PROCEEDINGS OF SCIENCE

Can we go beyond Wilks theorem for significance calculation? Estimating p-values with importance sampling

Francisco Matorras

Instituto de Física de Cantabria, Universidad de Cantabria-CSIC,
Los Castros, Santander, Spain

E-mail: francisco.matorras@cern.ch

P-value estimation is particularly relevant in High Energy Physics for assessing the significance of potential discoveries. The reliance on Wilks' theorem, often used for p-value calculations, can be problematic for very small p-values where the asymptotic conditions may not hold. On the other hand, standard Monte Carlo simulations require prohibitively large sample sizes (~10^8). Importance sampling (IS) offers an alternative by focusing computational effort on the tails of the test statistic distribution.

The core idea of IS involves sampling from a more convenient probability density function (pdf) and applying weights to ensure that the estimated expectations converge to the true values. In this work the problem is addressed on how to choose such pdf to permit a MC estimation with attainable sample sizes.

It is shown that for models with a single parameter of interest (POI) exhibiting a monotonic relationship with the test statistic, importance sampling from the signal-plus-background (S+B) model's pdf, with the parameter value set to its maximum likelihood estimate (MLE), significantly reduces the required number of pseudo-experiments to achieve a given precision, potentially by several orders of magnitude for 5σ significance.

To address more complex scenarios, such as two-sided hypotheses or models with multiple POIs, a method based on mixtures of pdfs corresponding to different parameter values that yield similar test statistic values is proposed. While less efficient and requiring a scan of the parameter space, this approach provides a path to estimate global p-values and move beyond the limitations of local p-value approximation.

In summary, importance sampling presents a promising approach for calculating very small p-values when asymptotic methods are unreliable and direct Monte Carlo simulations are impractical. By strategically sampling from appropriately chosen pdfs and reweighting the events, this technique provides a more accurate and computationally feasible means to assess the significance of new physics signals. The methods discussed offer solutions for both simple and more complex models, paving the way for more robust statistical inference in high-energy physics analyses.

The XVIth Quark Confinement and the Hadron Spectrum Conference (QCHSC24)
19-24 August, 2024
Cairns Convention Centre, Cairns, Queensland, Australia

ISSN 1824-8039 . Published by SISSA Medialab.

PoS(QCHSC24)253

PROCEEDINGS OF SCIENCE

1. Motivation

It is well known that the convention in High Energy Physics to claim a discovery is based on the 5σ criterion. Sometimes having lengthy discussions on whether the 5σ threshold is reached or rather the significance is *only* 4.9, ignoring that usually the calculation relies on p-value calculations based on Wilks' theorem [1], in situations where the conditions for the theorem to be valid might not always be met, particularly when calculating very small p-values around 10^{-7}. In such cases, the asymptotic regime may not be reliable. The alternative, running a large number of Monte Carlo (MC) simulations, typically $O(10^8)$, can be computationally impractical.

This motivates the use of techniques like importance sampling (IS), which focuses on generating Monte Carlo datasets in the tails of the test statistic distribution, the most important region for precise p-value estimation. For example [2], says: *"The procedure above may be improved by resorting to techniques such as importance sampling which concentrates on generating Monte Carlo datasets that lie in those tails.",* but there is no indication on how to generate datasets in such a way.

In this document a method is proposed to sample in such a way that these tails are populated and efficiently calculate p-values as small as desired with only a handful of events.

This procedure provides also a solution to the problem known as the Look-Elsewhere Effect (LEE) where *local* p-values are calculated fixing some of the model parameters to avoid Wilks' conditions breaking and *global* p-values are got from approximate calculations.

2.Importance sampling

The basic idea behind importance sampling is to sample from a more convenient probability density function (pdf) and then reweight the samples so that the expectation values asymptotically converge to the desired value. If done correctly, this can lead to faster convergence, requiring fewer simulations.

