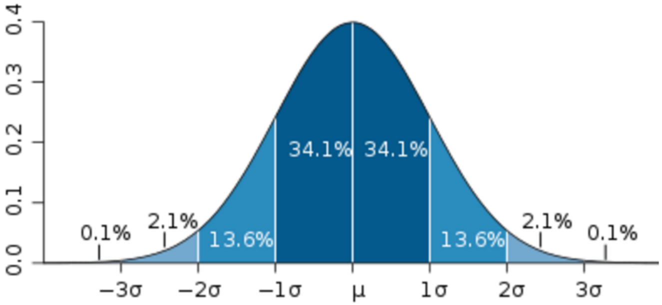
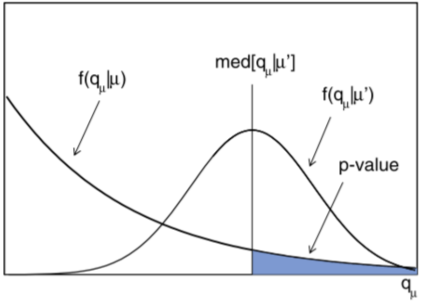




Asymptotics and Confidence Intervals

Applied Multi-Messenger Astronomy
Hans Niederhausen

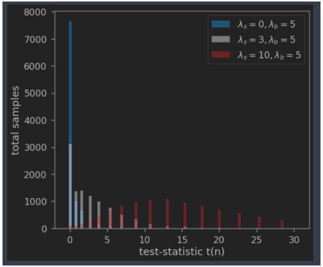
TUM - winter term 2020/21

Summary of 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 $\lambda\,(\mathrm{x})$ is defined as

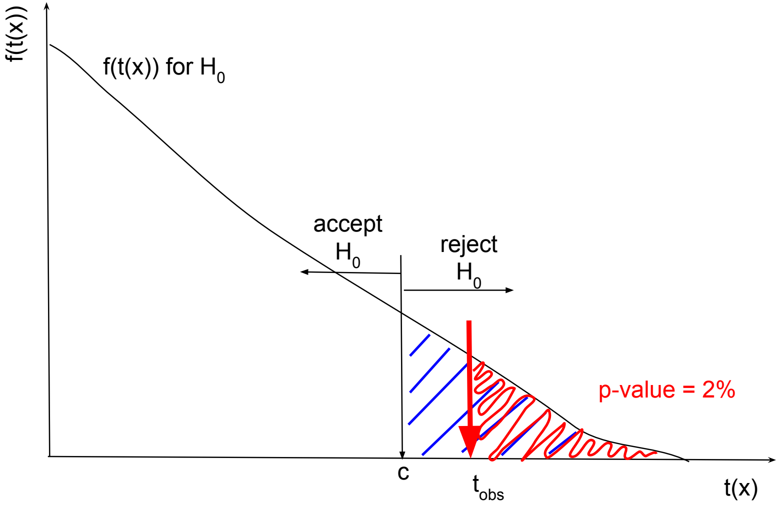
$$\lambda\,(\mathrm{x}) = -2\log\Lambda\,(\mathrm{x}) = -2\log\left\{\frac{sup_{\nu}\ L\,(\boldsymbol{\theta}_0,\,\boldsymbol{\nu}\,|\,\mathrm{x})}{sup_{\nu,\,\theta}\ L\,(\boldsymbol{\theta},\,\boldsymbol{\nu}\,|\,\mathrm{x})}\right\} \tag{1}$$



We discussed how to

- compute the distribution $f_\lambda(\lambda;\theta)$ using random number generation.
- calculate a critical value c give a desired value for α (the Type-1 error rate)
- calculate the observed test-statistic value and compare to the critical value

p-values



Questions about last week?

Questions about last week?

Next topic: large sample theory (asymptotics)

... or how large samples make life sooo much easier.

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)

Large Sample Theory: Point Estimation

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

$$E_\theta W(X) - \theta \ \text{(bias)} \tag{2}$$

$$E_\theta \left[\{W(X) - E_\theta W(X)\}^2 \right] \ \text{(variance)} \tag{3}$$

$$E_\theta \left(W(X) - \theta \right)^2 = Var_\theta W + \left(Bias_\theta W \right)^2 \ \text{(mean squared error)} \tag{4}$$

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

$$lim_{n\to\infty} P_\theta(|W_n - \theta| < \epsilon) = 1 \tag{5}$$

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)

Large Sample Theory: Point Estimation

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**

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathrm{X})\right)^2}{E_\theta\left([\frac{\partial}{\partial\theta}\log f(\mathrm{x}|\theta)]^2\right)} \tag{6}$$

which in the iid situation simplifies to

$$Var_\theta \geq \frac{\left(\frac{d}{d\theta}E_\theta W(\mathrm{X})\right)^2}{nE_\theta\left([\frac{\partial}{\partial\theta}\log f(x|\theta)]^2\right)} \tag{7}$$

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)

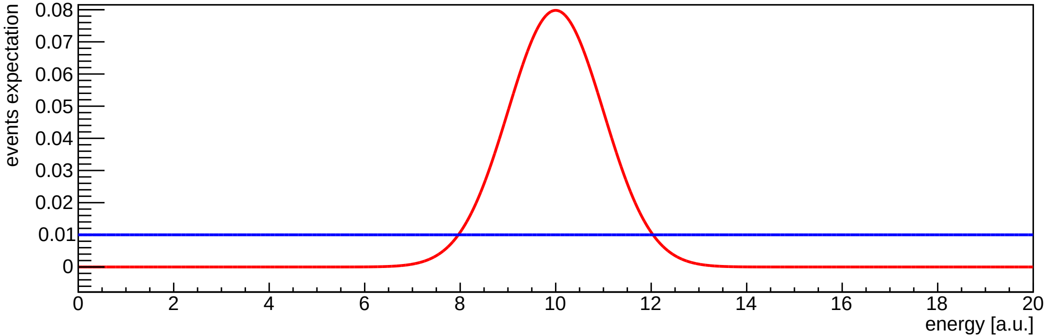
In summary: The MLE is ...

