Notes on χ^2: What is it and why the hell are we using it?

(Dated: January 1, 2025)

THE FUNDAMENTALS: PDFS, CDFS, QUANTILE FUNCTIONS

- PDF: $f(x)$ The Probability Density Function. The thing you integrate to 1 over its domain D.

$$f : D \to \mathbb{R} \tag{1}$$

$$\int_D f(x)dx = 1 \tag{2}$$

- CDF: $F(x)$. The Cumulative Distribution function. The integral up to some point x of the PDF;

$$F : D \to [0, 1] \tag{3}$$

$$F(x) = \int_{min\{D\}}^{x} f(x')dx' \tag{4}$$

  It represents the probability that the variate X is less than x. In the multivariate case, it is

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots X_n \leq x_n) \tag{5}$$

- The quantile function: F^{-1} the inverse of the CDF F

$$F^{-1} : [0, 1] \to D \tag{6}$$

I. WHY χ^2? CONNECTION TO THE GAUSSIAN NULL HYPOTHESIS

We have to introduce the idea of a *Hypothesis*. A Hypothesis is a model we have that tries to explain data under certain assumptions. In particle physics usually we have two hypotheses that we want to compare:

- The Null Hypothesis, H_0: The hypothesis that no new physics is present; the data is described by known physics that we assume is true
- The test Hypothesis, H_1: The hypothesis that the data is described by H_0 *and* some additional physical process that we are interested in finding out evidence for

Caveat: "new physics" could mean lots of things. For example, I might be interested in detecting the CEνNS process in my experiment. For me, then, *new physics* is defined by just the events measured from coherent neutrino nucleus recoils in the detector (the thing I'm trying to prove exists) and the *known physics* comprises all the backgrounds and any other events I should expect to see from running the experiment. In another analysis, I might be looking for evidence of dark matter, and in that case CEνNS could be a background/known-physics for me (part of H_0) and the dark matter signal is my H_1.

Now let us imagine our observable for our experiment is particle counts distributed into 3 energy bins (see Fig. 1). Considering our null hypothesis which we use to calculate the expectation or prediction for the observable, μ_i (blue), binned over energy bins i = 1,2 and 3 we can then go measure the data over these bins, getting observed particle counts d_i for which we suppose the errors are known (black). The question we are asking is how well separated our prediction is from the data. More explicitly, the χ^2 helps us answer the question **"Is the data consistent with normally-distributed fluctuations about the expectation value in each bin?"**

See Fig. 1 for example. You can imagine that each data point is a different *degree of freedom*, and we can represent the observed data in its totality by a point in the (bin1, bin2, bin3) vector space on the right plot of Fig. 1. The prediction coordinates are represented by the blue vector while the data coordinates are represented by the black vector and an associated error ellipsoid[1]. The distance-squared between the two points is $d^2 = (\mu_1 - d_1)^2 + (\mu_2 - d_2)^2 + (\mu_3 - d_3)^2$. The χ^2 has a similar structure, except we divide each of the coordinates by the expected error in each - in this sense, you can really think of the χ^2 as a distance measure between the observation (data) and the expected distribution.

Now, here's the premise for using a χ^2:

Some observable is distributed according to Gaussian statistics

This is because of a theorem: if a random variable Z follows a normal distribution, then Z^2 follows a χ^2 distribution. As a corollary, if the data lives in the α-quantile (or alternatively $n\sigma$ away from the central

[1] We take an ellipsoid for the errors because at this stage I have assumed that all the bin counts are *independent* random variables – if they have correlations, that will add a layer of complexity.