Mathematically, if we want to estimate the expectation h of an observable $H(\vec{x})$ under a pdf $\rho(\vec{x})$ we calculate $h = E[H(\vec{x})] = \int H(\vec{x})\,\rho(\vec{x})d\vec{x}$ or with sampling $h = \sum H(\vec{x_i})$ with $\vec{x_i}$ drawn from $\rho(\vec{x})$. With importance sampling, we use an alternative pdf $\tilde{\rho}(\vec{x})$ and reweight:

$$h = \int H(\vec{x})\frac{\rho(\vec{x})}{\tilde{\rho}(\vec{x})}\tilde{\rho}(\vec{x})d\vec{x} = \int H(\vec{x})W(\vec{x})\tilde{\rho}(\vec{x})d\vec{x} = E_{\sim}[H(\vec{x})W(\vec{x})],$$

where we average or integrate using a weight $W(\vec{x}) = \frac{\rho(\vec{x})}{\tilde{\rho}(\vec{x})}$. It is important to note that if calculated with sampling, the method proposes to generate events from $\tilde{\rho}(\vec{x})$ and estimate $h = \sum W(\vec{x}_i)H(\vec{x}_i)$, what we commonly know as *weighted events*. The method guarantees asymptotically correct results for any $\tilde{\rho}(\vec{x})$ provided some regularity conditions, but not all choices improve the sampling efficiency. An optimal sampling pdf [3]

ISSN 1824-8039 . Published by SISSA Medialab.

PoS(QCHSC24)253

$$\rho^*(\vec{x}) = \frac{H(\vec{x})\rho(\vec{x})}{\int H(\vec{x})\rho(\vec{x})d\vec{x}}$$

can be derived, but it is impractical as it depends on the quantity we want to estimate. Alternatively, one can use a family of pdfs $\tilde{\rho}(\vec{x},\vec{\alpha})$ with the parameter(s) $\vec{\alpha}$ chosen such that they minimize the variance or maximize the cross-entropy. This will not guarantee a global best solution but can guarantee a solution that improves if the family is sufficiently broad.

3.Importance sampling in discoveries

Let's now pose our p-value problem in the context of discoveries as an importance sampling problem. Our null hypothesis H_0, corresponding to background-only (B), driven by a pdf $\rho_B(\vec{x})$ and an alternative hypothesis H_1, corresponding to signal + background (S+B), driven by $\rho'(\vec{x})$, usually in a parametric way and such H_0 is contained $\rho'(\vec{x}) = \rho(\vec{x}, |\vec{\alpha})$. Often it is only dependent on a signal strength μ and $\rho'(\vec{x}) = \rho(\vec{x}|\mu)$ such that $\rho_B(\vec{x}) = \rho(\vec{x}|\ \mu = 0)$ represents H_0. A test statistic based on the likelihood ratio $q(\vec{x}) = -2\log\left(\frac{\rho(\vec{x}|\mu=0)}{\rho(\vec{x}|\mu=\mu_0)}\right)$ can be defined, with μ_0 representing the best fit to our data. Given an observed data $\vec{x}_0$ with $\mathrm{q}_0 = q(\vec{x}_0)$ the p-value is defined as $p = \int_{q(\vec{x})>q_0} \rho(\vec{x})d\vec{x}$ and discovery is claimed when p is sufficiently small.

This p-value can be expressed using the Heaviside step function $\theta(\mathrm{q}(\vec{x}) - q_0)$

$p = \int_{q(\vec{x})>q_0} \rho(\vec{x})d\vec{x} = \int_{\vec{x}} \theta(\mathrm{q}(\vec{x}) - q_0)\rho(\vec{x})d\vec{x}$ (note that in the first case the integral is over the possible sets, pseudo-experiments, that fulfill $q(\vec{x}) > q_0$ while the second is integrated over the whole space).

The p-value can be obtained as the expectation of the observable $\mathrm{H}(\vec{x}) = \theta(\mathrm{q}(\vec{x}) - q_0)$, and hence we can transform our p-value calculation into an expectation calculation which we can deal with importance sampling.