- **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 $\lambda_s + \lambda_b = N$

$$f_X(x;\mu,\sigma,\lambda_s,\lambda_b) = \frac{1}{\lambda_s+\lambda_b}\left[\lambda_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \lambda_b \cdot \frac{1}{20}\right] \tag{8}$$

becomes

$$f_X(x;\mu,\sigma,p_s) = \left[p_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + (1-p_s) \cdot \frac{1}{20}\right] \tag{9}$$

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

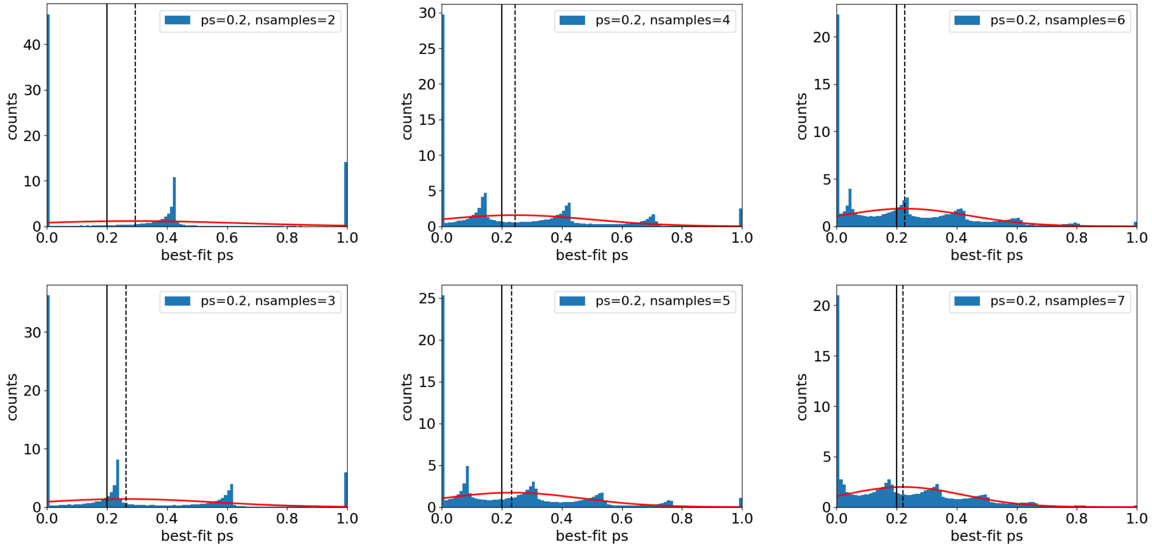
Large Sample Theory: The Toy Problem

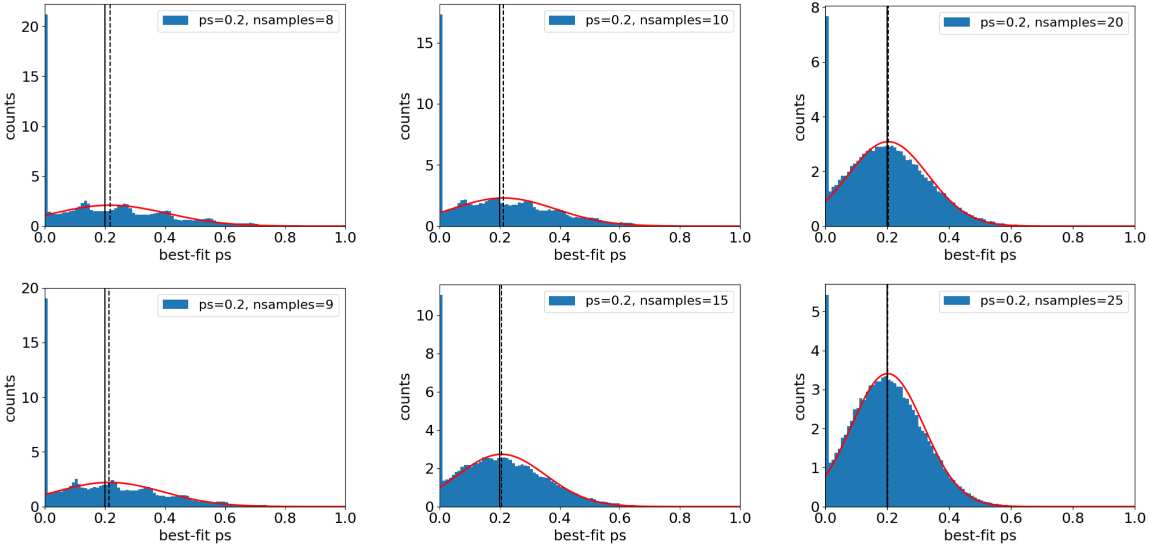
Here we compare the **distribution of the MLE** $\hat{p}_s$ of $p_s = 0.2$ for different sample sizes $N \in \{2, 3...10, 15, 20, ...100\}$.

