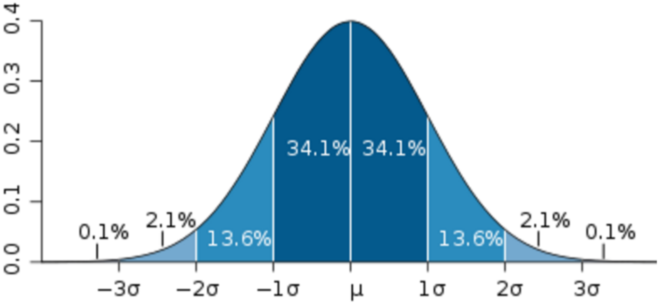
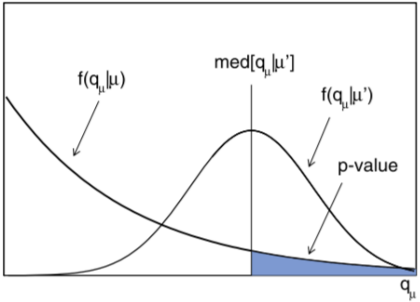




PH2282 part 6: Confidence Intervals

Applied Multi-Messenger Astronomy 2:
Statistical and Machine Learning Methods in Particle and Astrophysics

Hans Niederhausen and Matteo Agostini
TUM - summer term 2019

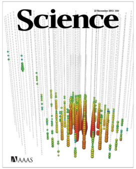
Topics of this block of lectures

About my lectures (upcoming three Fridays):

- Introduction to IceCube (and relevant physics)
- Statistical models: describing the detection process
- Monte Carlo Generation: understanding importance weights
- **Example application**: discovering diffuse astrophysical neutrinos
- Asymptotic properties of maximum likelihood methods
- **for today:** Interval estimation and confidence regions
- **Example application**: Searching for a point source of neutrinos in the sky (bonus topic, to be added at a later time)

Outline of today's lecture

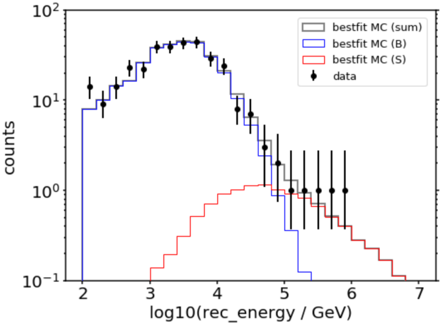
- summary of last lecture
- concept of **power** of a hypothesis test
- concept of **confidence sets**
- concept of **coverage**
- example: CI on normal mean using a pivot
- general Construction: **Inversion of Likelihood Ratio Tests**
- example: our toy-problem (gaussian signal on uniform bg)
- example and exercise: IceCube diffuse flux measurement

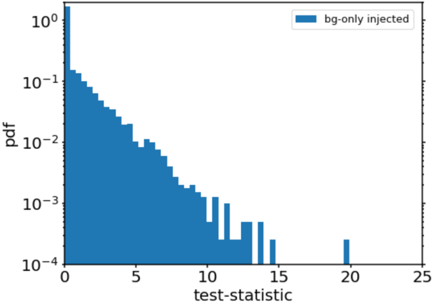


Summary of Previous Lectures: Diffuse Neutrino Flux

ingredients for the IceCube discovery analysis:

- maximum likelihood fitting, hypothesis testing using likelihood ratio ($H_0 : \Phi_{astro} = 0$ and $H_1 : \Phi_{astro} > 0$, with $\lambda = -2 \log L0/L1$ as TS)
- weighted Monte Carlo simulation to predict expected number of counts in each bin (for some assumption about the signal and background flux)





Summary of Previous Lectures: Large Sample Theory and Point Estimation

The distribution of the MLE converges to a normal distribution (with variance given by the CRB bound). The MLE is ...

- **a consistent estimator**. bias and variance converge to 0.
- **an asymptotically efficient estimator**. smallest possible variance as n grows large.
- **asymptotically normal**.

These are the reasons why maximum likelihood is so popular.

Summary of Previous Lectures: Large Sample Theory and Likelihood Ratio Testing

Reminder

given two hypotheses $H0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ and $H1 : \boldsymbol{\theta} \neq \boldsymbol{\theta_0}$
the likelihood ratio test-statistic $\lambda\left(\mathbf{x}\right)$ is defined as

$$\lambda\left(\mathbf{x}\right) = -2\log\Lambda\left(\mathbf{x}\right) = -2\log\left\{\frac{sup_{\nu}\ L\left(\boldsymbol{\theta_0},\ \boldsymbol{\nu}\,|\,\mathbf{x}\right)}{sup_{\nu,\,\theta}\ L\left(\boldsymbol{\theta},\ \boldsymbol{\nu}\,|\,\mathbf{x}\right)}\right\} \tag{1}$$

Wilk's Theorem

As the sample size increases, the distribution of the likelihood ratio test-statistic (**??**) converges to a χ^2 distribution with number of degrees of freedom k equal to the difference in number of free parameters specified by each hypothesis. In our notation $k = dim\,\boldsymbol{\theta}$.