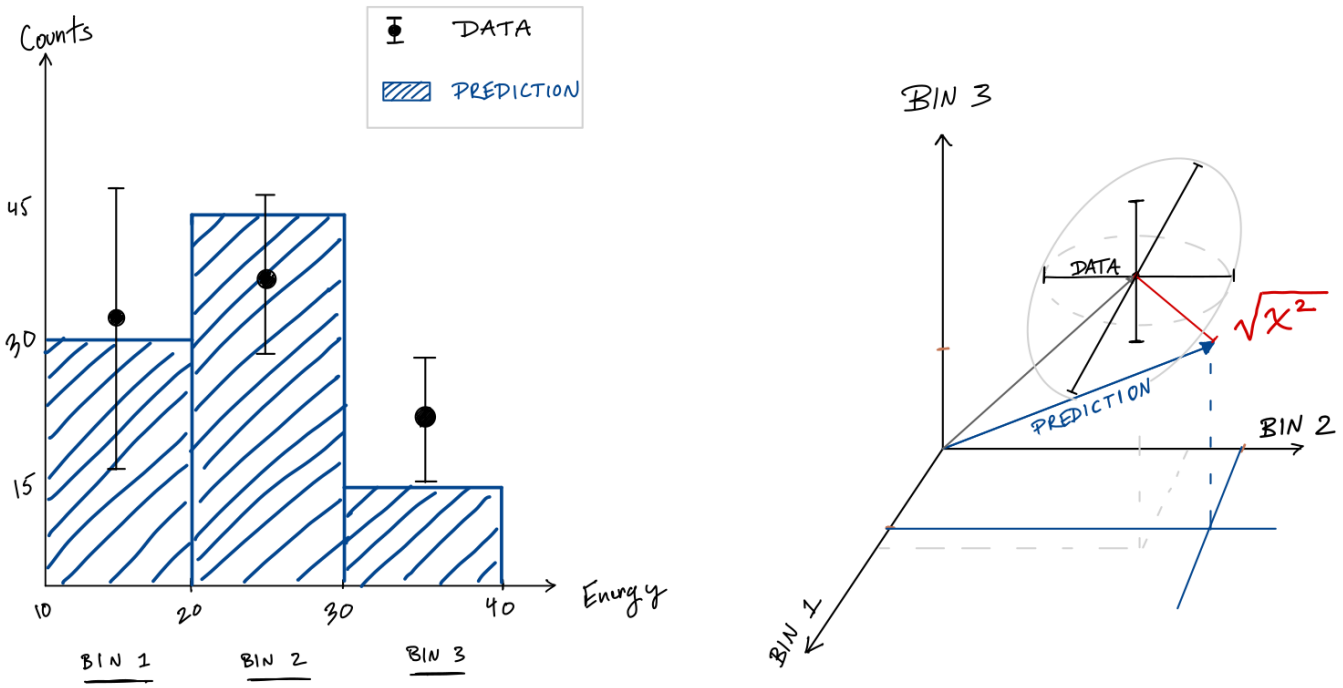

FIG. 1. The geometric interpretation of the χ^2 for 3 degrees of freedom.

value) of a gaussian distribution about our test hypothesis, then the χ^2 will also live in the α-quantile (or $n\sigma$) of the χ^2 distribution. Mathematically,

$$\chi^2 \leq F^{-1}(\alpha, \nu) \tag{7}$$

where F^{-1} is the inverse CDF or quantile function of the χ^2 distribution.

II. SETTING LIMITS / SENSITIVITY CURVES IN A MODEL PARAMETER SPACE

What do we do when we set limits? Given data (d), signal (s), and background (b), we evaluate the binned χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(d_i - s_i - b_i)^2}{\sigma_i^2} \tag{8}$$

But for a sensitivity plot you don't have data, instead you want to ask the question:

"if the experiment runs and measures data consistent with the null hypothesis H_0, then how much parameter space can be ruled out as being inconsistent with the measurement?"

In this case you set $d_i = b_i$, and take the error $\sigma_i = \sqrt{b_i}$ which is the Poisson error. This means effectively computing

$$\chi^2 = \sum_{i=1}^{N} \frac{(s_i)^2}{b_i} \tag{9}$$

and since the $\Delta\chi^2 = \chi^2 - \chi^2_{min}$, and $\chi^2_{min} = 0$ in our case of $d_i = b_i$, then the above is equivalent to the $\Delta\chi^2$. This identification is super handwavy and will get criticized a lot by people that know stats. In a more realistic analaysis we would instead simulate pseudoexperiment data for d_i based on a Poisson-distributed set of events $d_i \sim P(\mu = b_i)$, but this quick method above circumnavigates it for theorists' simplicity. A more genuine approach is illustrated in § VI.

