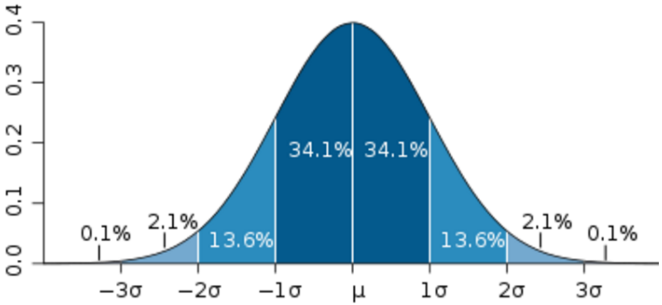
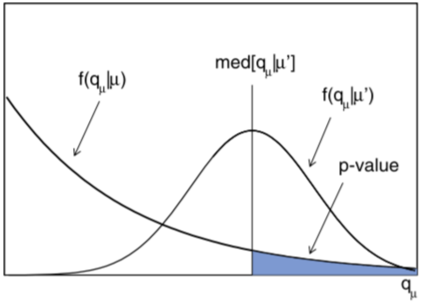




PH2282 part 5: Large Sample Theory ("Asymptotics")

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):

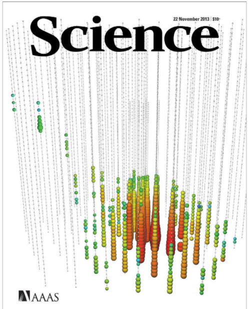
- Introduction to IceCube (and relevant physics)
- Statistical models: describing the detection process
- Monte Carlo Generation: understanding importance weights
- **Example application**: discovering diffuse astrophysical neutrinos
- Interval estimation and confidence regions
- Asymptotic properties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

Outline of today's lecture

- Summary of last lecture
- More theory of random variables (multiple random variables, transformations)
- Another look at Monte Carlo simulations
- Asymptotic properties maximum likelihood methods (MLE and LRT)

Summary of last lecture: IceCube experiment

- IceCube is a Cherenkov detector at the South Pole
- Goal: astronomy with neutrinos
- Neutrinos interact through DIS (CC and NC)
- Neutrino properties (direction + energy) inferred from light signals
- Observables used in analyses: reconstructed quantities
- Example: Discovery of the diffuse flux



Summary of last lecture: Monte Carlo simulations

Ingredients

full statistical model

model parameters (physics)

stochasticity inherent to physics model
(generation of neutrino energies)

true particle properties (latent random variables)
marginalized out of final model

stochasticity inherent to detection model
(e.g. particle interactions, electronics)

observables (random variables)
to go into likelihood

(could explain relationship to hierarchical models, marginalization on blackboard)

Summary of last lecture: Monte Carlo simulations

- want pdf $f(x)$ of observables x
- problem: typically no closed form expression for $f(x)$ available
- solution: it is easier to construct $f(x, \xi_1, ..., \xi_N)$ where ξ_i are latent random variables

$$f(x) = \int d\xi_1 ... \int d\xi_2 \ f(x, \xi_1, ..., \xi_N) = \int d\xi_1 ... \int d\xi_2 \ f(x \,|\, \xi_N, ...\xi_1) f(\xi_N | \xi_{N-1}, ..., \xi_1) .... f(\xi_2 | \xi_1) \quad (1)$$

We simply simulate the entire hierarchy:
for i in range(Ns)

1) draw ξ_1 from $f(\xi_1)$
2) draw ξ_2 from $f(\xi_2|\xi_1)$ (using the value ξ_1 from the prev. step)
3) ...
4) draw observable x from $f(x \,|\, \xi_N, ...\xi_1)$ (typically your gaussian resolution)

recording the values x_i ($i \in \{1...Ns\}$) implicitely solves the marginalization integral.

Summary of last lecture: Monte Carlo simulations

problems with multiple random variables

- bivariate case $f(x,y) = f(x|y)f(y) = f(y|x)f(x)$
- bivariate case cont'd $f(x) = \int dy\ f(x,y)$, $f(y) = \int dx\ f(x,y)$
- this generalizes to higher dimensions

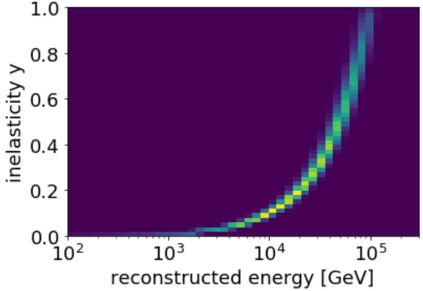
$$f(x_1, ..., x_N) = f(x_1, ..., x_k \mid x_{k+1}, \ ... x_N) f(x_{k+1}, \ .... x_N) \tag{2}$$

