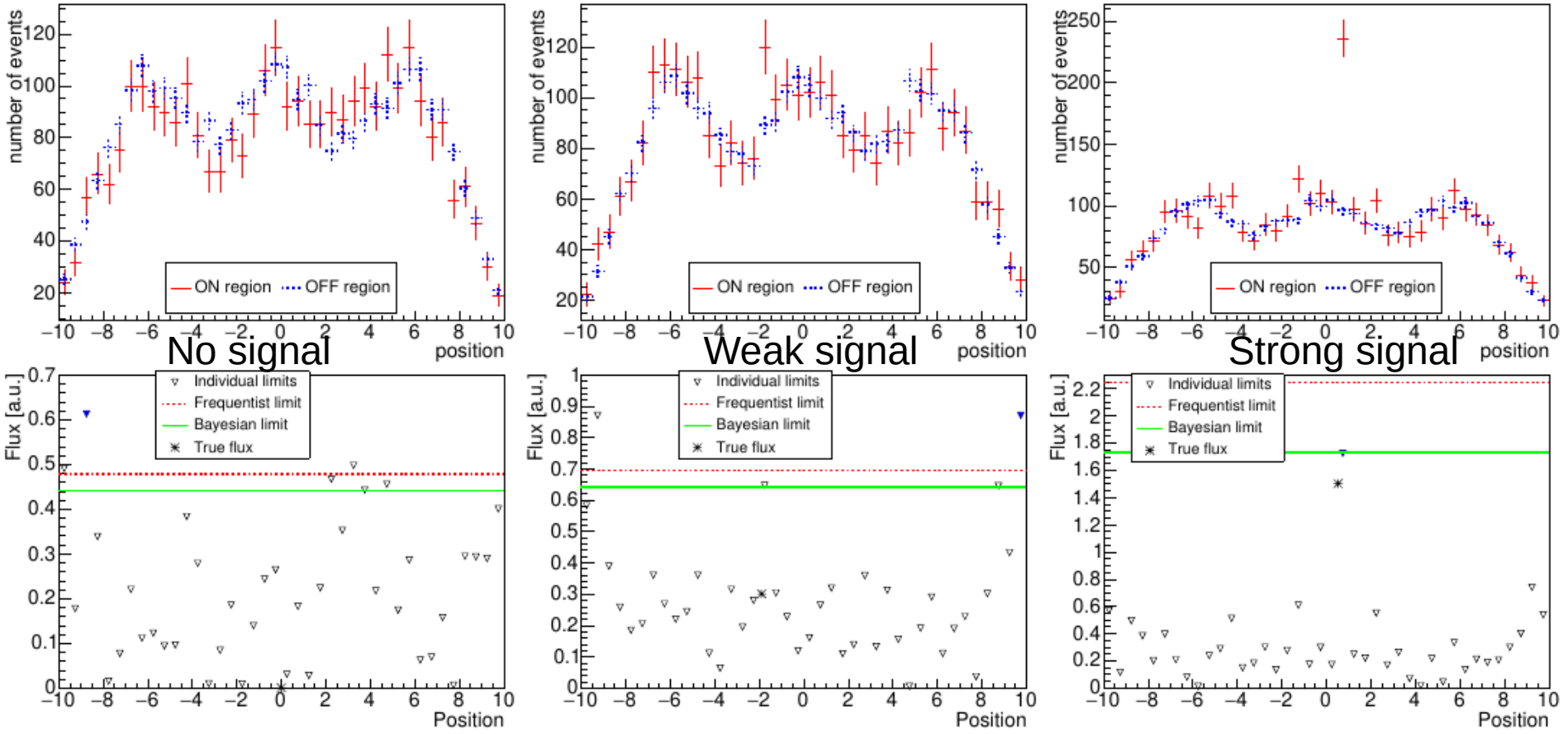






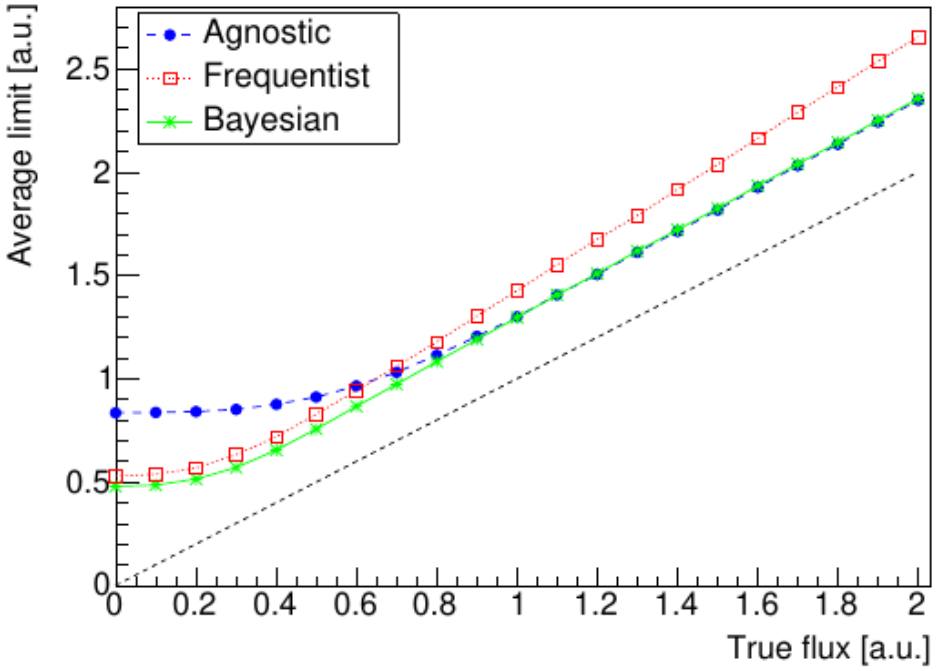


3 example realizations



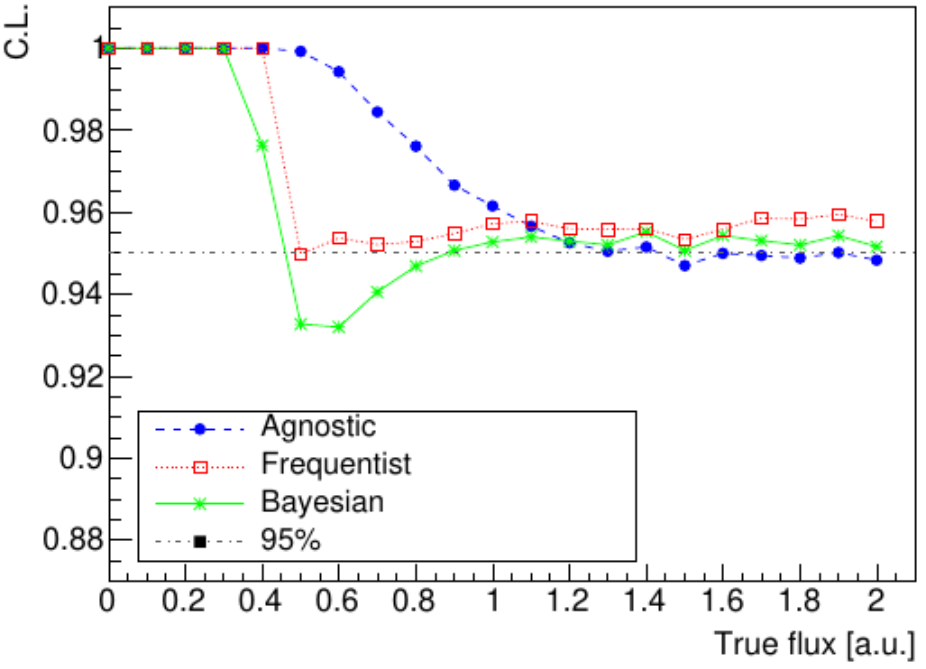
Average limits

- For low fluxes both frequentist and Bayesian methods provide on average much better limits than agnostic
- For high fluxes agnostic merge with Bayesian (posterior determines the position of the excess)
- Frequentist at strong fluxes is suboptimal (it uses simulations over the whole ROI, even while we have a clear detection), but it does not matter much in this case