Example 1: Consider a simple scenario where we predict background counts b_i over energy bins in the interval $E \in [1, 100]$ in arbitrary units. We might want to test a signal model that takes on a Gaussian energy distribution;

$$s(E_i) = 10^4 g^2 \exp\Big(-(E_i - m)^2/200 \Big) \; : \text{ Counts in energy bin } E_i \tag{10}$$

Why did I pick a gaussian distribution? No reason in particular. I just want to have a toy signal distribution to play with. If you want, you could substitute this shape for a realistic distribution e.g. from dark matter or neutrino scattering. For us the parameter m changes the peak of the signal and g changes the normalization (suggestive of a mass and coupling, respectively). Now I can make up a background distribution and define this signal function in python like so;

```
import matplotlib.pyplot as plt
import numpy as np

# Energy bins
energy_bins = np.linspace(5.0, 100.0, 10)
energy_centers = (energy_bins[1:] + energy_bins[:-1])/2

# Background data
bkg = np.array([25000.0, 2700.0, 1000.0, 120.0, 170.0, 145.0, 90.0, 110.0, 118.0])

# Signal model
def signal_event_rate(g, m):
    return 1e4 * g**2 * np.exp(-(energy_centers - m)**2 / 200.0)
```

A sample signal distribution plotted against a sample background in Fig. 2.

Now we can scan across the parameter space. In python, we could do this in a brute-force way by instantiating a grid of (m, g) values to loop over;

```
# Parameter scan
g_array = np.logspace(-3, 0, 100)
m_array = np.linspace(1.0, 100.0, 100)
g_list = []
```

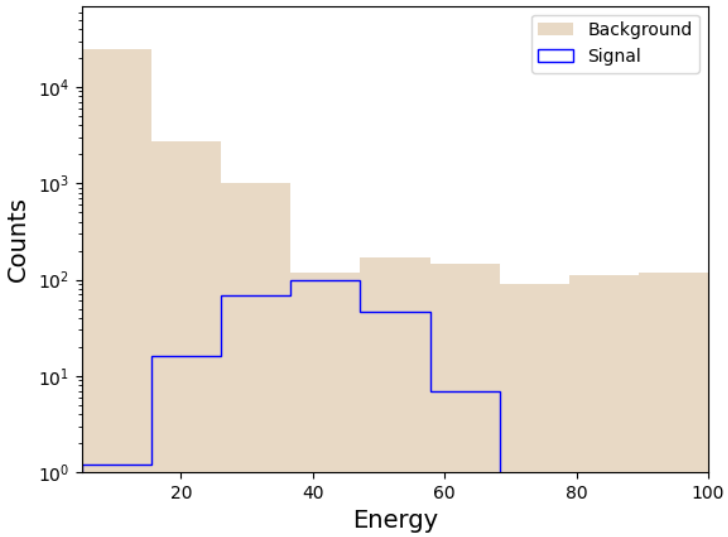

FIG. 2. Example background spectrum and signal event distribution over energy bins.

```
m_list = []
chi2_list = []
for m in m_array:
    for g in g_array:
        signal = signal_event_rate(g, m)
        chi2 = np.sum(signal**2 / bkg)
        g_list.append(g)
        m_list.append(m)
        chi2_list.append(chi2)
```

Together, `g_list`, `m_list`, and `chi2_list` form a map of $\chi^2(m, g)$. We can then go ahead and plot iso-contours of constant χ^2 in the (m, g) space for the 1σ, 2σ, and 3σ contours corresponding to constant $\Delta\chi^2$ values of 2.3, 6.18, amd 11.83, respectively; I plot this in Fig. 3 using the following code:

```
    M_GRID, G_GRID = np.meshgrid(m_array, g_array)
    CHI2_GRID = np.reshape(chi2_list, (100, 100)).transpose()
    plt.contour(M_GRID, G_GRID, CHI2_GRID, levels=[2.3, 6.18, 11.83], colors=["maroon",
        "indianred", "lightcoral"], linewidths=1.0)
    plt.text(20.0, 0.2, "Consistent with Background/H0\n(Experiment is sensitive to this
        region)", fontsize=12)
    plt.text(30.0, 0.005, "Experiment not sensitive\nto this region", fontsize=12)
    plt.ylim((1e-3, 1.0))
    plt.xlim((1.0, 100.0))
    plt.yscale('log')
    plt.ylabel(r'$g$', fontsize=14)
```

```
    plt.xlabel(r"$m$", fontsize=14)
    plt.show()
```