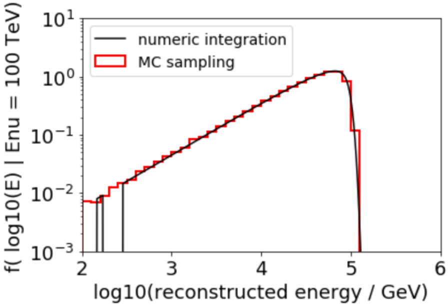
$$f(x_{k+1}, \ .... x_N) = \int dx_1 ... \int dx_k \ f(x_1, ..., x_N) \tag{3}$$

Summary of last lecture: Monte Carlo simulations

check ipython notebook on github (ex4) for a simple example
(understanding-marginalization-of-latent-variables-through-sampling)

IceCube: inelasticity (energy transfer to nucleon target) of the interaction is a latent random variable
inelasticity strongly impacts reconstructed neutrino energy ($E_\nu = 100\,TeV$ const.)





Summary of Previous Lectures: Importance Weights

- Monte-Carlo simulations take a long time
- detector simulation is the main challenge
- can simulate the full model (including the detector) for some arbitrary initial physics model (IceCube: particle fluxes!)
- importance weights allow to *reweight* the MC to describe different physics models (effectively reusing the detector simulation)

example: calculate an average over pdf $f(x)$ using samples from a different pdf $g(x)$

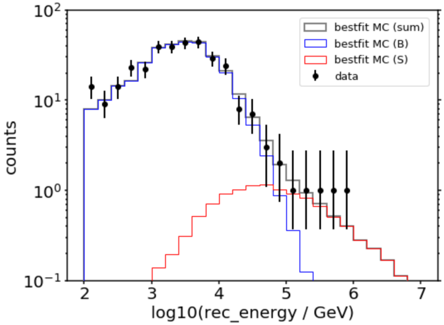
$$\mu = \int g(x) f(x) dx = \int \frac{g(x)f(x)}{h(x)} h(x) dx = E_h\left(\frac{g(x)f(x)}{h(x)}\right) \tag{4}$$

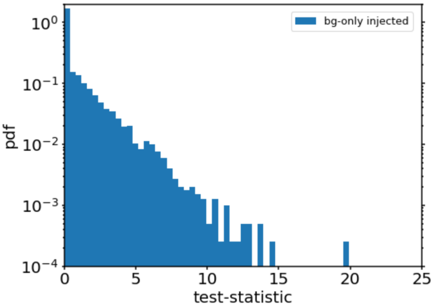
$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \frac{g(x_i)f(x_i)}{h(x_i)}, \qquad w_i = \frac{f(x_i)}{g(x_i)} \tag{5}$$

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)





Questions about last lecture?

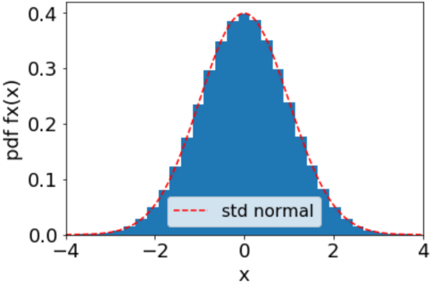
More Theory of Random Variables: Transformations

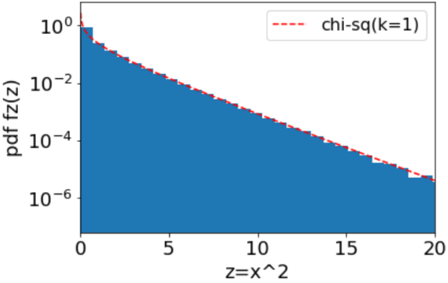
in many problems it can be useful to work with transformed random variables.
assume $x \sim f_X(x)$ - what is the distribution $f_Z(z)$ of $z = g(x)$ ($g(x)$ some function)?
If $g(x)$ is monotone, then

$$f_Z(z) = \begin{cases} f_X(g^{-1}(z))\left|\frac{d}{dz}g^{-1}(z)\right|, & z \in Z \\ 0 & \textit{otherwise} \end{cases} \tag{6}$$

in the multivariate case, the transformation factor is given by the determinant of the Jacobian matrix.

example: square of a std. normal rv $g(x) = x^2$, $X \sim N(0, 1)$
caution: square is not monotone.
solution: partition the sample space in regions where transformation is monotone (here: $x < 0$ and $x > 0$) apply law in each region separately. sum transformed pdf over the contributions from each partition.