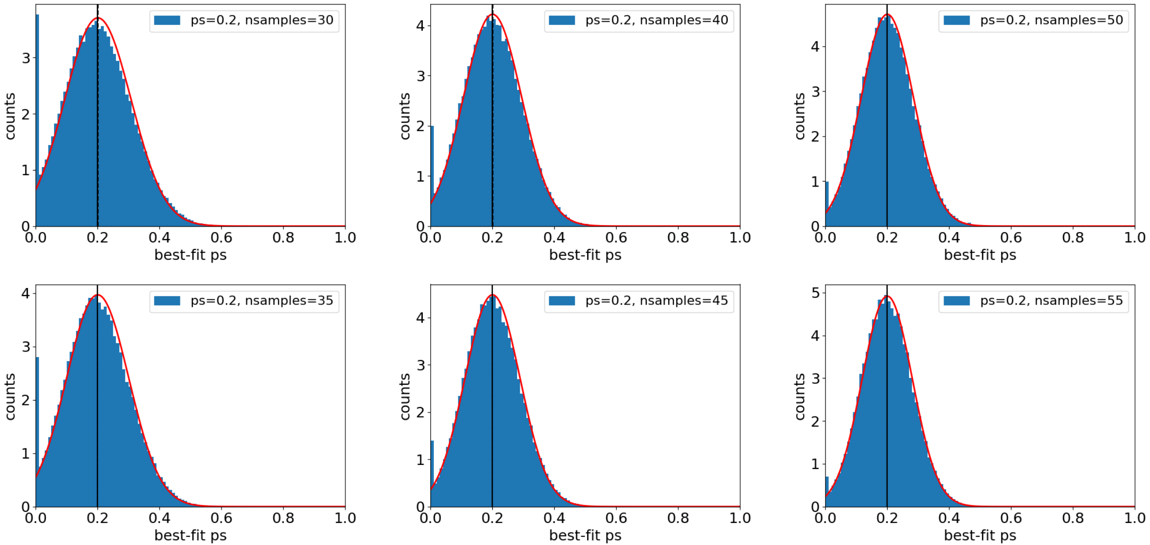
We also calculate numerically (using both, numeric integration, and sampling) the corresponding CRB

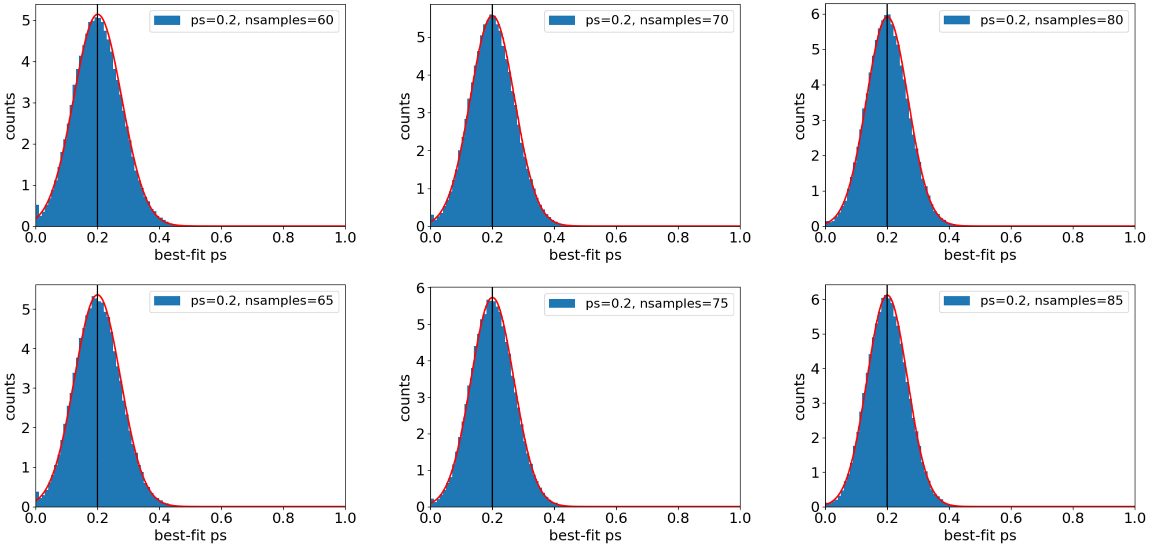
The CRB is compared to the observed variance $Var\hat{p}_s$ ($p_s = 0.2$).
Similiar to exercise 2, we need to generate many pseudo-datasets (here: for fix N) for this.

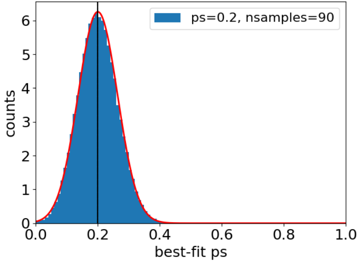
The MLE $\hat{p}_s$ clearly shows the expected convergence!

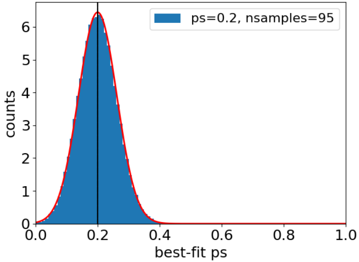

ps=0.2, nsamples=2
ps=0.2, nsamples=4
ps=0.2, nsamples=6
ps=0.2, nsamples=3
ps=0.2, nsamples=5
ps=0.2, nsamples=7
counts
best-fit ps

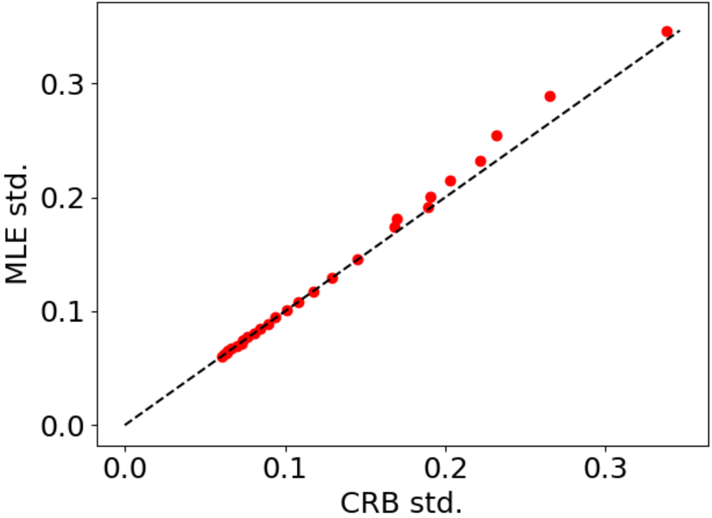

ps=0.2, nsamples=8
ps=0.2, nsamples=10
ps=0.2, nsamples=20
ps=0.2, nsamples=9
ps=0.2, nsamples=15
ps=0.2, nsamples=25
counts
best-fit ps


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.