Fraction of limits above the true flux

- At low fluxes all methods converge to C.L.=1 (for frequentist thanks to the cap at median)
- At higher fluxes for frequentist (nearly) by definition reaches assumed 95%.
- Agnostic is very conservative throughout the whole flux range
- In specific range of true fluxes the “effective” C.L. of the Bayesian can be slightly smaller than requested 95%

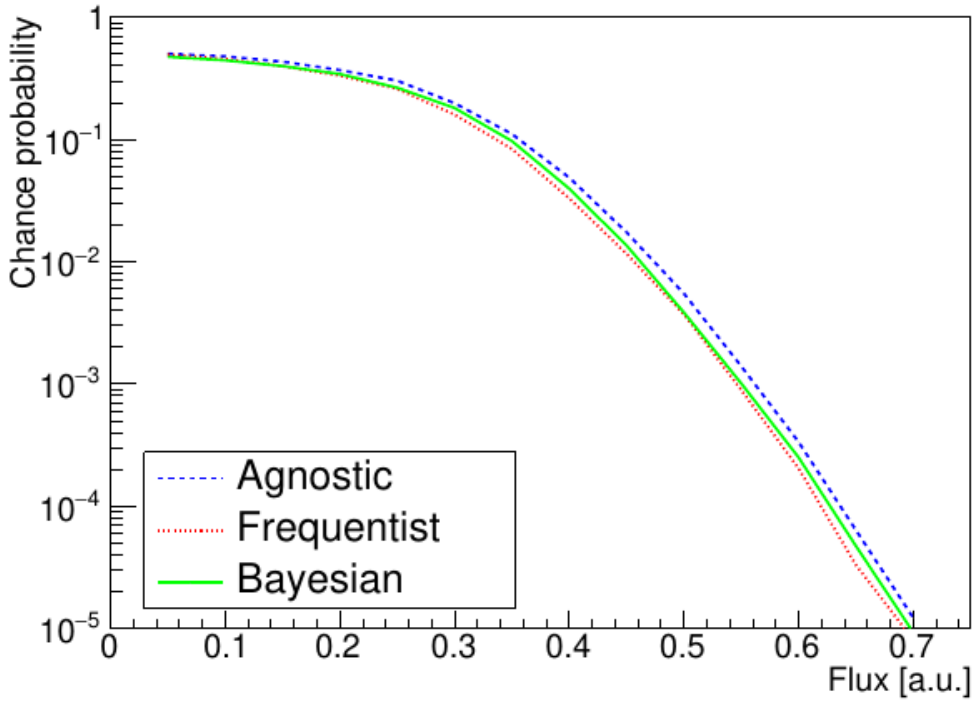


Why Bayesian goes below 95%?