More Theory of Random Variables: Transformations

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \tag{7}$$

$$g^{-1}(z) = \begin{cases} -\sqrt{z}, & x < 0 \\ \sqrt{z}, & x > 0 \end{cases} \tag{8}$$

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-(-\sqrt{z})^2/2\right) \frac{1}{2\sqrt{z}} + \frac{1}{\sqrt{2\pi}} \exp\left(-(\sqrt{z})^2/2\right) \frac{1}{2\sqrt{z}} \tag{9}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{z}} \exp\left(-z^2/2\right) \tag{10}$$

Questions about transformations of random variables?

Questions about transformations of random variables?
Next topic: large sample theory (asymptotics)

Large Sample Theory

- large samples ($n \longrightarrow \infty$) are typically easier to analyze than small samples ($n \longrightarrow 0$)!
- behavior/performance of statistical methods is (often) well defined in large samples
- maximum likelihood methods can be shown to have nice properties in large samples
- this lecture: develop some of the important concepts (and apply to our toy example)

Remember the following metrics to judge quality of a point estimator $W(X)$:
bias, variance, mean squared error

$$E_\theta W(X) - \theta \ \text{(bias)} \tag{11}$$

$$E_\theta \left[\{W(X) - E_\theta W(X)\}^2 \right] \ \text{(variance)} \tag{12}$$

$$E_\theta \left(W(X) - \theta \right)^2 = Var_\theta W + (Bias_\theta W)^2 \ \text{(mean squared error)} \tag{13}$$

Remember the following metrics to judge quality of a point estimator $W(X)$:
bias, variance, mean squared error

need to decide how to choose an estimator:

1) require no bias (unbiased estimator) and then choose the one which minimizes variance
2) choose the one which minimizes MSE (not necessarily the same criterion as above)

The MLE asymptotically satisfies 1) (i.e. as $n \longrightarrow \infty$) (*exclusions apply)

definition
A sequence of estimators $W_n = W_n (X_1, ..., X_n)$ is a *consistent* sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$ we have

$$lim_{n \to \infty} P_\theta (|W_n - \theta| < \epsilon) = 1 \tag{14}$$

i.e. for any small region around the true value, the probability to find the estimator inside converges to one!
The MLE is a consistent estimator!
(its bias and variance converge to 0)

How about the rate of convergence (*efficiency*)? Let's look at the variance. The smallest possible variance (i.e. the one that no estimator can beat) is well defined by the **Cramer-Rao Inequality**

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X})\right)^2}{E_\theta\left([\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)]^2\right)} \tag{15}$$

which in the iid situation simplifies to

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X})\right)^2}{nE_\theta\left([\frac{\partial}{\partial\theta}\log f(x|\theta)]^2\right)} \tag{16}$$

The MLE is an asymptotically efficient estimator!.
Its variance attains the Cramer-Rao lower bound (CRB).

How about the rate of convergence (*efficiency*)? Let's look at the variance. The smallest possible variance (i.e. the one that no estimator can beat) is well defined by the **Cramer-Rao Inequality**

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathbf{X})\right)^2}{E_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right]^2\right)} \tag{17}$$

which in the iid situation simplifies to

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathbf{X})\right)^2}{nE_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2\right)} \tag{18}$$

The MLE is an asymptotically efficient estimator!.
Its variance attains the Cramer-Rao lower bound (CRB).

Large Sample Theory: Point Estimation

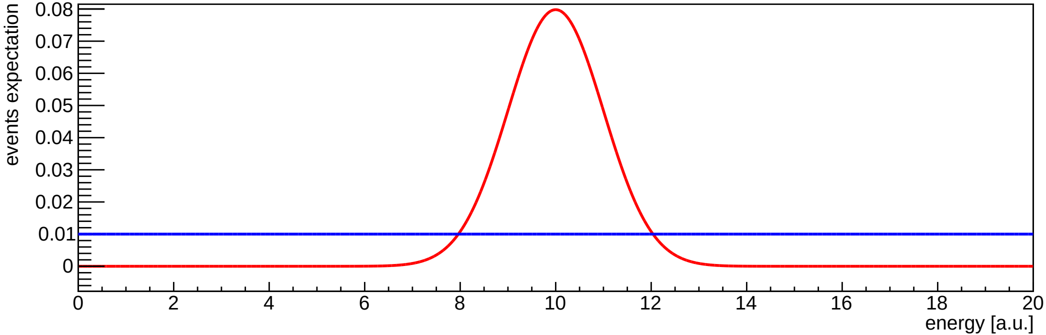
The distribution of the MLE converges to a normal distribution (with variance given by the CRB bound)
In summary: The MLE is ...