Large Sample Theory: The Toy Problem

Questions?

Large Sample Theory: The Toy Problem

Questions?

Behavior of Likelihood Ratio Tests in large samples.

Reminder

given two hypotheses $H0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $H1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$
the likelihood ratio test-statistic $\lambda\left(\mathrm{x}\right)$ is defined as

$$\lambda\left(\mathrm{x}\right) = -2\log\Lambda\left(\mathrm{x}\right) = -2\log\left\{\frac{sup_{\nu}\ L\left(\boldsymbol{\theta}_0,\ \boldsymbol{\nu}\,|\,\mathrm{x}\right)}{sup_{\nu,\,\theta}\ L\left(\boldsymbol{\theta},\ \boldsymbol{\nu}\,|\,\mathrm{x}\right)}\right\} \tag{10}$$

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

$$\lambda \sim f_{\lambda}\left(\lambda;\ \boldsymbol{\theta},\ \boldsymbol{\nu}\right) \tag{11}$$

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!)

Large Sample Theory: Likelihood Ratio Testing

Wilk's Theorem

As the sample size increases, the distribution of the likelihood ratio test-statistic (11) 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 $k = dim\,\boldsymbol{\theta}$.

$$f_\lambda\left(\lambda;\, \boldsymbol{\theta}_0\right) \underset{n\to\infty}{\longrightarrow} \chi^2\left(k\right) \tag{12}$$

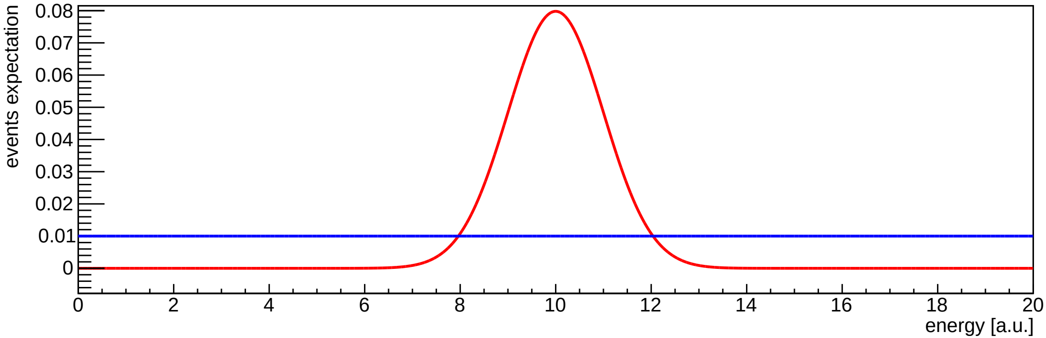
Wilk's Theorem (cont'd)
Unfortunately there are strict regularity conditions. Here are the two most important ones

- $\boldsymbol{\theta}_0$ needs to be an interior point of Θ
- 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

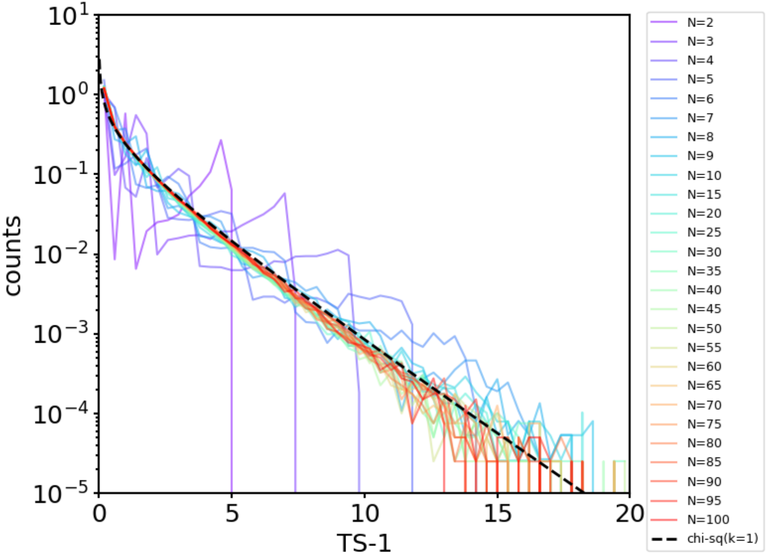
Application to our standard toy problem (with 2 parameters: p_s, μ_s)



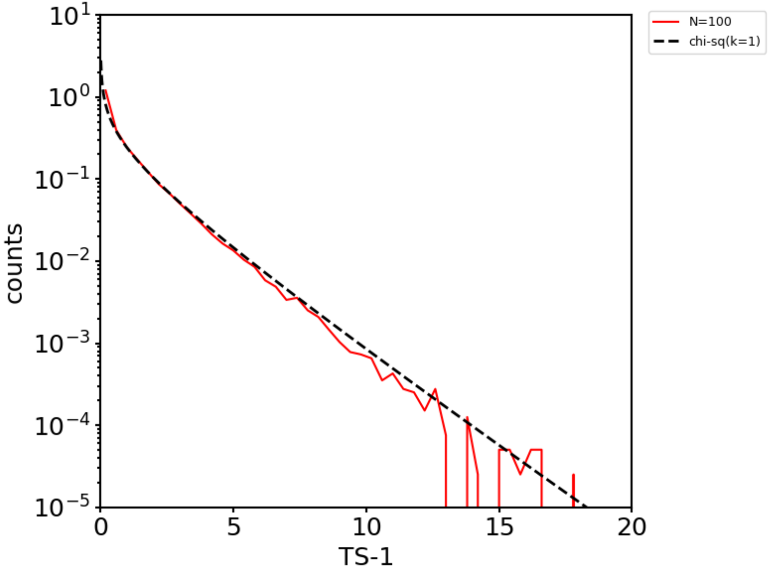
Two different hypothesis tests satisfying Wilk's theorem

Case 1: $H0 : p_s = 0.2$ and $H1 : p_s \neq 0.2$ (k=1) (μ_s is nuisance!)