$$f_{\lambda}\left(\lambda;\ \boldsymbol{\theta_0}\right) \underset{n\rightarrow\infty}{\longrightarrow} \chi^2\left(k\right) \tag{2}$$

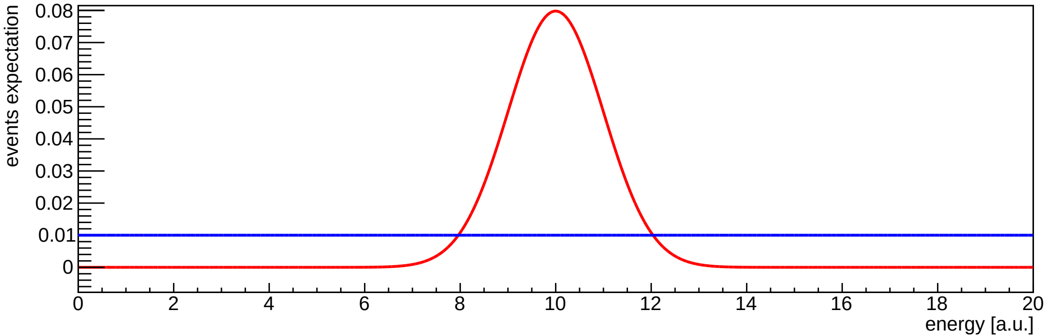
Wilk's Theorem (cont'd)
Unfortunately there are strict regularity conditions. Here are the two most important ones

- $\boldsymbol{\theta_0}$ needs to be an interior point of $\boldsymbol{\Theta}$
- nuisance parameters $\boldsymbol{\nu}$ that are only present under H1 are another issue
- ... several minor ones (typically not important)

Some extensions exists that might be useful (see Chernoff 1954, Gross, Vitells 2010) in such situations.

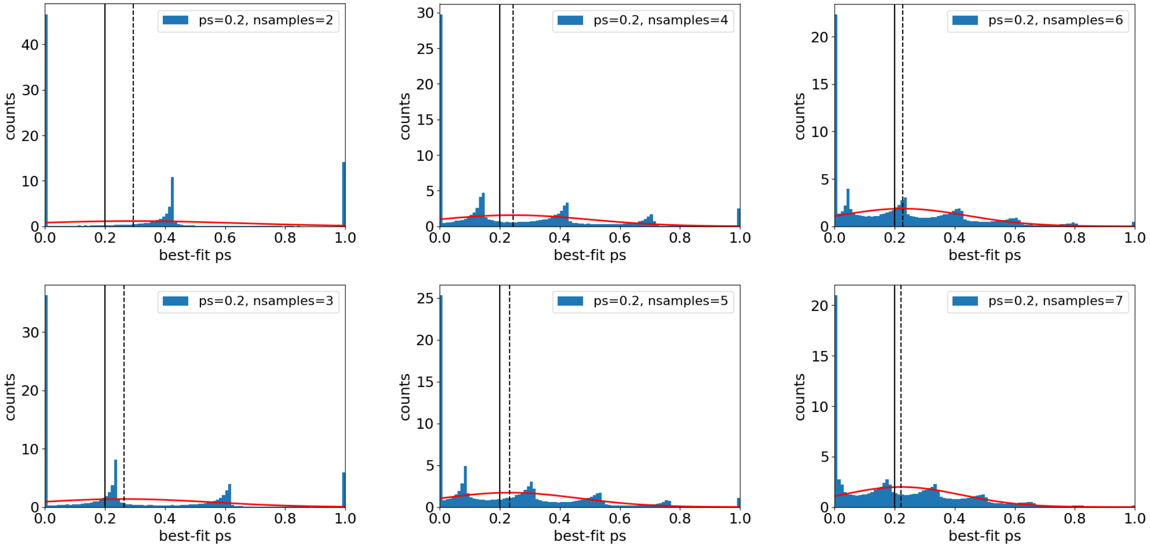
Summary of Previous Lectures: Large Sample Theory: The Toy Problem

Example: Our standard toy problem (with fixed sample size)