- **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.

Large Sample Theory: The Toy Problem

Example: Our standard toy problem



We will use one simplification:
We keep the sample size fixed! (no poisson fluctuations)
This corresponds to running the experiment until N counts have been observed (not for some fixed amount of time).

Large Sample Theory: The Toy Problem

For fixed sample size N, we can use the *signal fraction* $p_s = \lambda_s/N$ as parameter and eliminate λ_b through $\lambda_s + \lambda_b = N$

$$f_X(x;\mu,\sigma,\lambda_s,\lambda_b) = \frac{1}{\lambda_s+\lambda_b}\left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20}\right] \tag{19}$$

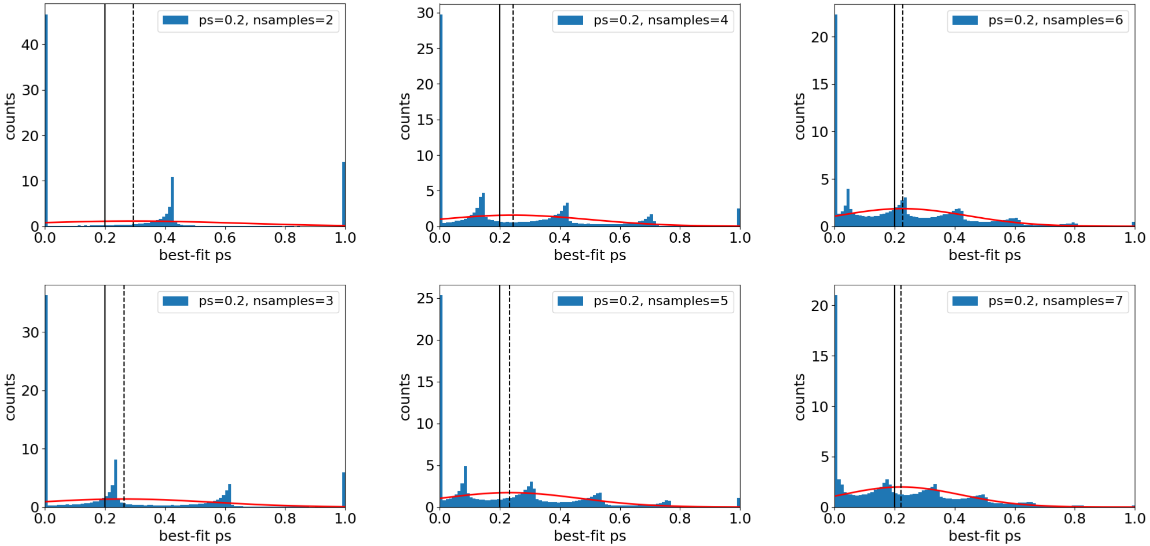
becomes

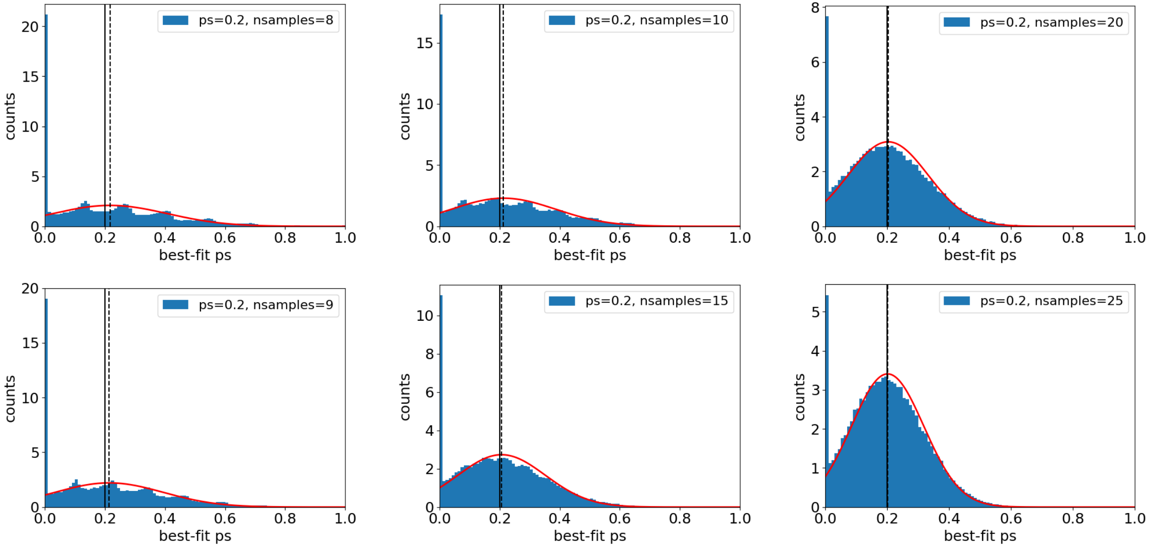
$$f_X(x;\mu,\sigma,p_s) = \left[p_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + (1-p_s) \cdot \frac{1}{20}\right] \tag{20}$$