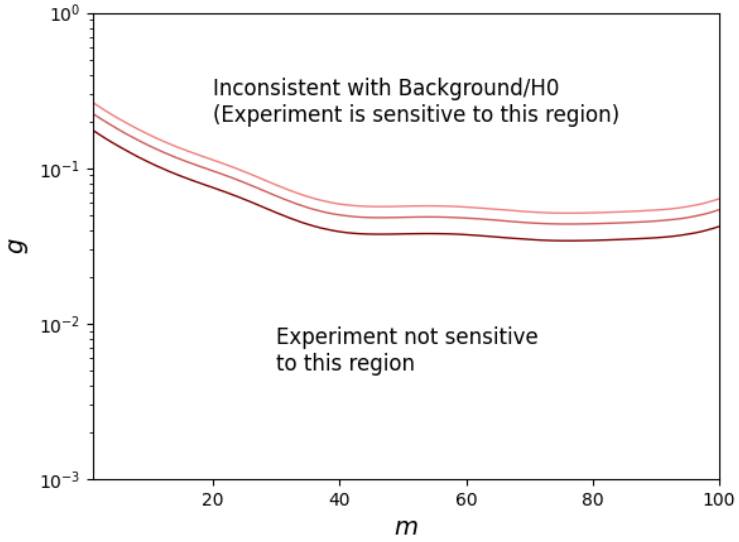

FIG. 3. Sensitivity curves: the 1σ (dark red), 2σ (red) and 3σ (light red) contours for $\chi^2 = 2.3, 6.18$, and 11.83, respectively. A larger χ^2 is interpreted as being more inconsistent with the background expectation to a higher degree of confidence, hence these are called confidence levels (CLs).

Notice how the sensitivity curves are roughly flat above $m = 40$ and start losing sensitivity below $m = 40$; this reflects the fact that the background distribution I chose is much larger at lower energies. So for lower values of m, most of the signal will start peaking in those low-energy bins and it will be statistically much harder to tell if we are seeing the signal. Mathematically, this is because the signal rate in those bins has to exceed 1σ in the background error which is $\sigma = \sqrt{b_i}$; the larger the background, the larger the signal has to be to be statistically significant.

III. PARAMETER ESTIMATION USING REAL (FAKE) DATA

Now let's put in some real fake data into the picture. I'll make up some fake data over the energy bins E_i by drawing normally distributed random numbers, as shown in Fig. 4, with a simulated excess of events above the null expectation using our signal model;

```
from scipy.stats import norm
test_signal = signal_event_rate(0.08, 75.0)
data = norm.rvs(loc=bkg+test_signal, scale=np.sqrt(bkg+test_signal))
errors = np.sqrt(data)
```

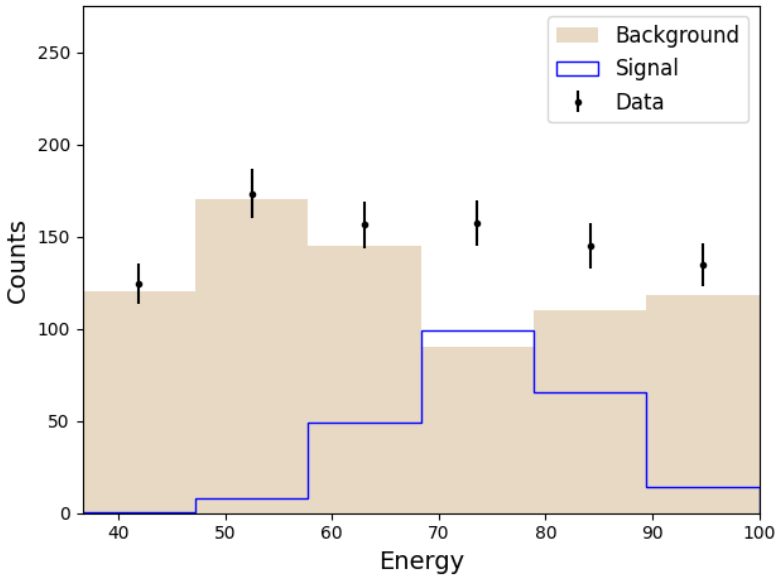

FIG. 4. Data with an excess. The first 3 bins of Fig. 2 with large counts are omitted to fit the error bars onto a linear scale.