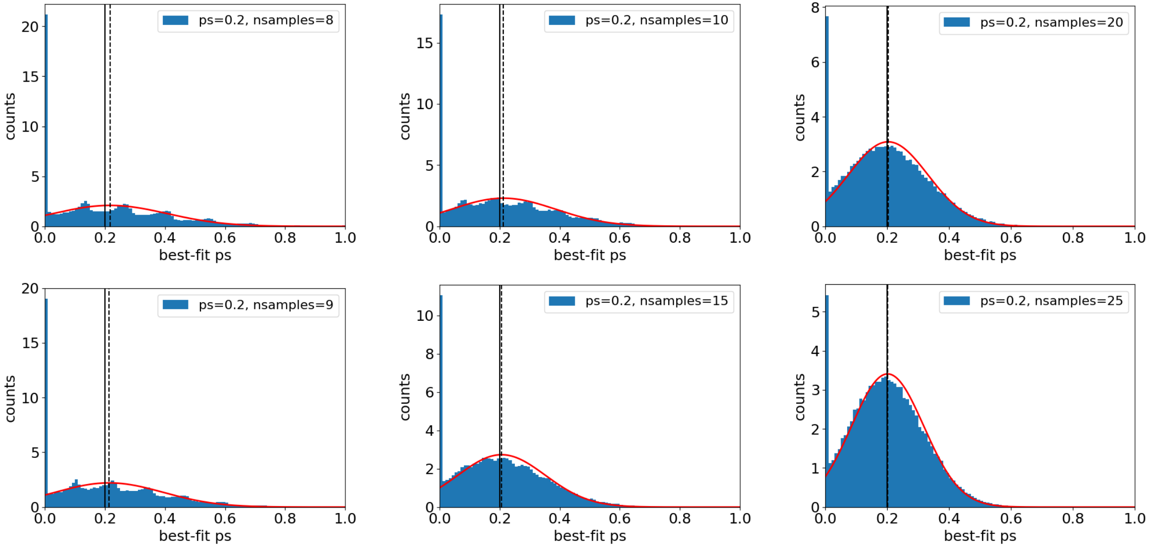


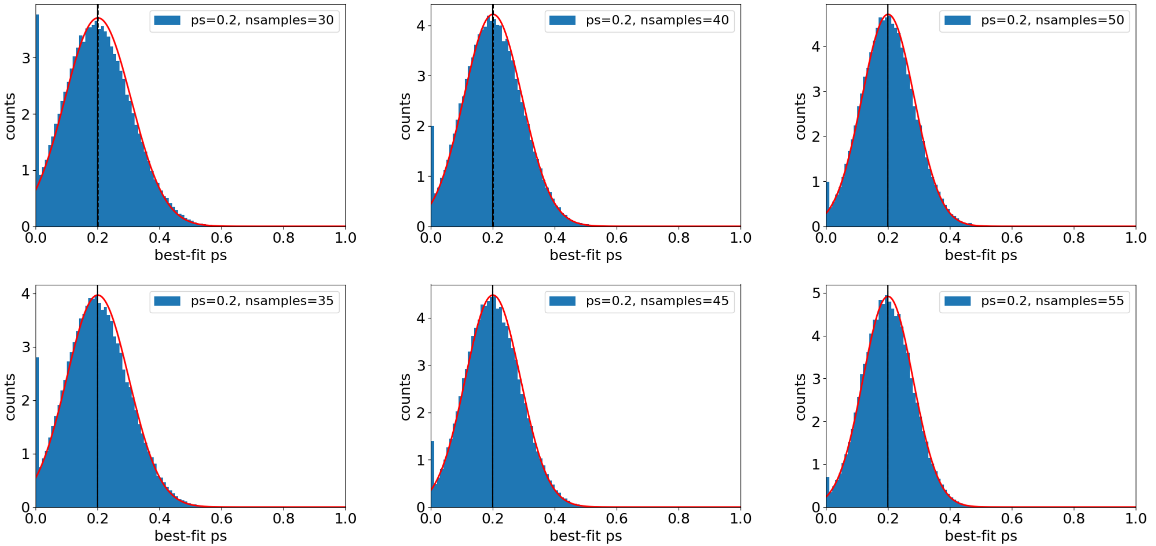
$$f_X(x;\mu,\sigma,p_s) = \left[p_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + (1-p_s) \cdot \frac{1}{20} \right] \tag{3}$$

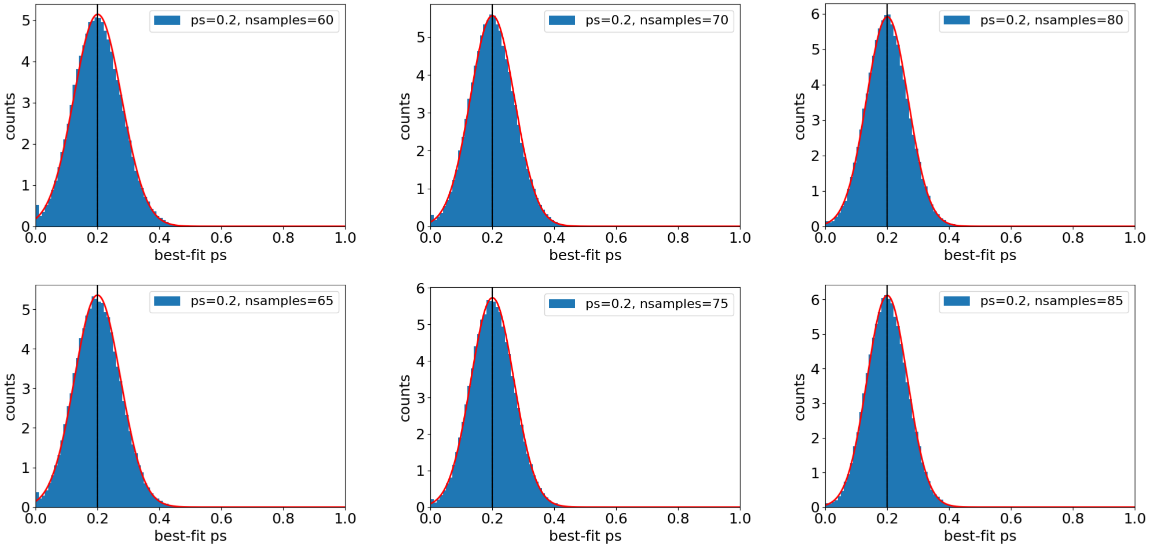
In the following treat p_s as the only unknown in the problem - and thus as a parameter.