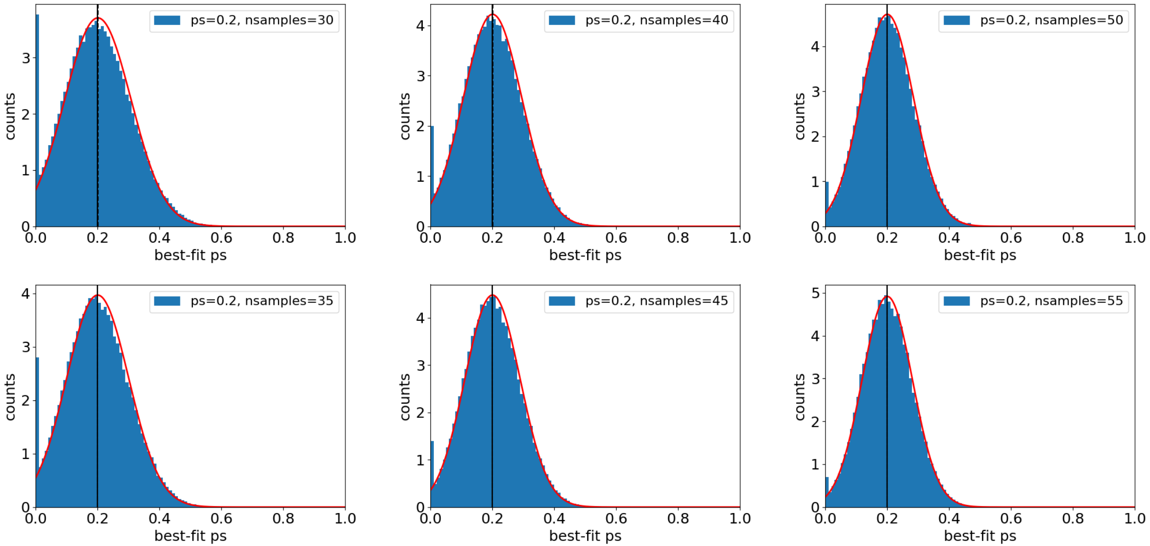
In the following treat p_s as the only unknown in the problem - and thus as a parameter.