Now we wish to scan over the parameter space just as before, except now that we have data we should modify our χ^2 formula accordingly;

```
for m in m_array:
    for g in g_array:
        signal = signal_event_rate(g, m)
        chi2 = np.sum((signal+bkg-data)**2 / bkg)
        g_list.append(g)
        m_list.append(m)
        chi2_list.append(chi2)
```

Since we are interested in the $\Delta\chi^2$ to see if a region in our parameter space is a good explanation of the data, we have to subtract off the minimum χ^2:

```
M_GRID, G_GRID = np.meshgrid(m_array, g_array)
CHI2_GRID = np.reshape(chi2_list - min(chi2_list), (100, 100)).transpose()
```

The resulting plot is shown in Fig. 5, where you can see a set of closed contours at the 1σ, 2σ, and 3σ levels. In this case the $\chi^2_{min} = 6.06$, which is an excellent fit for the number of degrees of freedom we have (9 bins). To check this, evaluate the the CDF of the χ^2 at the minimum;

```
from scipy.stats import chi2
print(chi2.cdf(6.06, df=9))
```

we find $F(\chi^2_{min}, \nu = 9) = 0.26$, which means that our fit is within the 26% quantile, well under 1σ ($\simeq 68\%$).

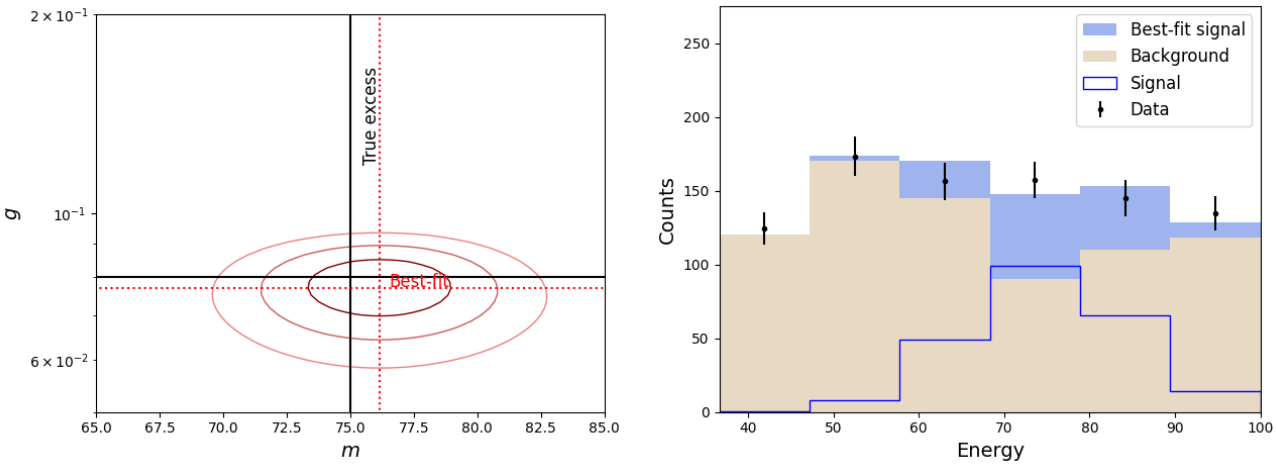

FIG. 5. Left: Sensitivity contours in constant $\Delta\chi^2$; the 1σ (dark red), 2σ (red) and 3σ (light red) contours for $\chi^2 = 2.3, 6.18$, and 11.83, respectively. Right: the energy spectrum shown again with the signal model (m, g) that gives the best-fit χ^2_{min} added as the blue histogram stacked onto the background model. The first 3 bins are omitted.

We show this best-fit in Fig. 5, right. It's important to check both the absolute χ^2 for a given number of d.o.f. and the $\Delta\chi^2$ because they tell you different things; while the $\Delta\chi^2$ tells you where the best-fit is, it doesn't tell you how good of a fit it is. There might be a point in the signal model parameter space that minimizes the χ^2, doing a better job than the null hypothesis, but if the χ^2_{min} lies in the $> 99\%$ quantile ($> 3\sigma$) then it might not be the most attractive solution to the excess. Perhaps another signal model would fit better.