- Different questions = different answers:
  - Frequentist: what flux on average produces TS' quantile of 1-95% (same flux, different event statistics)
  - Bayesian: what is the probability of different fluxes (with a given prior) to produce the event statistics as we see them (different fluxes, same statistics)
- The test of "effective C.L." is done frequentist-style (same flux), so it is a consistency check for that method
- Bayesian-style test is completely infeasible to be performed (only one in each ~10^{100} realization would reproduce the exact event statistics), but in this case we would expect opposite behaviour (perfect for Bayesian and deviations for frequentist)

How to detect the flux?

- Agnostic:
  - Calculate Li&Ma sigma and correct for number of trials
- Frequentist:
  - make huge amount of null simulations and check chance probability of given TS' value
- Bayesian:
  - No strict notion of significance, one can compute Bayes ratio between different hypotheses.
  - One can modify the prior to allow also negative fluxes and integrate the probability of negative flux, but this one does not have a flat distribution for no flux case and needs to be calibrated in "frequentist" way



Frequentist and Bayesian are quite comparable and better than agnostic

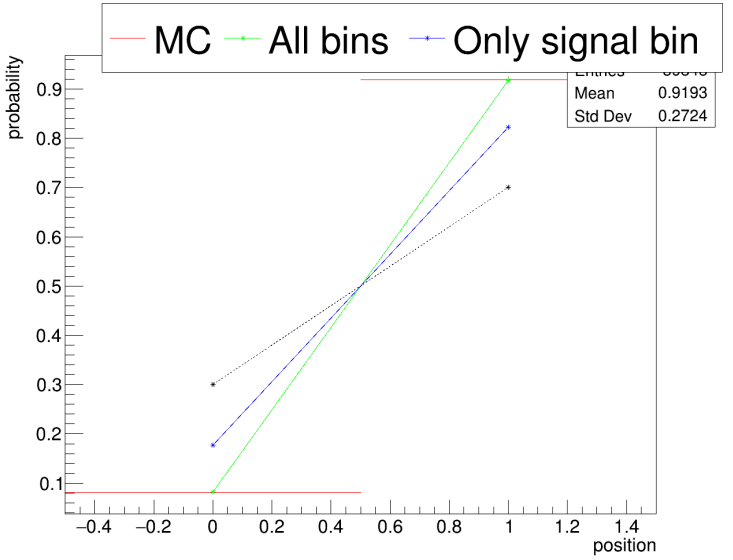
Can we use only signal bins in Bayesian?

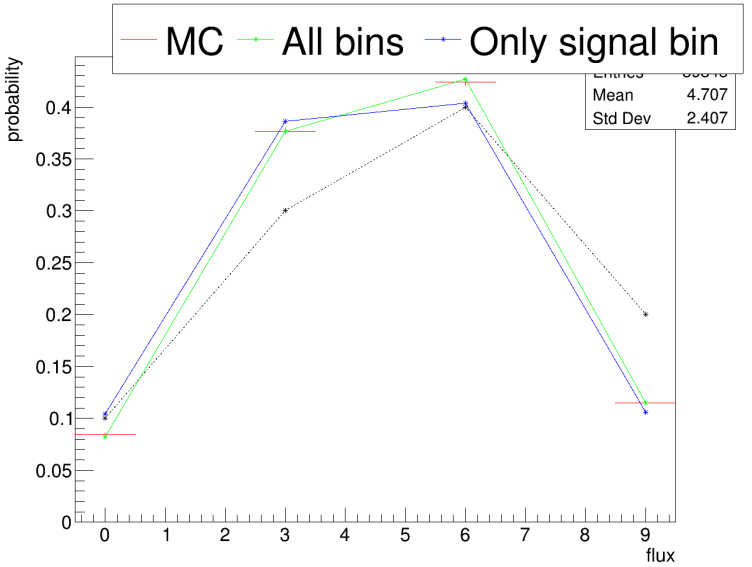
$$p(x, f) = A \prod_{x'} (Ps(N_{ON}(x'), \mathrm{PSF}(x, x')\mu + b(x')) \times Ps(N_{OFF}(x'), b(x')\beta) \times p_{GW}(x')),$$

- In Bayesian method we integrate the probability of all the bins in the skymap, but most of them are just noise.
- Do we need to do it or can we limit it to only to the bin(s) where we consider the signal?
- Vianello et al. (2017) limits it

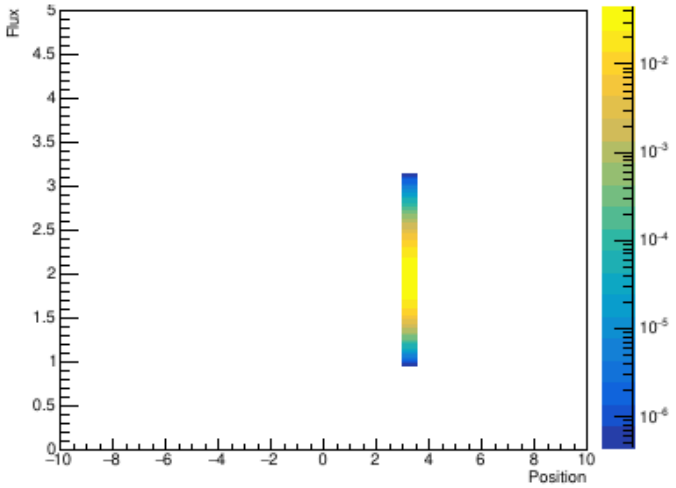
Simple test

- 2 position bins and 4 flux states with single realization drawn from it.
- Validation is done by huge number of MCs simulated according to the priors, but selecting only the realizations that have exactly the same statistics in all the bins
- For “all bins” this is a validation of the Bayes theorem
- Limiting to only signal bin diverges the results from the MC test.