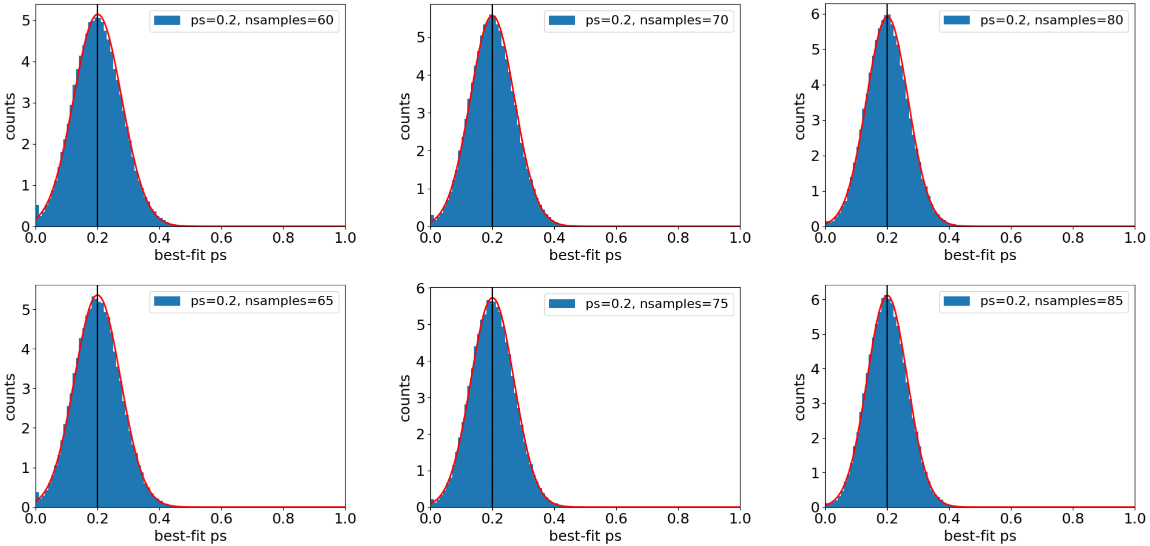
Large Sample Theory: The Toy Problem

Here we compare the distribution of the MLE $\hat{p}_s$ of $p_s = 0.2$ for different sample sizes $N \in \{2, 3...10, 15, 20, ...100\}$.
We also calculate numerically (using both, numeric integration, and sampling) the corresponding CRB
The CRB is compared to the observed variance $Var\hat{p}_s\ (p_s = 0.2)$.
Similiar to exercise 2, we need to generate many pseudo-datasets (for fix N) for this.
(discuss how to evaluate/estimate the CRB using eq. (16))
The MLE $\hat{p}_s$ clearly shows the expected convergence!

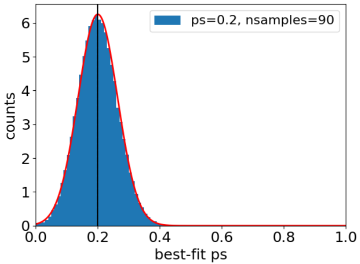
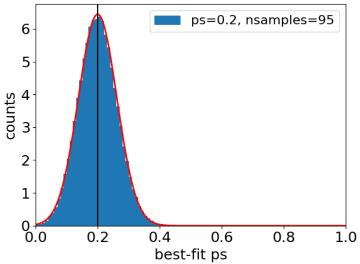
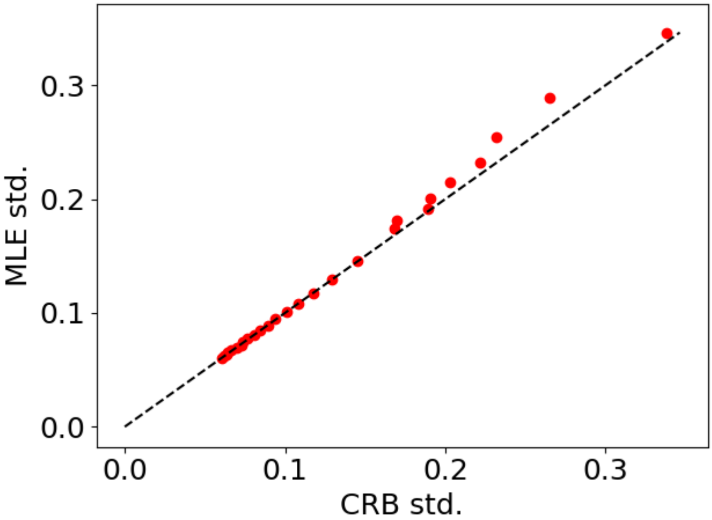

ps=0.2, nsamples=2
ps=0.2, nsamples=4
ps=0.2, nsamples=6
ps=0.2, nsamples=3
ps=0.2, nsamples=5
ps=0.2, nsamples=7
counts
best-fit ps


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.

Questions?

Large Sample Theory: The Toy Problem

Questions?

Behavior of Likelihood Ratio Tests in large samples.

Large Sample Theory: 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 $\lambda(\mathbf{x})$ is defined as

$$\lambda(\mathbf{x}) = -2 \log \Lambda(\mathbf{x}) = -2 \log \left\{ \frac{sup_{\nu} \; L(\boldsymbol{\theta_0}, \, \boldsymbol{\nu} \,|\, \mathbf{x})}{sup_{\nu, \, \theta} \; L(\boldsymbol{\theta}, \, \boldsymbol{\nu} \,|\, \mathbf{x})} \right\} \tag{21}$$

to perform the hypothesis test, we also need to know the sampling distribution of this test-statistic:

$$\lambda \sim f_{\lambda}(\lambda; \, \boldsymbol{\theta}, \, \boldsymbol{\nu}) \tag{22}$$

Often, this is non-trivial and one needs extensive Monte-Carlo computations (see example 3)
Luckily, as the sample size increases, the distribution is known to **converge**!
(beware of conditions!)