4. p-value calculation with importance sampling

It is interesting to note that the weights take the form of a likelihood ratio, and the variance minimization or cross entropy can take a form closely resembling that of a maximum likelihood fit. In fact, using the definition of [3] for the cross-entropy, our observable $\theta(\mathrm{q}(\vec{x}) - q_0)$ and using as sampling pdf our S+B model $\rho(\vec{x}|\mu)$, it can be seen that the optimal importance sampling pdf is obtained when

$$D_{CE} = \int_{q(\vec{x})>q_0} \rho(\vec{x}|\mu = 0)\log\left(\frac{\rho(\vec{x}|\mu=0)}{\rho(\vec{x}|\mu)}\right)d\vec{x}$$

is minimal (note this is just a function of μ)
Choosing the S+B model is particularly convenient, as the model is either analytically known or obtainable through simulation. Moreover, given the similarity between variance minimization and likelihood estimation, one might expect the optimal pdf for importance sampling to coincide with the best fit to our data. However, as will be shown later, this is not generally the case, though it provides a useful handle for solving our problem.

5. Single- and one-sided parameter of interest

Let's start with the common case where our S+B model depends on a single parameter and it is one sided, a signal strength $\mu \geq 0$. Let's also assume that dependence is monotonic, in the sense that larger μ imply smaller p-values.

Our experiment consists of a set of N measurements $\vec{x}_0$ that should follow a law $\rho(\vec{x}|\,\mu = 0)$ in the absence of signal, or $\rho(\vec{x}|\,\mu)$ if signal is present. Often each measurement is independent, and the pdf can be factorized as product of event pdf's which simplifies the practical applications, although this assumption is not needed for the following discussion. When we perform our maximum likelihood estimation to our data, a value of $\mu = \mu_0$ is obtained and one wants to establish if it is large enough to reject the background-only hypothesis or not.

It can be shown that under some general conditions the optimal sampling pdf is precisely $\rho(\vec{x}|\mu = \mu_0)$, in other words, the optimal sampling is obtained if we sample from the S+B distribution with a signal strength equivalent to what we observe in the data. This selection provides improvements of several orders of magnitude on the required number of pseudo-experiments needed to estimate the p-value. It is also found that the improvement is large for a wide range of values around μ_0 which means that even if the sampling is not done with the optimal pdf, it is still unbiased and significantly better than direct unweighted sampling. In practical cases, this implies the possibility to use already existing MC simulation with the nearest value to μ_0.

One would proceed according to the following algorithm:

1. Fit the **data** to H_1, S+B, model and get μ_0
2. Generate a handful (M) of pseudo-experiments according to this model
   - $\{\vec{x}\}_j \sim \rho(\vec{x},\, \mu_0)$
   - Or sample from the available full MC sample closer to μ_0
3. Fit each set $\{\vec{x}\}_j$ get μ_j and q_j (repeat the full analysis on this pseudodata)
4. Calculate the weights of **each pseudo-experiment** $w_j = \frac{\rho(\vec{x}^j, \mu=0)}{\rho(\vec{x}^j,\, \mu=\mu_0)}$
   - if independent, ρ is factorized and become products of N **event** weights
5. Calculate p as $\frac{1}{M}\sum_{q_j>q_0} w_j$

As illustration, this procedure is applied to the simple example of a binned Poisson fit to a low statistics histogram where the mean of each bin is given by $\lambda_i = b_i + \mu s_i$ with μ positive and for small number of counts. To highlight the power of the method, the p-value is sampled using only 100 pseudo-experiments (each pseudo-experiment being a histogram). The results are compared to unweighted sampling of 10^5 pseudo-experiments and to the asymptotic calculation according to Wilks theorem. The estimated p-values as a function of q, for different amounts of signal are shown in Fig. 1.

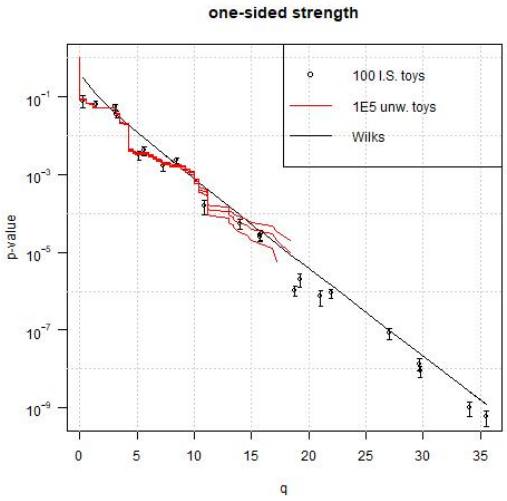

Figure 1: Estimated p-values for a binned Poisson case according to the method proposed in this work (points), unweighted toys (red line) or Wilks (black line). Error bars and bands show the statistical uncertainty on the sampling-estimated quantities.