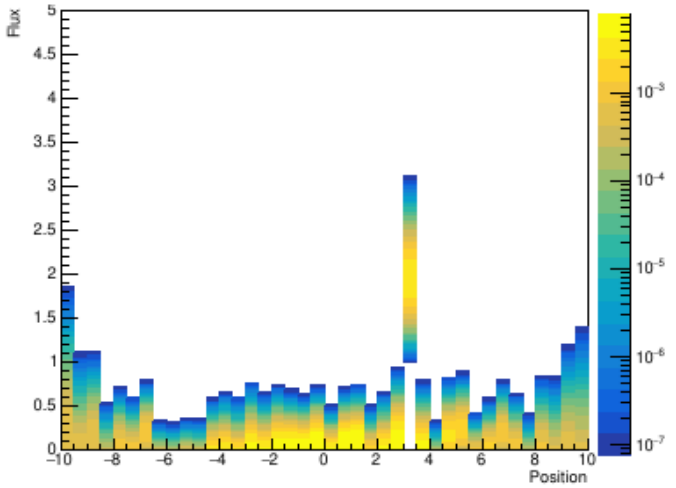




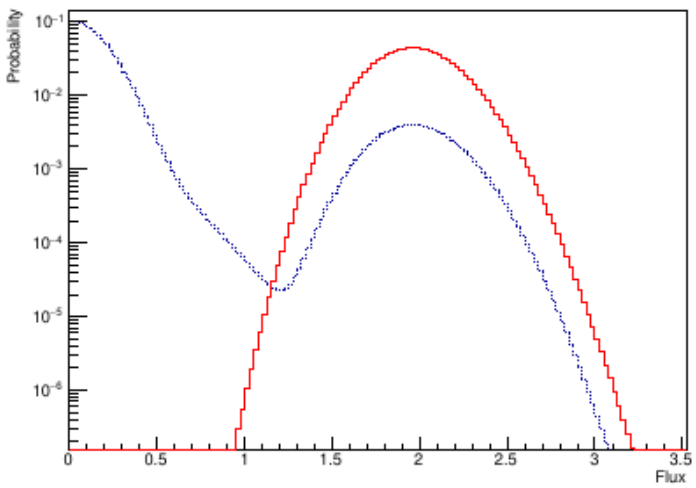
Can we use only signal bins in Bayesian?



All bins considered: probabilities of having signal anywhere else are strongly suppressed



Only signal bin considered: probabilities of having weak signal are high in all non-signal bins.