Wilk's Theorem
As the sample size increases, the distribution of the likelihood ratio test-statistic (22) 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 $k = dim\,\boldsymbol{\theta}$.

$$f_\lambda\left(\lambda;\, \boldsymbol{\theta_0}\right) \underset{n\to\infty}{\longrightarrow} \chi^2\left(k\right) \tag{23}$$

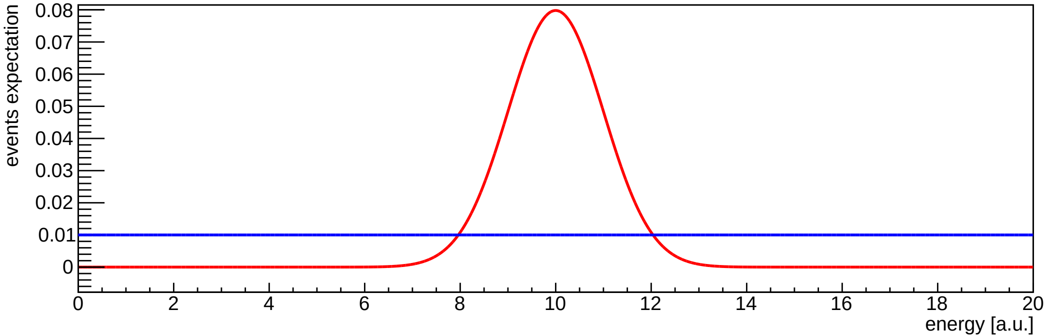
Wilk's Theorem (cont'd)
Unfortunately there are strict regularity conditions. Here are the two most importance ones

- $\boldsymbol{\theta_0}$ 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.

Large Sample Theory: The Toy Problem

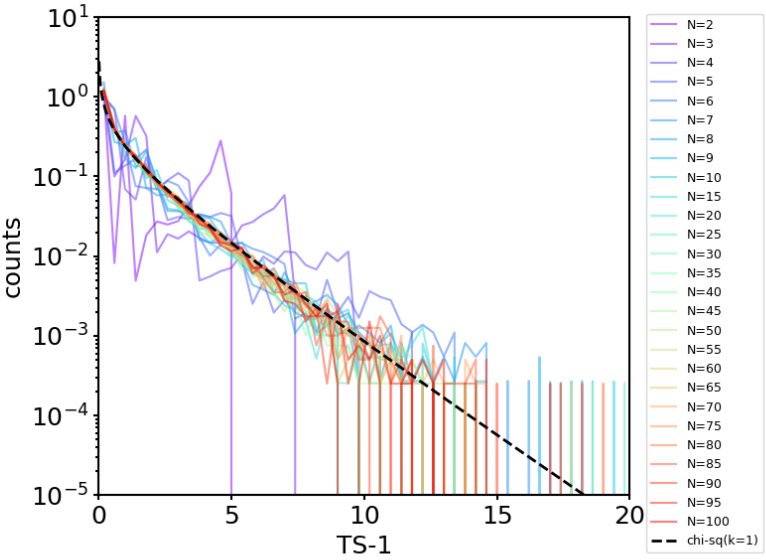
Application to our standard toy problem (with 2 parameters: p_s, μ_s)



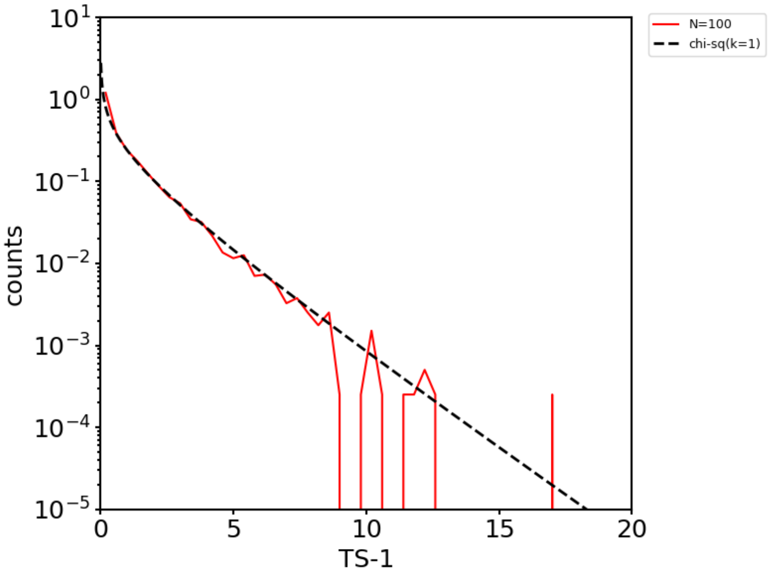
Two different hypothesis tests satisfying Wilk's theorem