We can appreciate that the importance sampling, perfectly reproduces the wiggles induced by the discrete nature of Poisson statistics even going to very small p-values and despite using very few simulations. Incidentally, it can be observed that the Wilks-theorem-based approximation is not extremely good for this case.

A test was performed to determine how much the choice of the μ affects the results. Fig. 2 shows the uncertainty obtained with importance sampling for different values of μ, relative to that of unweighted sampling when estimating the p-value of a 5σ excess. It can be seen that the improvement for the optimal μ is about a factor 1000, which means a factor 1000000 in number of simulated pseudo-experiments. It can also be seen that this improvement is still nearly valid for a wide range of μ, showing that the method is robust.

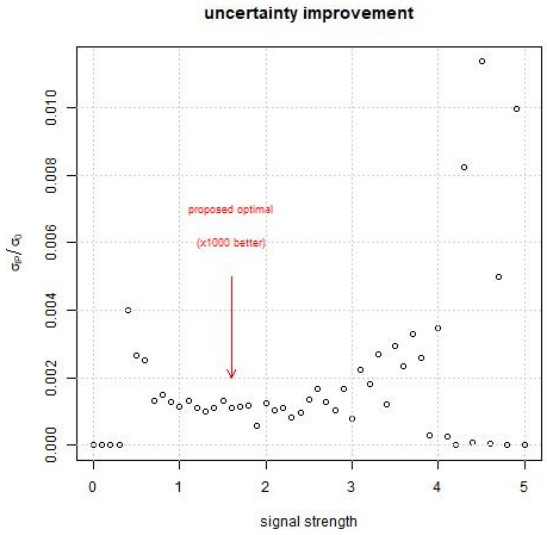

Figure 2: Relative uncertainty in the p-value estimation through sampling comparing importance or unweighted sampling for a 5σ excess as a function of the sampling μ

6.More general cases

Given the success and simplicity of the method, one can be tempted to generalize and assume it works for any situation, but unfortunately that is not the case. Fig. 3 illustrates the problem, when the dependence of q with μ is not monotonic, for example if we permit μ to be negative and we can have new physics both from an excess or a defect of data.

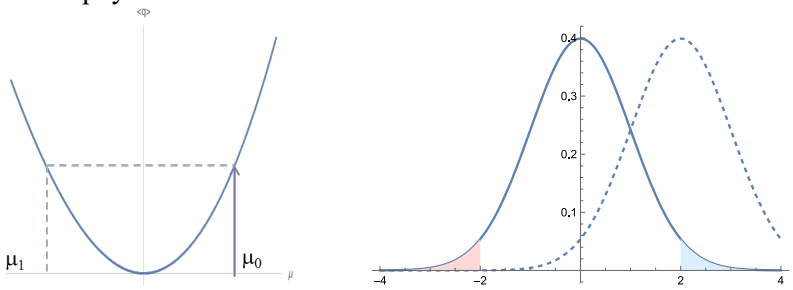

Figure 3: Schematic view of the situation for a two-sided problem

Suppose that the data sees an excess compatible with a μ_0, if we follow the previous method, we would sample from the pdf represented by the dotted line on the right-hand side figure, that will populate the positive-side (blue) tail, but not the negative-side tail (pink). Depending on the number of simulations, the results will be biased (only blue tail) or with a huge variance (pink tail estimated with very few events with huge weights). To properly account for both tails, events should also be sampled from the similarly unlike case of a negative μ_1.

To overcome this problem the proposed solution is to sample from an admixture
$\tilde{\rho} = \frac{1}{2}\rho(\vec{x}|\mu = \mu_0) + \frac{1}{2}\rho(\vec{x}|\mu = \mu_1)$,
where μ_0 is the best fit from the data and μ_1 is the alternative signal strength giving the equivalent $< q >$. This procedure can be proven to work, although it requires twice as many simulations and a scan to obtain μ_1.