Large Sample Theory: The Toy Problem

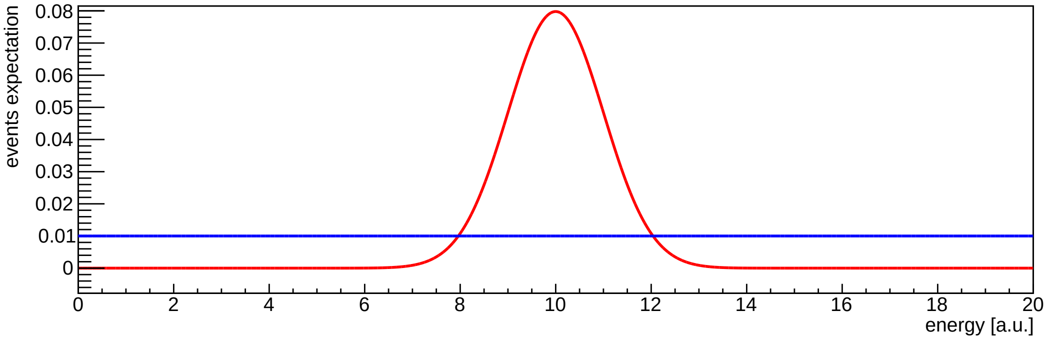


Large Sample Theory: The Toy Problem



Large Sample Theory: The Toy Problem

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

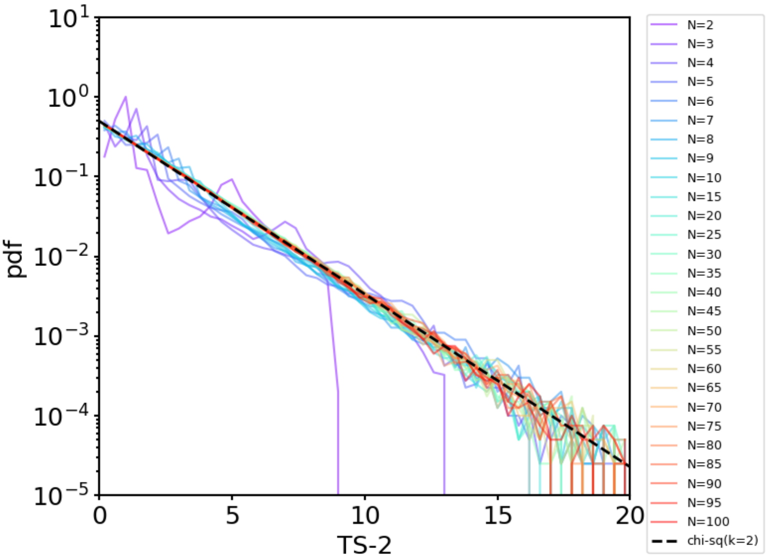


Two different hypothesis tests satisfying Wilk's theorem

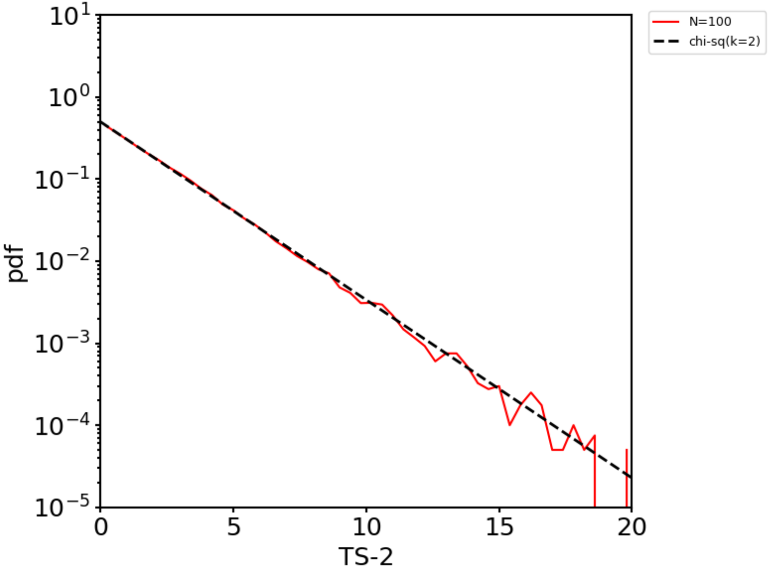
Case 1: $H0 : p_s = 0.2$ and $H1 : p_s \neq 0.2$ (k=1)

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

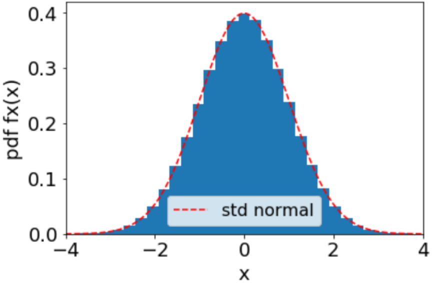


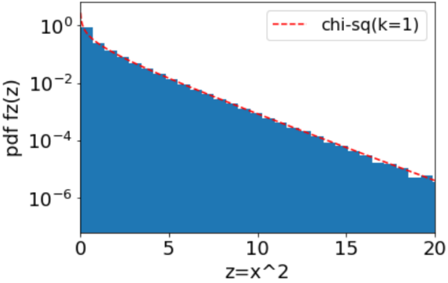
Large Sample Theory: The Toy Problem



Aside: Normal to χ^2

square of a std. normal rv $g(x) = x^2$, $X \sim N(0, 1)$





Critical Values from Chi-squared Distribution

for different number of degrees of freedom ($n = \Delta\mathrm{dim}\theta$)
and various choices of common levels α.