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

Large Sample Theory: The Toy Problem

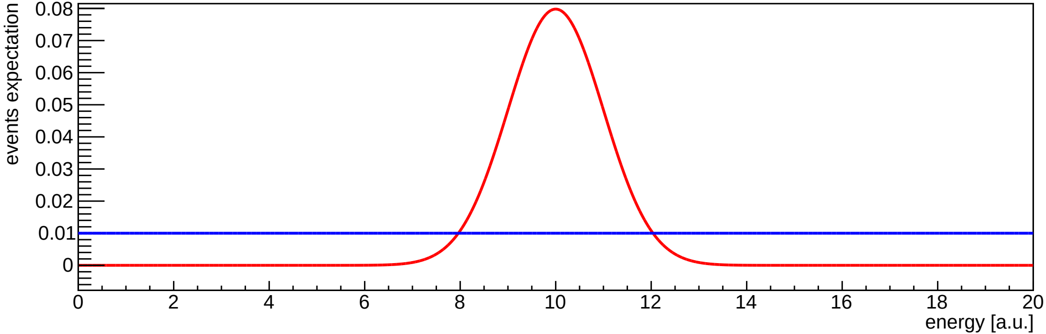


Large Sample Theory: The Toy Problem



Large Sample Theory: The Toy Problem

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

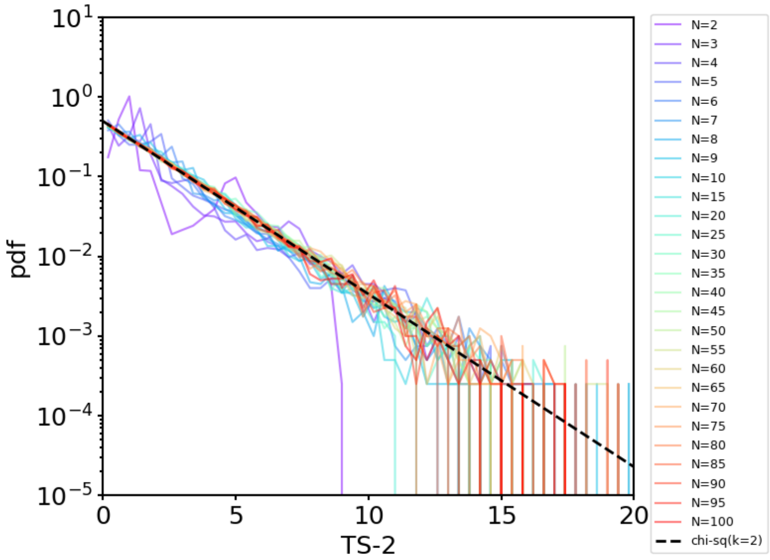


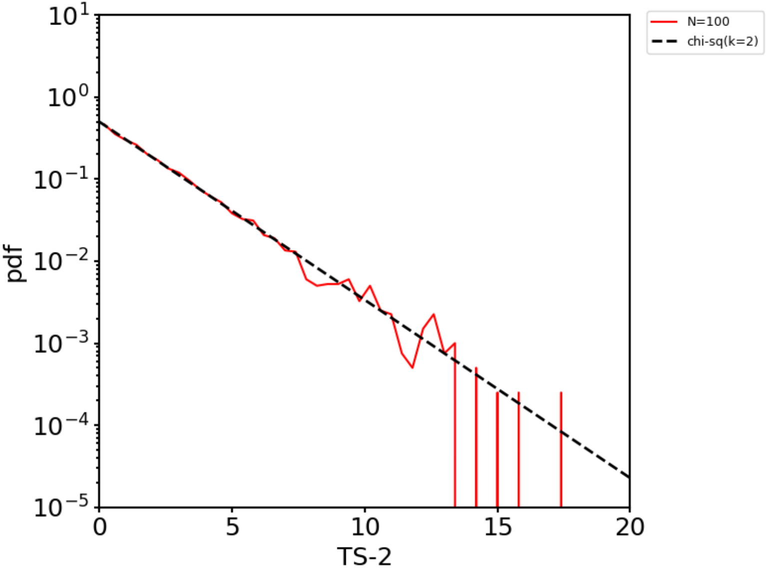
Two different hypothesis tests satisfying Wilk's theorem

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




N=100
chi-sq(k=2)
pdf
TS-2
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
0
5
10
15
20

Exercises

- Exercise 1: Study MLE as function of sample size (e.g. compute CRB bound)
- Exercise 2: Study LRT test-statistic distribution as function of sample size and observe Wilk's behavior
- Exercise 3: redo exercise 2 for the case of the IceCube disovery