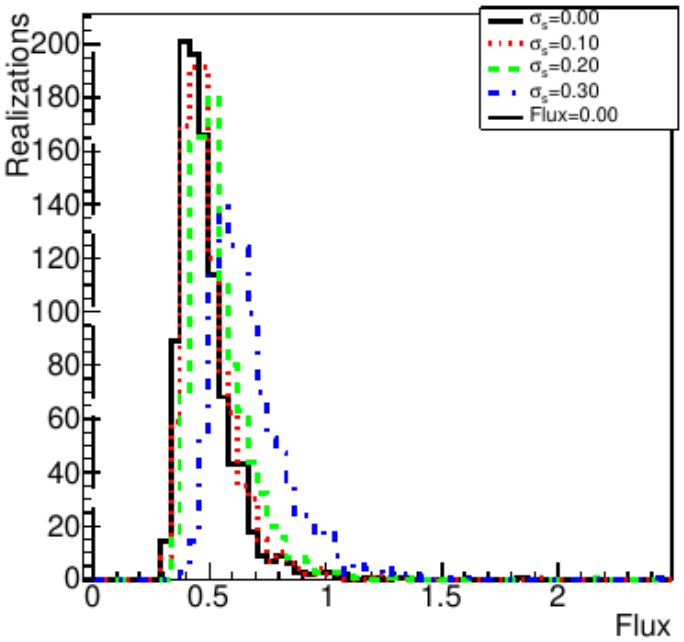


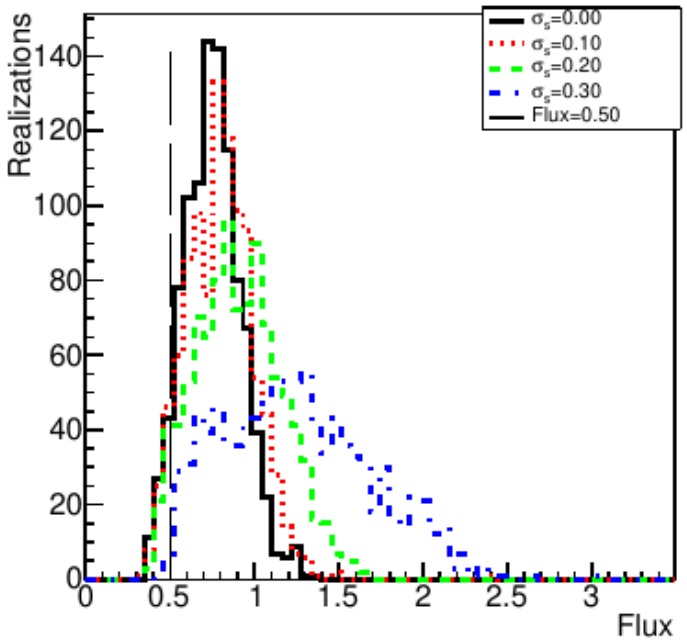
Marginized probability: For considering only the the bin of the putative signal it shows a large probability for 0 flux

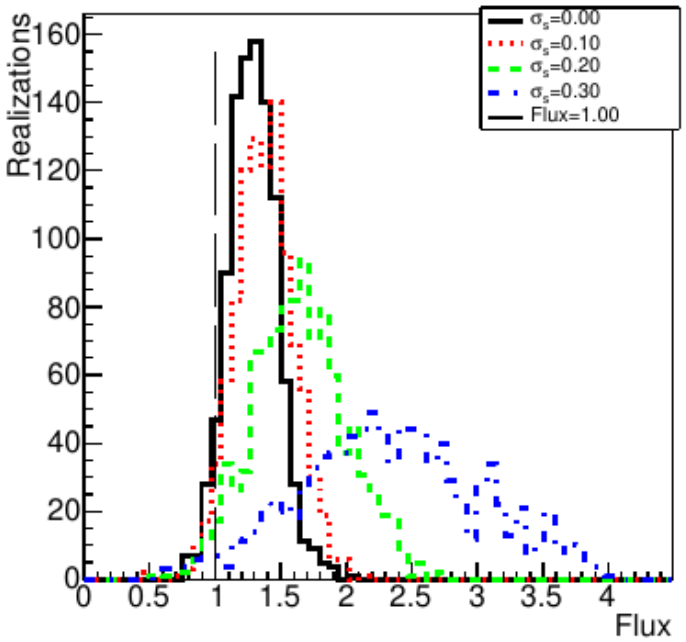
Systematics?

- How the systematics can be included in various methods:
  - Agnostic: Rolke 2005 already considers systematics, so also least constraining limit with systematics can be selected
  - Frequentist: systematics can be included when generating MC realizations at different flux levels
  - Bayesian: systematics can be added as a nuisance parameter (to be marginalized over)

Example of Bayesian with systematics





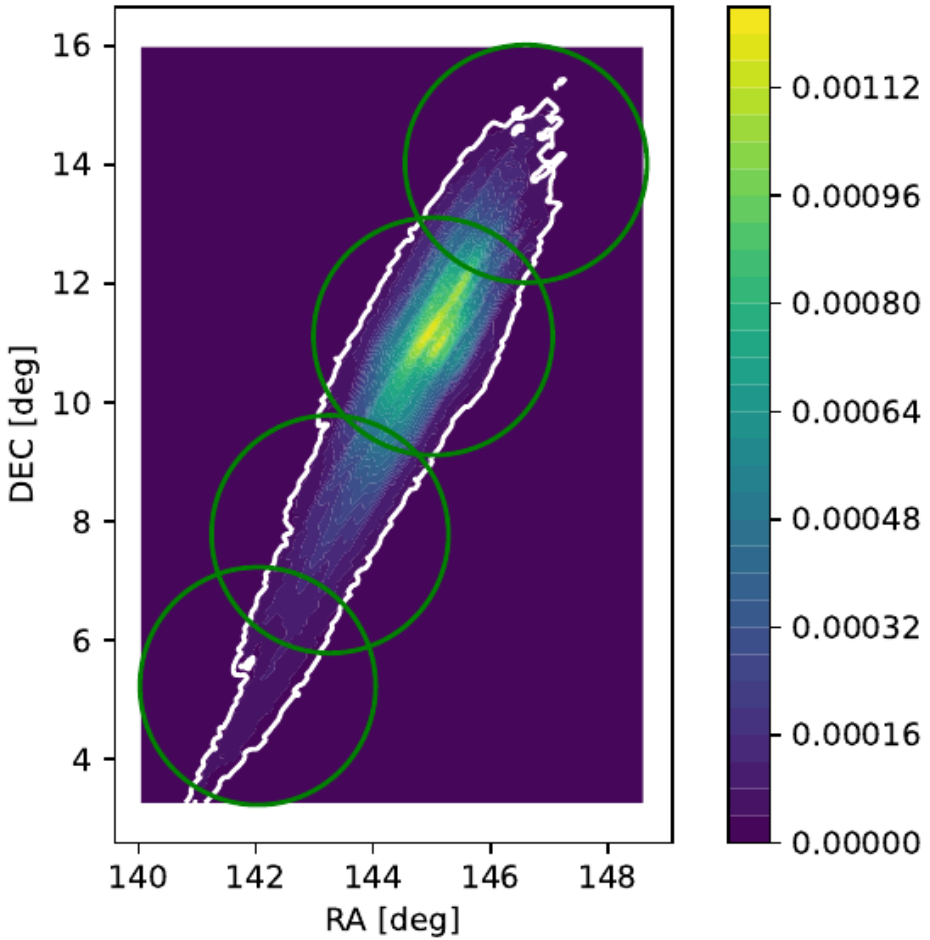


- Gaussian systematics on collection area with sigma of 10, 12 or 30%
- Large systematics can spoil considerably the limit!