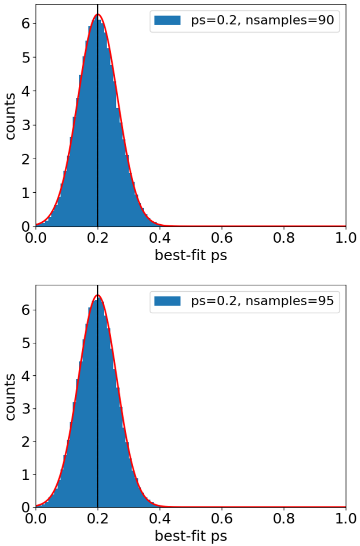
ps=0.2, nsamples=2
ps=0.2, nsamples=4
ps=0.2, nsamples=6
ps=0.2, nsamples=3
ps=0.2, nsamples=5
ps=0.2, nsamples=7
counts
best-fit ps


ps=0.2, nsamples=8
ps=0.2, nsamples=10
ps=0.2, nsamples=20
ps=0.2, nsamples=9
ps=0.2, nsamples=15
ps=0.2, nsamples=25
counts
best-fit ps

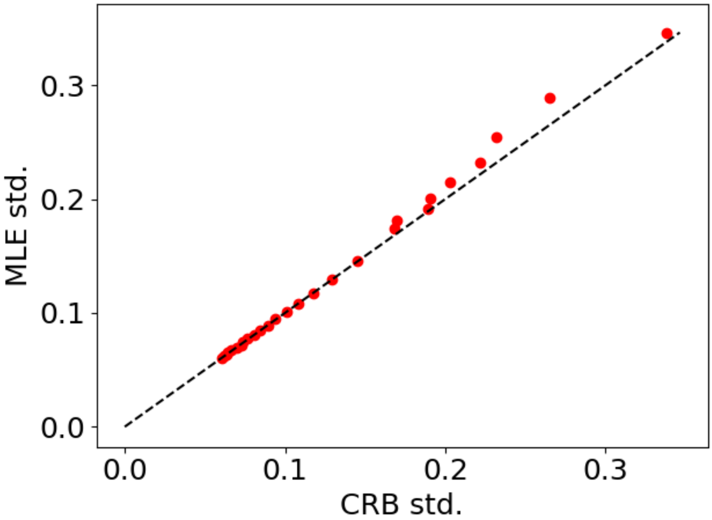

ps=0.2, nsamples=30
ps=0.2, nsamples=40
ps=0.2, nsamples=50
ps=0.2, nsamples=35
ps=0.2, nsamples=45
ps=0.2, nsamples=55
counts
best-fit ps



ps=0.2, nsamples=60
ps=0.2, nsamples=70
ps=0.2, nsamples=80
ps=0.2, nsamples=65
ps=0.2, nsamples=75
ps=0.2, nsamples=85
counts
best-fit ps



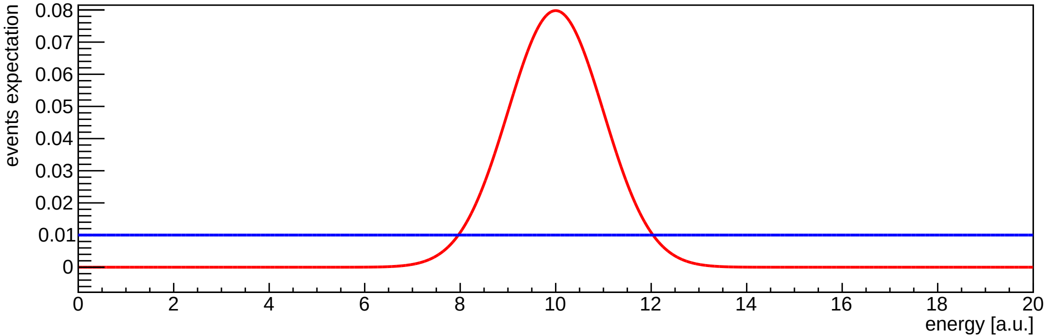
ps=0.2, nsamples=90
counts
best-fit ps
ps=0.2, nsamples=95
counts
best-fit ps



MLE std.
CRB std.