(here: critical value $c \equiv Q_\alpha$))

$1-\alpha$	Q_α				
	$n=1$	$n=2$	$n=3$	$n=4$	$n=5$
0.683	1.00	2.30	3.53	4.72	5.89
0.90	2.71	4.61	6.25	7.78	9.24
0.95	3.84	5.99	7.82	9.49	11.1
0.99	6.63	9.21	11.3	13.3	15.1

Questions?

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).

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 X we calculate function values $L(\mathrm{X})$ and $U(\mathrm{X})$ which are random variables.

$$[L(\mathrm{X}),\ U(\mathrm{X})] \quad (two - sided) \tag{13}$$
$$(-\infty,\ U(\mathrm{X})] \quad or \quad [L(\mathrm{X}),\ \infty) \quad (one - sided) \tag{14}$$

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.)

Confidence Intervals: Coverage

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

$$P_\theta\left(\theta \in [L(\mathrm{X}),\ U(\mathrm{X})]\right) \tag{15}$$

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

$$inf_\theta P_\theta\left(\theta \in [L(\mathrm{X}),\ U(\mathrm{X})]\right) = 1-\alpha \tag{16}$$

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 ($\mathrm{X} = \{X_1, ..., X_n\}$). This can be done using a **pivot** (a function of the parameter and observations, that has a distribution which is independent of the parameter).

$$Q(\mu, \mathrm{X}) = \frac{\bar{X} - \mu}{\sigma // \sqrt{n}} \tag{17}$$

$$Q \sim N(0, 1) \tag{18}$$

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

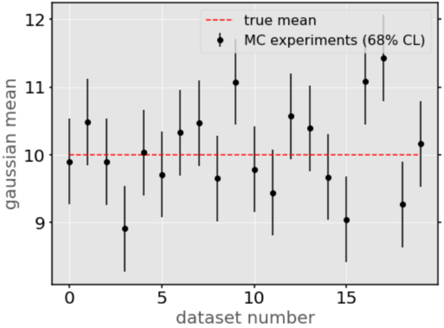
$$P_\mu \left(-a \leq Q \leq a \right) = 1 - \alpha \tag{19}$$

which corresponds to the following confidence set

$$\left\{ \mu : \; \bar{X} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}} \right\} \tag{20}$$

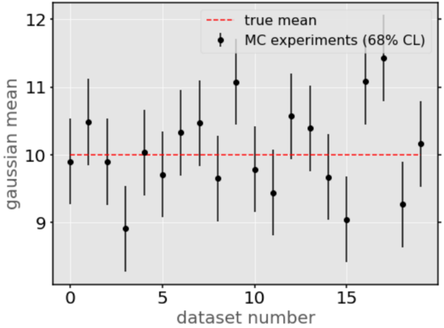
Confidence Intervals: Coverage in the normal mean problem

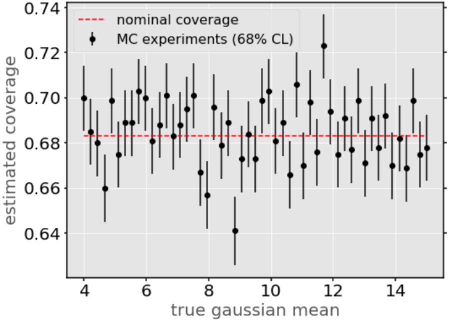
Check with Monte-Carlo (see ipython notebook)



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 $1 - \alpha$.

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)

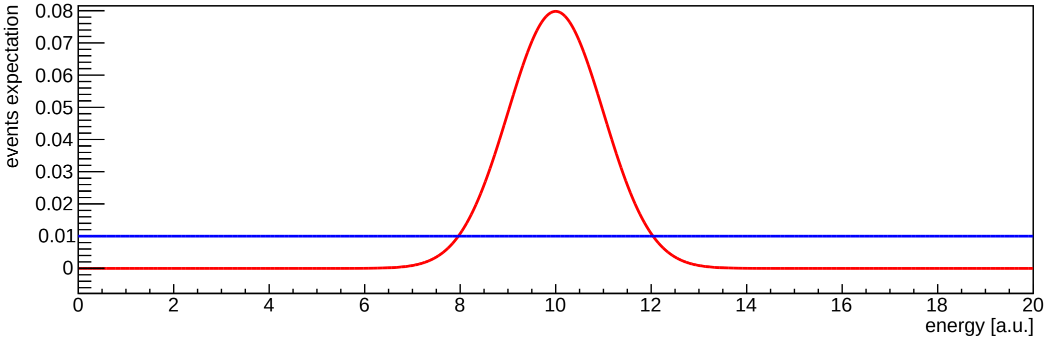
Confidence Intervals from inversion of hypothesis tests