Full simulations

Full simulations setup

- S250328ae probability map from tilepy
- 4xLST at zd=20 deg IRFs from Prod 5
- Pointing positions calculated with tilepy (using 2 deg effective FoV radius)
- Skymap is using with 0.07 deg x 0.07 deg binning. For possible position of the source we consider only the bins covered by 95% contour region of the GW and within 2 deg radius from pointing positions (~4000 bins)
- The energy binning is matching this in IRFs: 5 bins per decade in E_{est} and 10 bins per decade in E_{true}
- **Energy and position are treated as discrete values (however averaging over the skymap pixel area of the expected flux is done)**
- For the source model as example we use a HEGRA-like Crab straight power-law with index -2.62



Going from toy case to full simulations

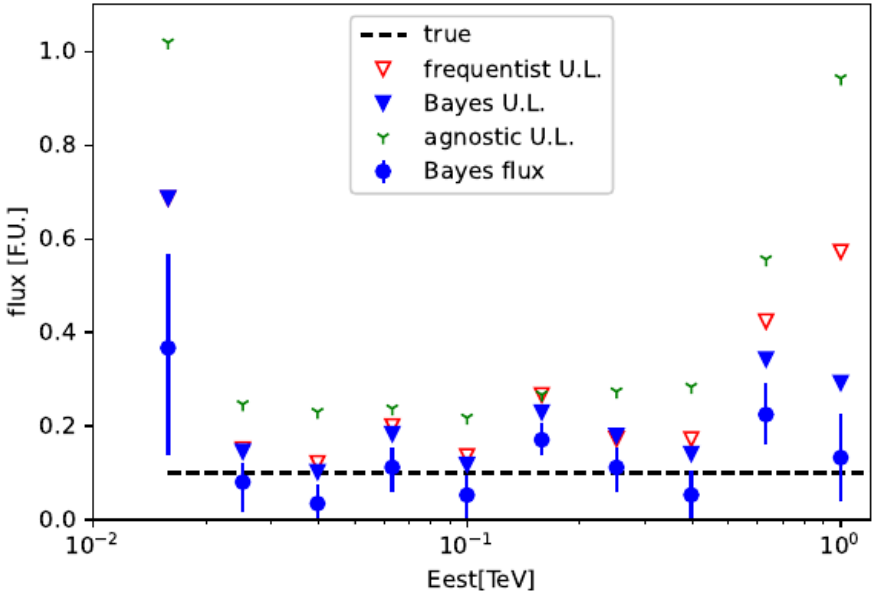
- Same formula can be used, but we need to fold the expected signal over energy migration (which depends on the location in the skymap!):

$$\mu_{i_{est}} = \sum_{i_{true}} A_{eff,i_{true}} t_{eff} F_{i_{true}} M_{i_{true},i_{est}}$$

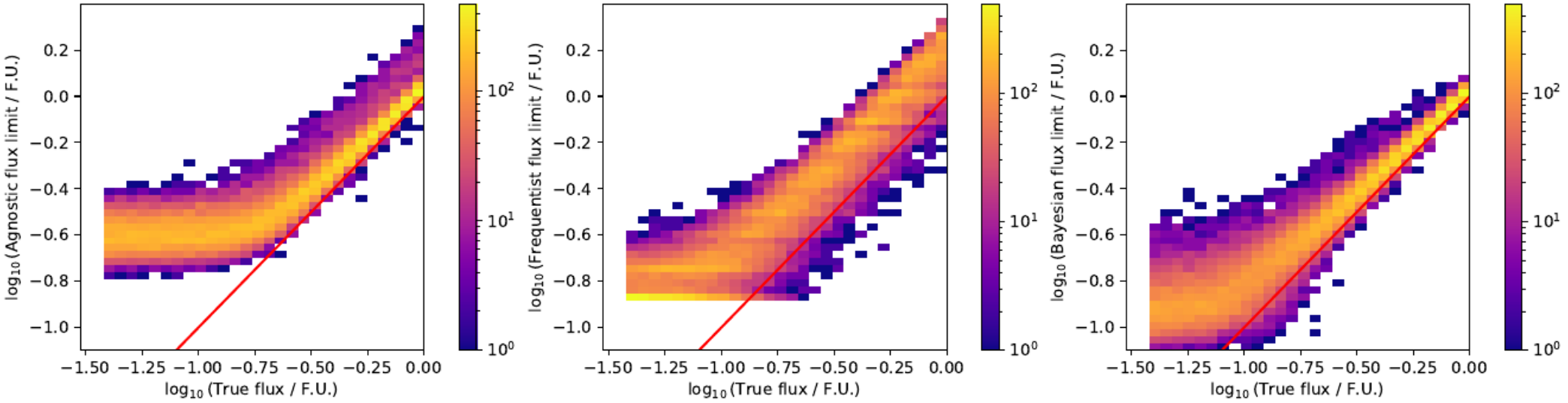
- The PSF is not anymore a delta function:
  - We fold over the skymap bins
  - To make calculations more accurate we also need to average PSF over the skymap bin
- Gammaness cuts are coming from the IRFs
- For agnostic and frequentist approach we need to specify θ^2 cut (used 75% efficiency)
- All the calculations are done in python using gammapy reading of CTAO IRFs (to simplify application to the data)