Large Sample Theory: The Toy Problem

Application to our standard toy problem (with 2 parameters: p_s, μ_s)

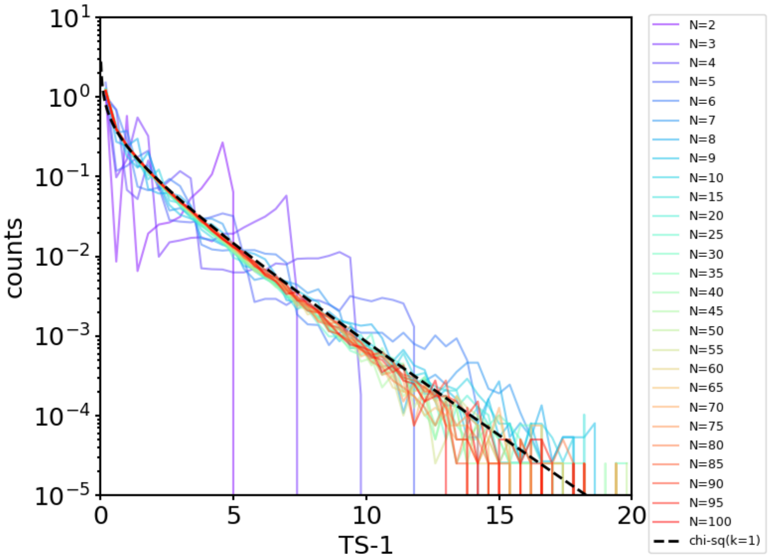


Two different hypothesis tests satisfying Wilk's theorem

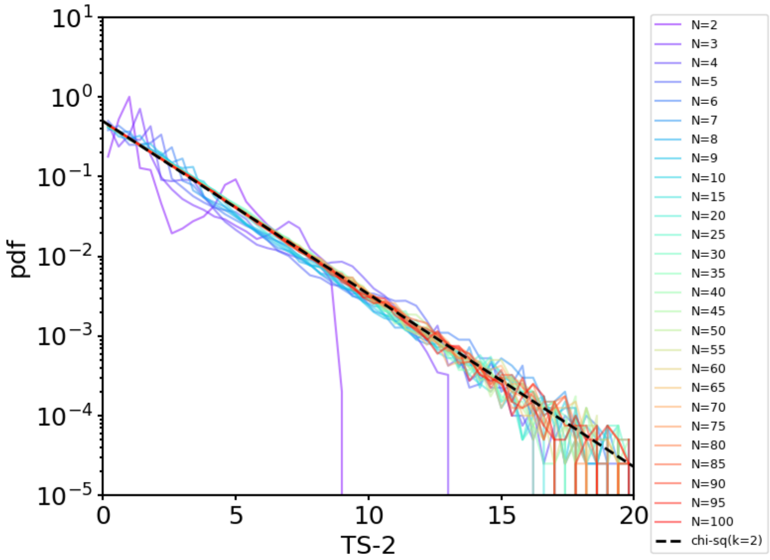
Case 1: $H0 : p_s = 0.2$ and $H1 : p_s \neq 0.2$ (k=1)

Case 2: $H0 : p_s = 0.2,\ \mu_s = 10.0$ and $H1 : p_s \neq 0.2,\ \mu_s \neq 10.0$ (k=2)

Large Sample Theory: The Toy Problem

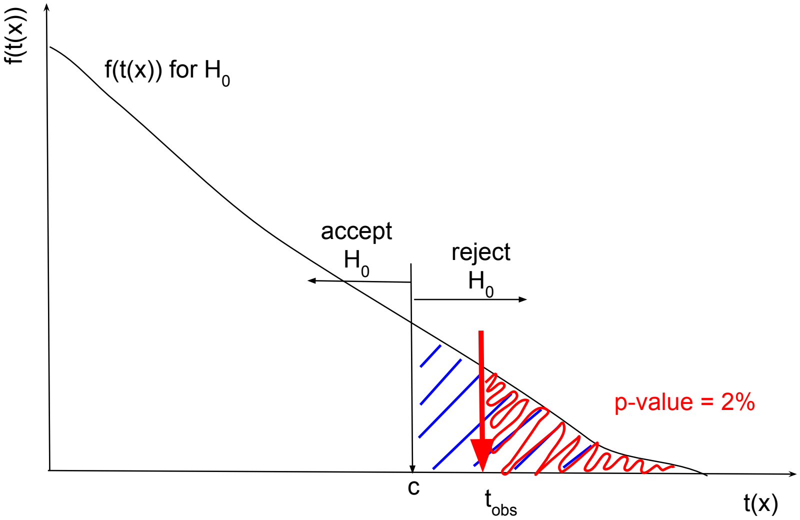


Large Sample Theory: The Toy Problem



Questions about previous lectures?

Hypothesis Tests: Statistical Power

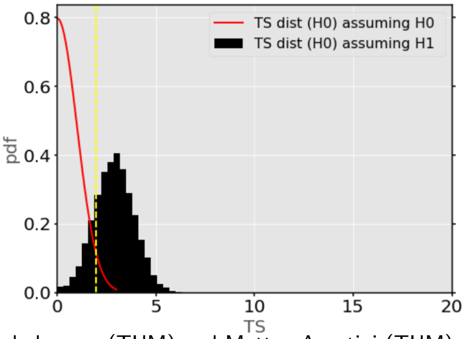


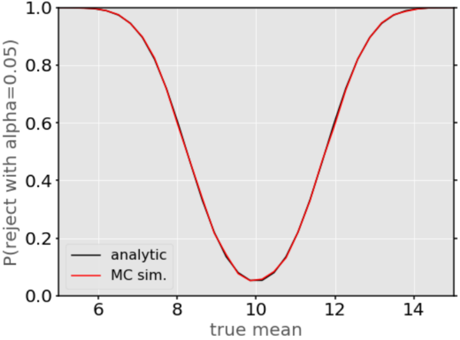
Hypothesis Tests: Statistical Power

The *power function* $\beta(\theta)$ of a hypothesis test with rejection region R is the probability of the test rejecting H0:$\theta = \theta_0$ as a function of the parameters θ in the model