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

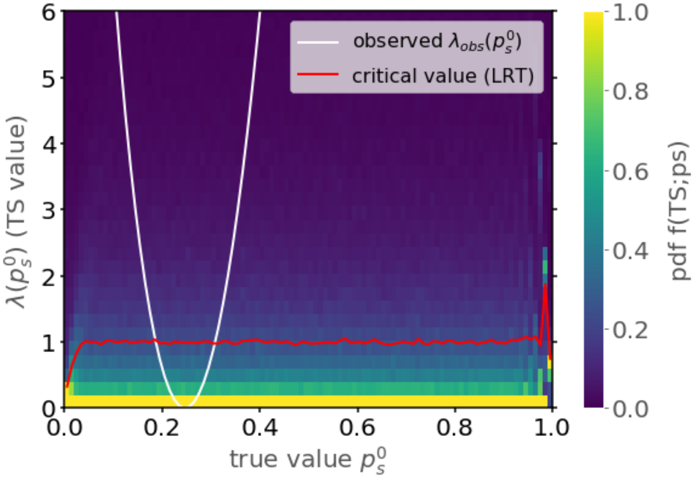
We have learned how to construct likelihood ratio tests. Let's invert a likelihood ratio test to obtain a confidence set on the signal fraction p_s in our toy model. (see ipython notebooks)



$$f_X(x;\mu,\sigma,p_s) = \left[p_s \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} + (1-p_s)\cdot\frac{1}{20}\right] \tag{21}$$

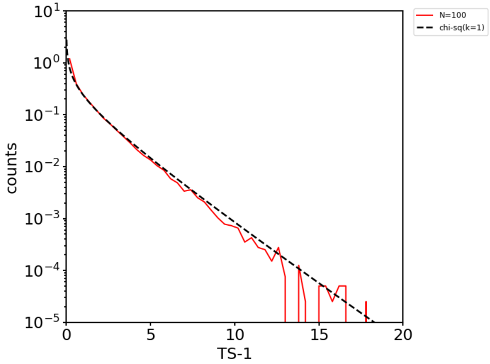
Confidence Intervals from inversion of LRT in toy problem.

$H_0 : p_s = p_s^0$ 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)



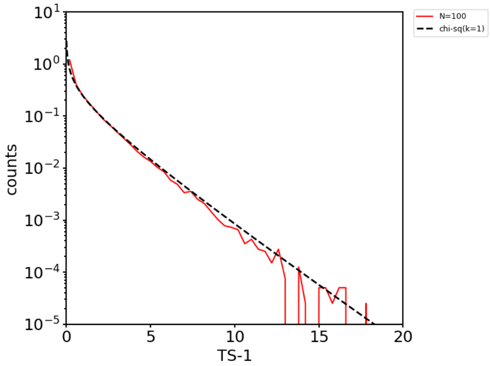
Reminder: Asymptotic Chi-Squared.

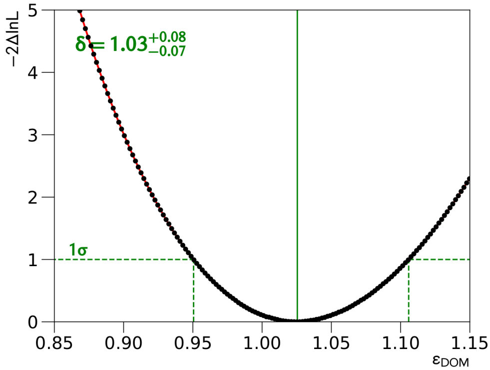
For **large n**:
The distribution of λ is independent from (true) p_s^0 with $\lambda \sim \chi^2(k = 1)$
We just have to find the points, where λ changes by 1 (for $1 - \alpha = 0.68$).
(because $p(\lambda \geq c) = \alpha$ yields $c = 1$ for k=1 and $1 - \alpha = 0.68$.



Reminder: Asymptotic Chi-Squared.

For **large n**:
The distribution of λ is independent from (true) p_s^0 with $\lambda \sim \chi^2(k=1)$
We just have to find the points, where λ changes by 1 (for $1-\alpha = 0.68$).
(because $p(\lambda \geq c) = \alpha$ yields $c = 1$ for k=1 and $1-\alpha = 0.68$.



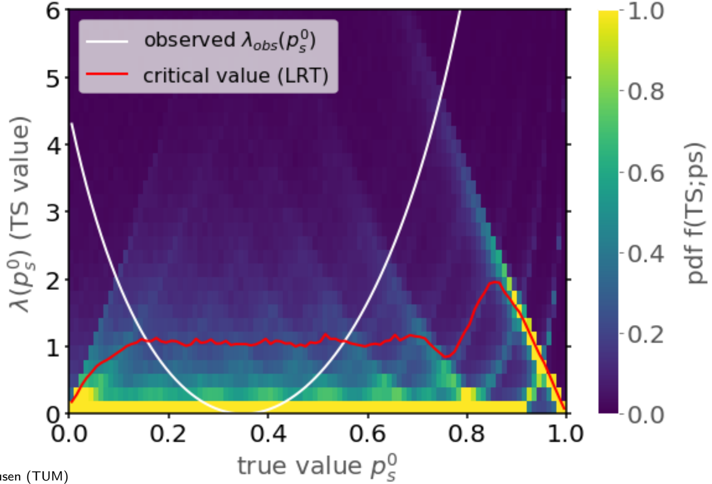


Questions?

Confidence Intervals from inversion of LRT in toy problem.

The problem is much harder, if Wilk's theorem does not apply.

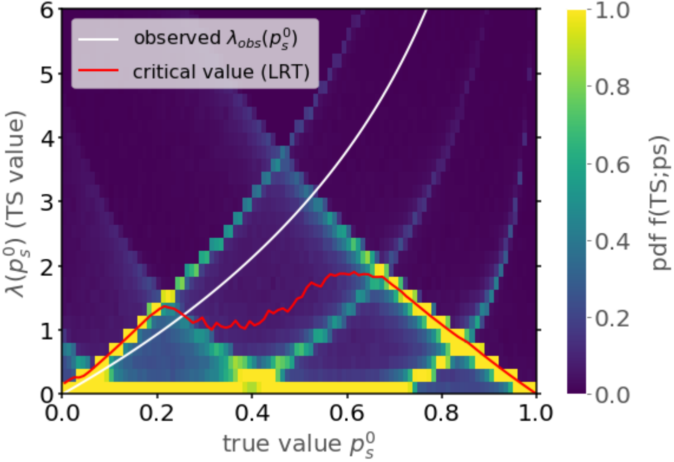
$H_0 : p_s = p_s^0$ and $H_1 : p_s \neq p_s^0$, sample size n=10
endpoints of interval: intersection points of obs. TS value (white) with critical value (red)



Confidence Intervals from inversion of LRT in toy problem.