Example calculations

- We can do the calculations independently using the statistics of each estimated energy bin (assuming the spectral shape and migrations from true energy, similar like in regular IACT analysis)
- Bayesian method $p_f(f)$ can be also converted into flux point with uncertainty

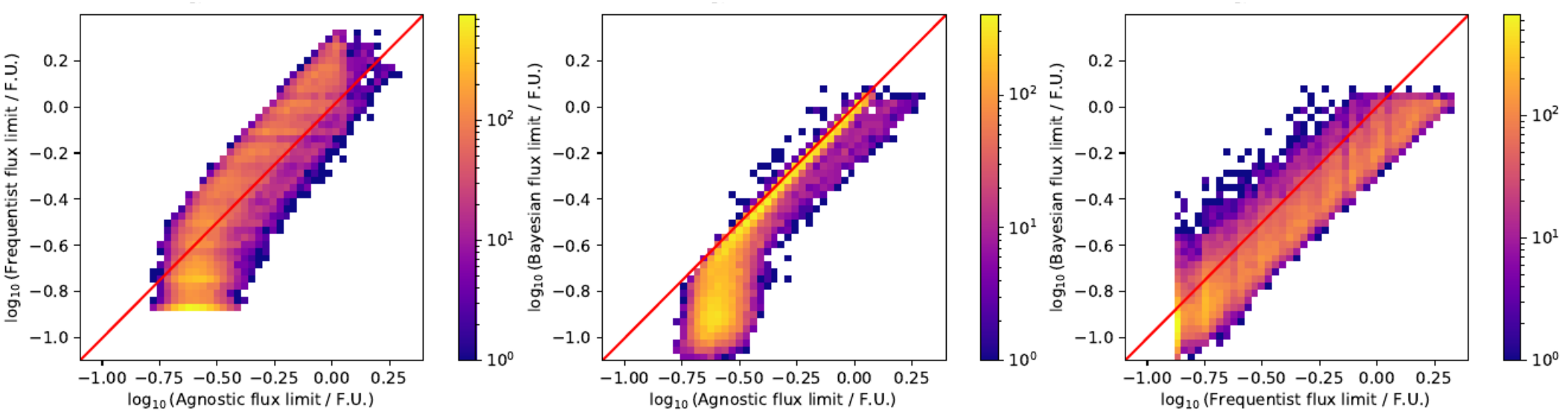


Limits of individual methods vs true flux



- Similar behavior like for toy MC case

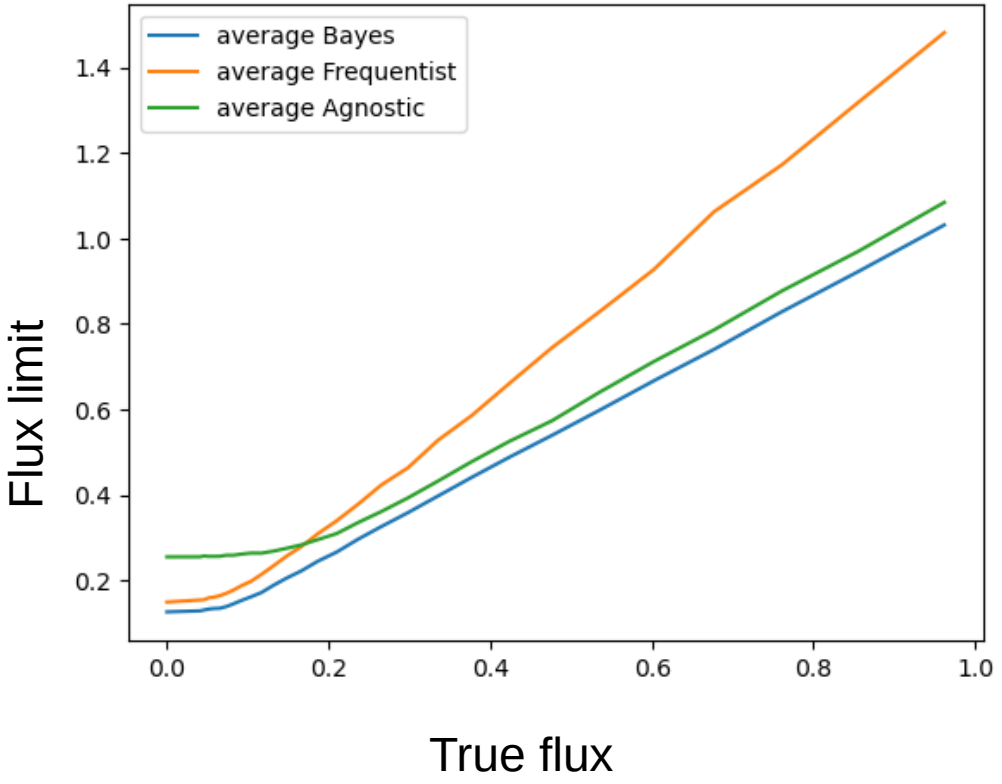
Comparison of limits of different methods