$$\beta(\theta) = P_\theta(TS(\mathbf{X}) \in R) \tag{4}$$

example: gaussian mean $H0 : \mu = \mu_0$ and $H1 : \mu \neq \mu_0$ (σ known)
(check ipython notebook)





Hypothesis Tests: Statistical Power

- the **ideal** power function would be equal to 1 through the parameter space of the alternative hypothesis and 0 throughout the null-space - **good metric to judge tests**
- if you have to choose between two tests with same type I error probability, take the one that has larger power in the parameter space of the alternative hypothesis.
- **discovery potential**: value of $\theta' \neq \theta_0$ with $\beta(\theta') = 0.5$ (need to define rejection region, e.g. the 5σ criterion)

(beware of tests with small power: rejection of H_0 would not make H_1 more plausible.)

Confidence Intervals

Goal: calculate some range/region that has some probability to contain the true (unknown) parameter/s.

- Probability does not refer to the parameter (the true parameter is a fixed constant, not a random variable.) but to the region/interval that we obtain from the data.
- Generally speaking: different data results in a different region/interval (albeit construction is the same).

Mathematically, from data $\mathbf{X}$ we calculate function values $L(\mathbf{X})$ and $U(\mathbf{X})$ which are random variables.

$$[L(\mathbf{X}),\ U(\mathbf{X})] \quad (two - sided) \tag{5}$$

$$(-\infty,\ U(\mathbf{X})] \quad or \quad [L(\mathbf{X}),\ \infty) \quad (one - sided) \tag{6}$$

in physics: two-sided intervals often called "uncertainties", one-sided intervals often called "limits". (sometimes gets mixed ... e.g. hard to tell difference on bounded parameter spaces. always check how the construction was done.)

coverage := probability that the random interval $[L(\mathbf{X}),\, U(\mathbf{X})]$ (or limit) happens to overlap with the unknown, true parameter value.

$$P_\theta\left(\theta \in [L(\mathbf{X}),\, U(\mathbf{X})]\right) \tag{7}$$

confidence coefficient of an interval (denoted by $1-\alpha$) defined by

$$inf_\theta P_\theta\left(\theta \in [L(\mathbf{X}),\, U(\mathbf{X})]\right) = 1-\alpha \tag{8}$$

Can not always guarantee exact coverage (hello nuisance parameters!) - strive to guarantee confidence coefficient (i.e. minimum coverage!). That is usually possible.

Confidence Intervals: Coverage in the normal mean problem

Consider the problem of constructing a confidence interval for the unknown mean μ of a normal distribution (variance σ^2 known) from n observations ($\mathbf{X} = \{X_1, ..., X_n\}$). This can be done using a **pivot** (a function of the parameter and observations, that has a distribution which is independent of the parameter).

$$Q(\mu, \mathbf{X}) = \frac{\bar{X} - \mu}{\sigma / / \sqrt{n}} \tag{9}$$

$$Q \sim N(0, 1) \tag{10}$$

i.e. here Q is a standard normal random variable. Thus can solve

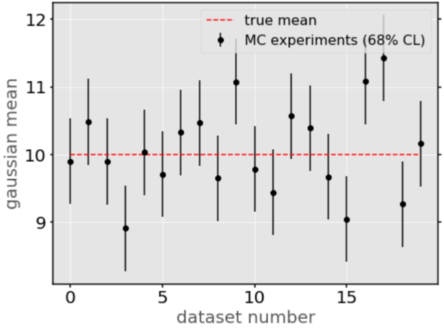
$$P_{\mu}(-a \leq Q \leq a) = 1 - \alpha \tag{11}$$

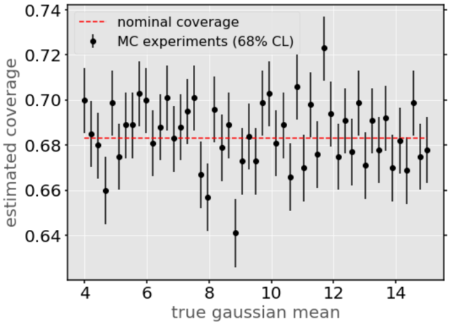
which corresponds to the following confidence set

$$\left\{\mu : \bar{X} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}\right\} \tag{12}$$

Confidence Intervals: Coverage in the normal mean problem

Check with Monte-Carlo (see ipython notebook)





Confidence Intervals from inversion of hypothesis tests

If you can construct a level α hypothesis test for the unknown parameter/s specified by H_0 it is always possible to use this test to construct a confidence interval with guaranteed confidence coefficient $1 - \alpha$ (see Theorem 9.2.2 in Casella and Berger). This is called **inverting a hypothesis test**. Whether you get two-sided or one-sided intervals depends on the alternative hypothesis

- $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$ produces two-sided intervals
- $H_0 : \theta = \theta_0$ and $H_1 : \theta < \theta_0$ produces one-sided intervals (upper-limit)
- $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$ produces one-sided intervals (lower-limit)

The more powerful the underlying hypothesis test, the better the interval (smaller range, more accurate).