IV. GENERALIZATION TO A LOG-LIKELIHOOD TEST STATISTIC

Sometimes we are dealing with data with low statistics, e.g. $N = \mathcal{O}(10)$ or so. In these cases the hypothesis that our observation is consistent with normally-distributed fluctuations about some expectation is no longer good. For low-stats expectations, we need to use **Poisson** statistics instead. In this case, instead of using the χ^2 formula, we'll use a binned Poissonian log-likelihood. For binned data d_i, background expectation b_i, and a signal hypothesis s_i over N bins, the Poisson log-likelihood is

$$\ln L(\vec{\theta}) = \sum_{i=1}^{N} d_i \ln\left(s_i(\vec{\theta}) + b_i\right) - \left(s_i(\vec{\theta}) + b_i\right) - \ln\left(d_i!\right) \tag{11}$$

In fact, because of the central limit theorem saying that Poisson distributions turn into Gaussian distributions in the large stats limit, so actually we can use this Poisson log-likelihood for both the low and high stats cases.

One modification should be made to this if you are making a sensitivity plot using the method in : since for sensitivity plots we take d_i to be our null expectation b_i, it

V. ADDING SYSTEMATIC UNCERTAINTIES

VI. MOTIVATING A SIMPLER CHI2 TEST

Suppose we have a set of independent background expectations B_i and an expected signal S_i across bins $i = 1, \ldots, \nu$. In the absence of data, we can ask the question: "If our experiment sees data consistent with background-only, does the median value of the χ^2 distribution tested against our $S+B$ model exceed the $1-\alpha$ quantile of the background-only χ^2 distribution?"

To do this we generate a set of independent, normally distributed fake data $d_i \sim N(B_i, \sqrt{B_i})$. In the limit of a large number of pseudo-experiments, the $1-\alpha$ quantile of this distribution approaches the theoretical limit $F^{-1}(1-\alpha, \nu)$. Now consider the expectation value of the χ^2 tested against our $S+B$ expectation;

$$\begin{aligned}
\mathrm{E}[\chi^2_{S+B}] &= \mathrm{E}\left[\sum_{i=1}^{\nu} \frac{(d_i - S_i - B_i)^2}{B_i}\right] \\
&= \sum_{i=1}^{\nu} \left(\mathrm{E}\left[\frac{d_i - S_i - B_i}{\sqrt{B_i}}\right]\right)^2 + \mathrm{Var}\left[\frac{d_i - S_i - B_i}{\sqrt{B_i}}\right] \\
&= \sum_{i=1}^{\nu} \left(\mathrm{E}\left[\frac{d_i}{\sqrt{B_i}}\right] - \mathrm{E}\left[\frac{S_i + B_i}{\sqrt{B_i}}\right]\right)^2 + \frac{1}{B_i}\mathrm{Var}\left[d_i\right] \\
&= \sum_{i=1}^{\nu} \left(\frac{B_i}{\sqrt{B_i}} - \frac{S_i + B_i}{\sqrt{B_i}}\right)^2 + \frac{1}{B_i} B_i \\
&= \sum_{i=1}^{\nu} \left(\frac{S_i^2}{B_i} + 1\right) \\
&= \nu + \sum_{i=1}^{\nu} \frac{S_i^2}{B_i}
\end{aligned} \tag{12}$$

Above we used the linearity of expectation values and the assumption that d_i are independent, and that $\mathrm{Var}\left[d_i\right] = (\sqrt{B_i})^2$. The test statistic that I used is the second term;

$$t \equiv \sum_{i=1}^{\nu} \frac{S_i^2}{B_i} = \mathrm{E}[\chi^2_{S+B}] - \nu \tag{13}$$

We can then make use of the following approximation of the median of a χ^2 distribution with k degrees of freedom

$$\mathrm{Median}[\chi^2] \simeq k\left(1 - \frac{2}{9k}\right)^3 = k - \frac{2}{3} + \mathcal{O}(k^{-1}) \tag{14}$$

Since $\mathrm{E}[\chi^2_k] = k$, for $k \gtrsim 13$ we have $\mathrm{Median}[\chi^2] \simeq \mathrm{E}[\chi^2_k]$ to within 5% error. Our original question requires that

$$\mathrm{Median}[\chi^2_{S+B}] > F^{-1}(1-\alpha, \nu) \equiv \chi^2_\alpha \tag{15}$$