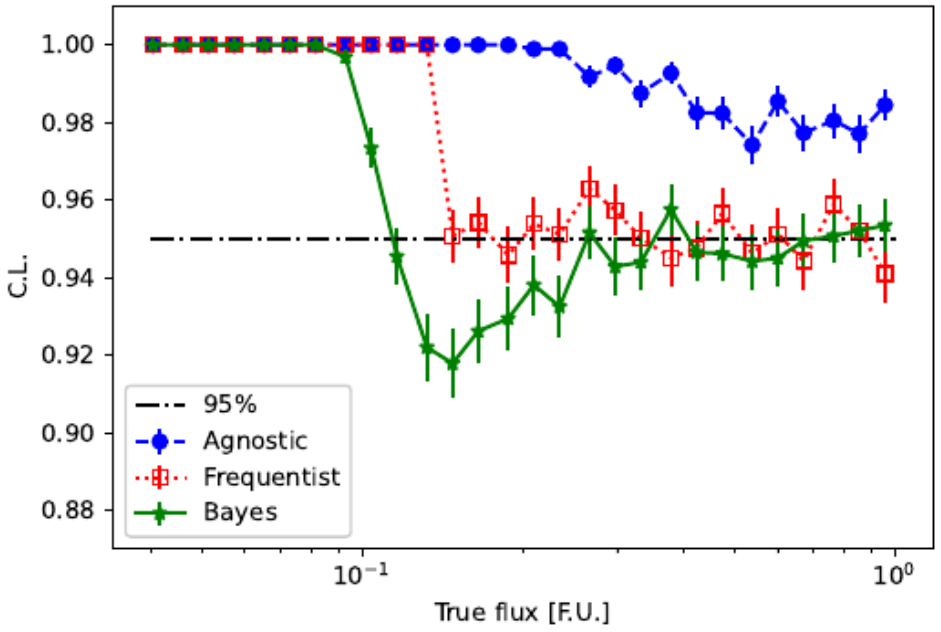
- For weak fluxes either frequentist of Bayesian perform well.
- Bayesian does not merge with agnostic at high fluxes anymore likely because of the using of PSF shape
- For strong fluxes frequentist methods overestimates limits (but it does not matter much since then you either way have a detection)

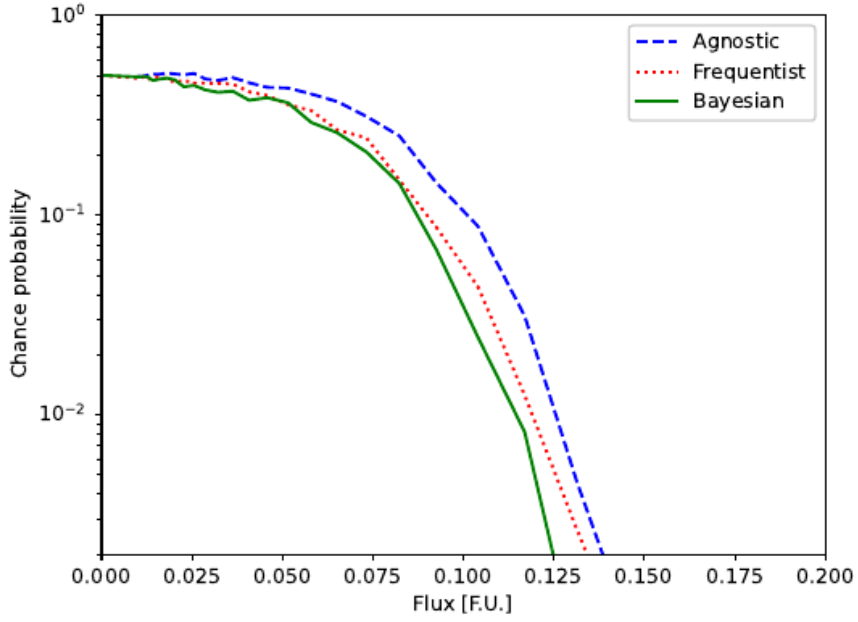
Best limit

- At low fluxes Bayesian and frequentist provide similar limits
- In this implementation Bayesian contrary to frequentist and agnostic includes also PSF information giving extra power to the method



Performance of methods (at 63 GeV)





- The behavior from toy MCs is reproduced

Comparison of the implemented methods

	Bayesian	**Frequentist**
Assumptions/tweeks	Need some (even if dummy) priors on the flux	Needs a special treatment to avoid negative upper limits
Used information	Full likelihood (using PSF)	Basic implementation uses aperture photometry (no PSF shape used)
Achieved limits	Best	For low fluxes nearly as good as Bayesian, for high fluxes worse (even than agnostic!)
Achieved C.L.	Mostly ok, but sometimes might be slightly lower.	By definition perfect
Technical issues	Calculations are relatively fast (~minutes), but used implementation needs considerable amount of memory	Based on MC simulations – quite CPU costly

Implementation

- All the code is in github: https://github.com/jsitarek/uncertain_position_limits

  (but since there are a few pecularities if you want to use it it is best that you contact me first)
- toy MC is ROOT based, full simulations is written in python with GADF IRFs

Take home message

- Statistical analysis allows us to put a probability on how likely we are wrong when claiming something
- There is a multitude of statistical methods. If they provide different answers, likely it is because they respond to a different question.