(Note, the Feldman-Cousins construction is a special case of this. They invert a likelihood ratio test (two-sided) concerning the poisson mean).

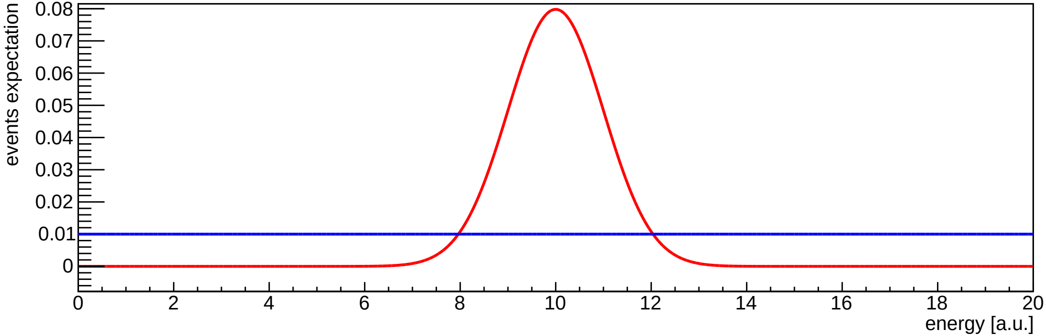
Confidence Intervals from inversion of hypothesis tests

Why does it work?

- perform the test on every possible point in parameter space
- if the test rejects the point, simply discard it
- if the point is accepted, add the point to your confidence set
- Whats the coverage of this strategy? (probability that the random set contains true parameter)
- Probability to rejected a parameter if it is true is $\leq \alpha$ by definition (size of test)
- Thus, probability for true parameter to contribute to set is $\geq 1 - \alpha$.
- Hence, probability for set to cover true parameter is $\geq 1 - \alpha$ by construction

Confidence Intervals from inversion of LRT

We have learned how to construct likelihood ratio tests. Let's invert a likelihood ratio test to obtain a confidence set on the signal fraction p_s in our toy model. (see ipython notebooks)

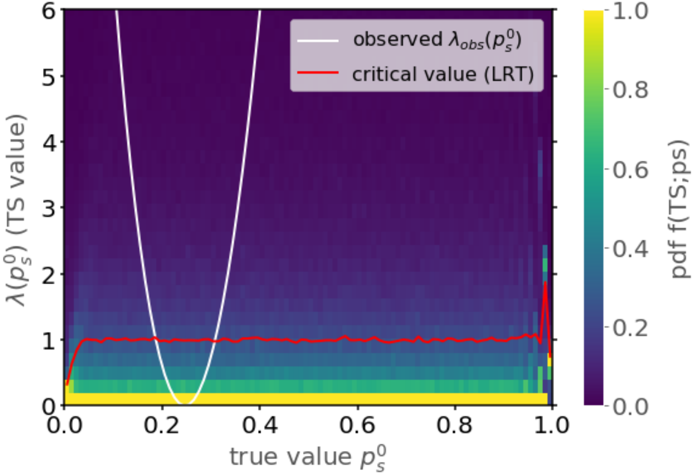


$$f_X(x; \mu, \sigma, p_s) = \left[p_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + (1 - p_s) \cdot \frac{1}{20} \right] \tag{13}$$

Confidence Intervals from inversion of LRT in toy problem.

$H_0 : p_s = p_s^0$ and $H_1 : p_s \neq p_s^0$, sample size n=100

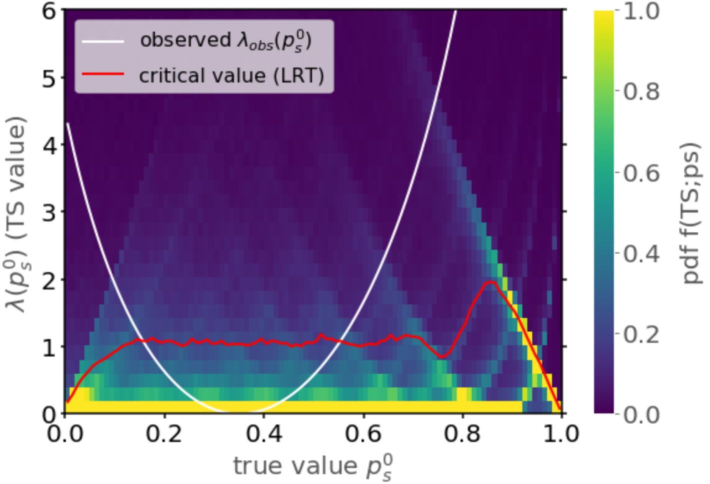
endpoints of interval: intersection points of obs. TS value (white) with critical value (red)



Confidence Intervals from inversion of LRT in toy problem.