The problem is much harder, if Wilk's theorem does not apply.

$H_0 : p_s = p_s^0$ 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)

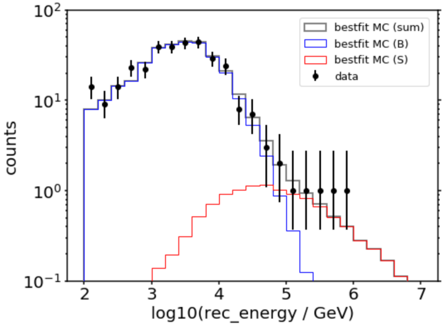


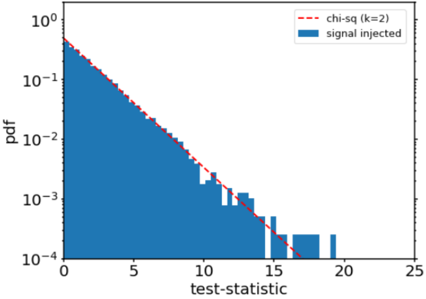
Summary: Confidence Intervals from inversion of LRT

- 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

Example: Inversion of LRT in the IceCube diffuse flux measurement

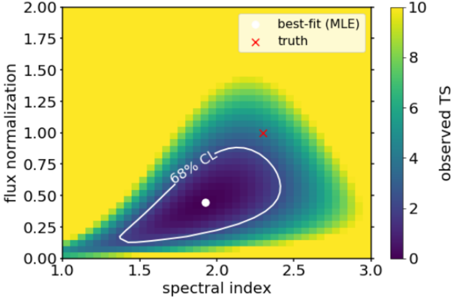
To construct a joint confidence interval for the normalization and spectral index of the astrophysical neutrino flux, we need to invert a LRT:
$H_0 : (\Phi, \gamma) = (\Phi_0, \gamma_0)$ and $H_1 : (\Phi, \gamma) \neq (\Phi_0, \gamma_0)$
The asymptotic expectation for the TS distribution would be χ^2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





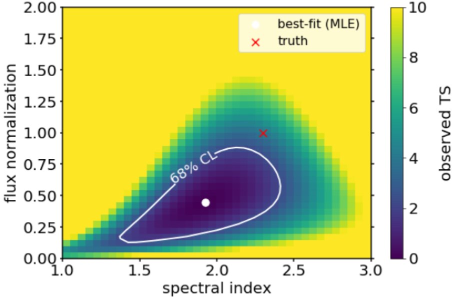
Example: Inversion of LRT in the IceCube diffuse flux measurement

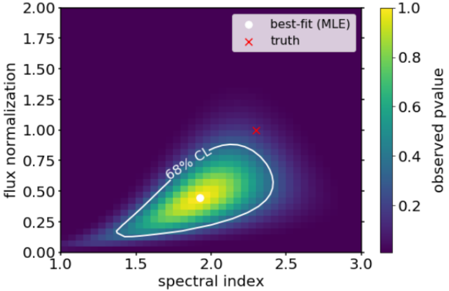
if we have sufficient data, we use the χ^2 pdf (left) otherwise we need to obtain (valid) p-values from MC simulations and use those to get the contours (right)



Example: 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)





$$p(x_{\text{obs}}) = \sup_{\theta \in \Theta_0} P_\theta \left(TS(X) \geq TS(x_{\text{obs}}) \right) \tag{22}$$

Questions?

BONUS

More Theory of Random Variables: Transformations

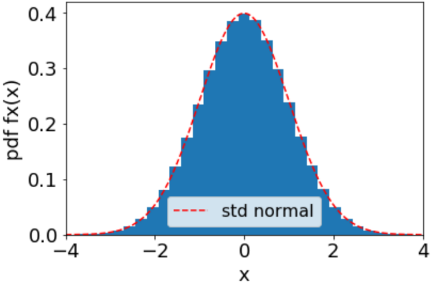
in many problems it can be useful to work with transformed random variables.
assume $x \sim f_X(x)$ - what is the distribution $f_Z(z)$ of $z = g(x)$ ($g(x)$ some function)?
If $g(x)$ is monotone, then

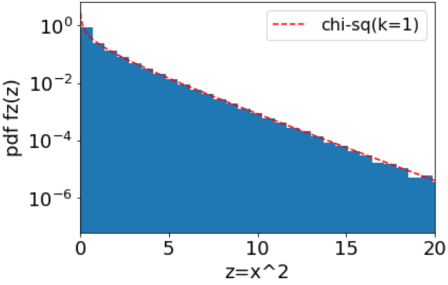
$$f_Z(z) = \begin{cases} f_X(g^{-1}(z))|\frac{d}{dz}g^{-1}(z)|, & z \in Z \\ 0 & \textit{otherwise} \end{cases} \tag{23}$$

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

More Theory of Random Variables: Transformations

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

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) \tag{24}$$

$$g^{-1}(z) = \begin{cases} -\sqrt{z}, & x < 0 \\ \sqrt{z}, & x > 0 \end{cases} \tag{25}$$

$$\begin{aligned} f_z(z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-(-\sqrt{z})^2/2\right) \frac{1}{2\sqrt{z}} + \frac{1}{\sqrt{2\pi}} \exp\left(-(\sqrt{z})^2/2\right) \frac{1}{2\sqrt{z}} && (26) \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{z}} \exp\left(-z^2/2\right) && (27) \end{aligned}$$