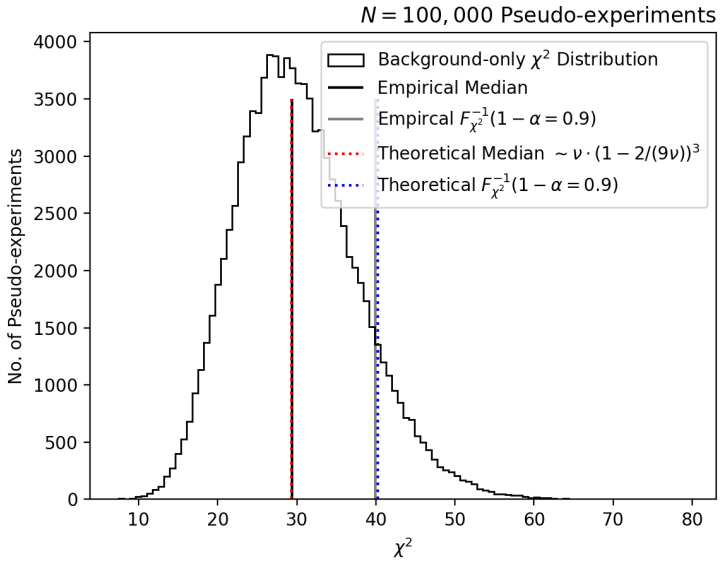

FIG. 6. Pseudo-experiment data for a given background over $\nu = 30$ bins, and a comparison of the empirically calculated and theoretical positions of the median χ^2 and $\chi^2_{90\%}$ values with 10,000 psuedo-experiments. The empirical median and theoretical lines lie on top of each other.

it follows that we need to find the smallest signal such that

$$t \gtrsim \chi^2_\alpha - \nu. \tag{16}$$

For example, if we take $\nu = 24$ bins, and $\alpha = 0.1$, then we have

$$t \gtrsim 33.2 - 24 \simeq 9.2 \tag{17}$$

This differs from the critical value I chose, $t = 4.61$ based on 2 free parameters. It seems, then, that Eq. 16 *could* be used for CL setting in lieu of running pseudo-experiments only under the following special conditions;

- If the data d_i can be safely assumed to be independent
- If the data d_i can be safely assumed to be normally distributed
- If the number of bins $\nu \gtrsim \mathcal{O}(10)$, otherwise the approximation for Median$[\chi^2]$ loses precision

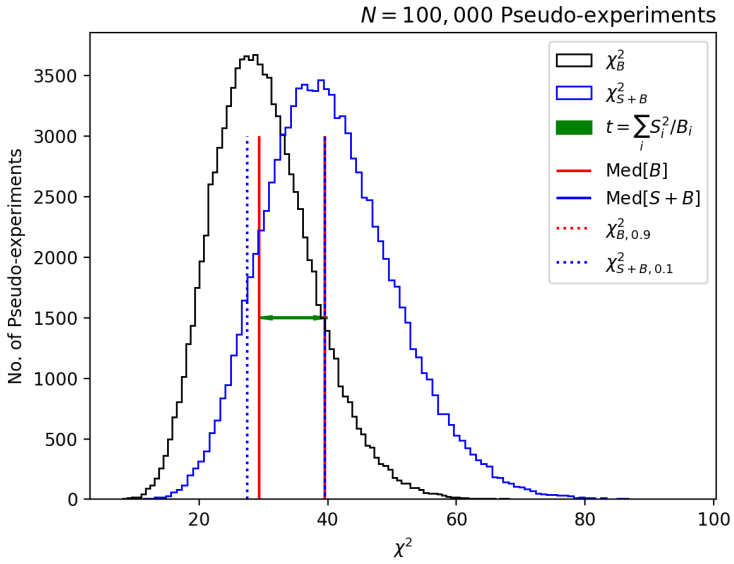

FIG. 7. Comparison between the background χ^2 and the signal-plus-background χ^2 distributions, their medians and p-value locations, and the distance t given in Eq. 13 for $\nu = 30$ bins. The $\chi^2_{B,0.9}$ and Med$[S+B]$ lines lie on top of each other. It is interesting that the requirement Median$[\chi^2_{S+B}] \to \chi^2_{B,0.9}$ is not the same as asking that Median$[\chi^2_B] \to \chi^2_{S,0.1}$; the latter would require a slightly more conservative distance measure between the two distributions. I think this is because the χ^2 distribution still has a noticeable asymmetry for $\nu = 30$ degrees of freedom; this discrepancy should vanish in the limit of large ν.