$H_0 : p_s = p_s^0$ and $H_1 : p_s \neq p_s^0$, sample size n=10

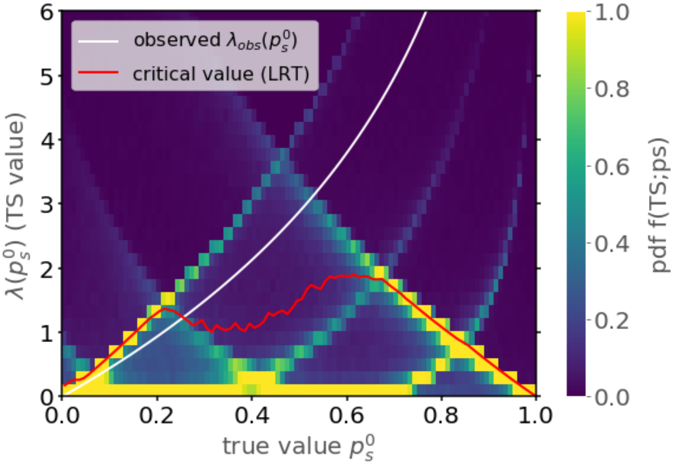
endpoints of interval: intersection points of obs. TS value (white) with critical value (red)



Confidence Intervals from inversion of LRT in toy problem.

$H_0 : p_s = p_s^0$ and $H_1 : p_s \neq p_s^0$, sample size n=3

endpoints of interval: intersection points of obs. TS value (white) with critical value (red)

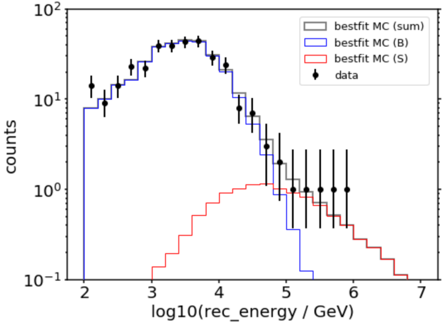


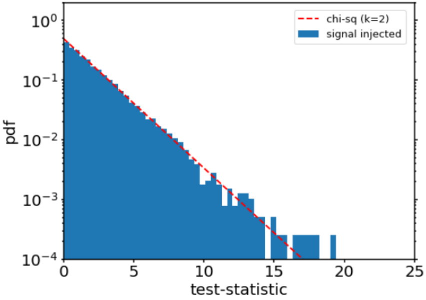
Confidence Intervals from inversion of LRT in toy problem.

- very simple if your measurement is in the asymptotic regime (lots of data!)
- obtain the critical value (red curve) from wilk's theorem (i.e. appropriate χ^2-pdf)
- generalizes well to high dimensions, if analysis remains asymptotic
- if asymptotics don't apply, you will run out of CPU quickly as the dimensionality increases (since you need to costruct the TS distributions for each point in parameter space)
- always check a few representative parameter combinations (and also a few extreme ones) first
- be mindful of the power in your tests!)

Inversion of LRT in the IceCube diffuse flux measurement

To construct a joint confidence interval for the normalization and spectral index of the astrophysical neutrino flux, we need to invert a LRT:
$H_0 : (\Phi, \gamma) = (\Phi_0, \gamma_0)$ and $H_1 : (\Phi, \gamma) \neq (\Phi_0, \gamma_0)$
The asymptotic expectation for the TS distribution would be χ^2 with 2 dof.

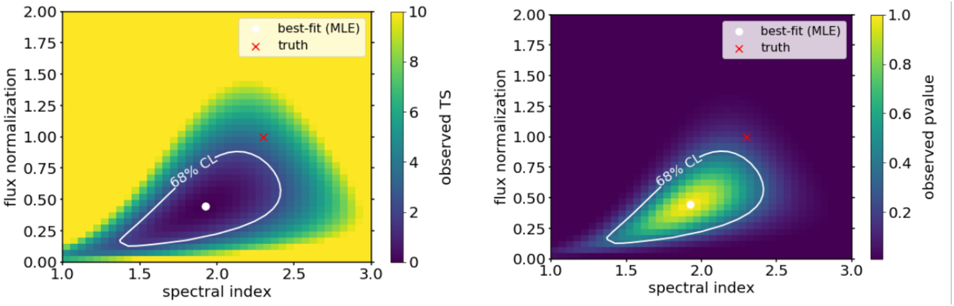




Inversion of LRT in the IceCube diffuse flux measurement

if we have sufficient data, we use the χ^2 pdf (left) otherwise we need to obtain (valid) p-values from MC simulations and use those to get the contours (right)

$$p(\mathbf{x_{obs}}) = \sup_{\theta \in \Theta_0} P_\theta \left(TS(\mathbf{X}) \geq TS(\mathbf{x_{obs}}) \right) \tag{14}$$



Exercise

- ...

Further Reading

- Casella and Berger
- Asymptotics by Cranmer et. al.
- Maybe FC paper