In this case, the algorithm would turn into:

1. Fit the data to H_1, S+B, model and get q_0 and μ_0
2. Scan μ values: for each, generate pseudo-experiments $\{\vec{x}\}_j$, $\sim\rho(\vec{x}, \mu)$ run the analysis, get q_j and calculate the average. It can be done for the available MC points.
3. From <q> as a function of μ get the two values μ_0 and μ_1 in your scan closer to q_0
4. Generate M pseudo-experiments $\{\vec{x}\}_j$, $\sim\frac{1}{2}\rho(\vec{x}^j, \mu=\mu_0)+\frac{1}{2}\rho(\vec{x}^j, \mu=\mu_1)$ or a combination of the **two** closer full MC samples
5. Fit $\{\vec{x}\}_j$ get μ_j and q_j (repeat the full analysis on this pseudo-data)
6. Calculate weights $w_j = \frac{\rho(\vec{x}^j,\mu=0)}{\frac{1}{2}\rho(\vec{x}^j,\mu=\mu_0)+\frac{1}{2}\rho(\vec{x}^j,\mu=\mu_1)}$
7. Calculate p as $\frac{1}{M}\sum_{q_j>q_0} w_j$

As an extreme example, this procedure is applied to the same low statistics histogram, but in this case permitting μ to be negative (requiring λ to be nonnegative $\lambda_i = \max(b_i + \mu s_i, 0)$). Figure 4 (left) shows the average q as a function of the signal strength for this extreme example as well as the results of the p-value calculation (right) with 200 importance sampling simulations compared with those with one million unweighted simulations. Again, we can see that the method reproduced well the true value despite the very few simulations and that Wilks' approximation does not reproduce it extremely well.

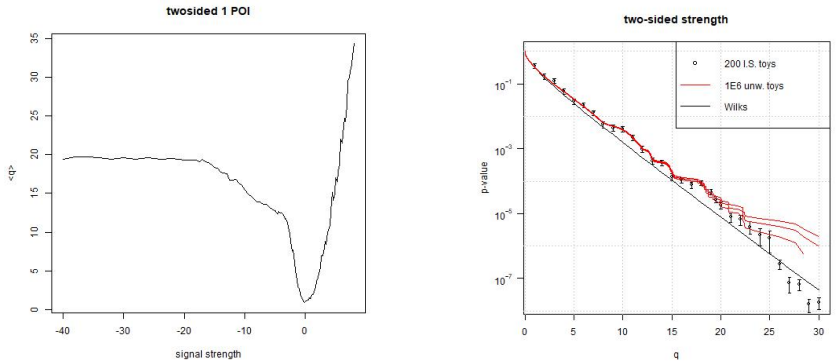

Figure 4 Average value of q as a function of the signal strength (left) and p-value estimation as a function of q (right) for an example with a quadratic dependence of q on μ.

This idea can be extended to more complex situations with many crossings, many signal strengths giving the same significance, using admixtures $\frac{1}{k}\sum\rho(\vec{x} \mid \mu=\mu_i)$ summing over all μ_i which provide the same q. An example for a quartic dependence is shown in Fig 5.

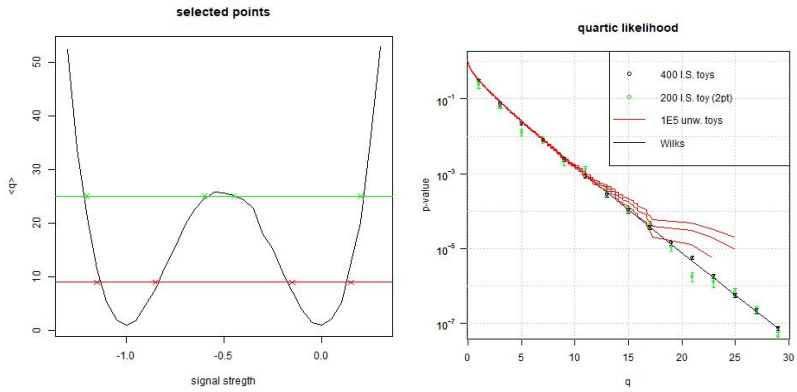

Figure 5 Average value of q as a function of the signal strength (left) and p-value estimation as a function of q (right) for an example with a quartic dependence of q on μ.

7. Application to cases with two or more POI

The immediate question is whether this method can be extended to a higher dimension problem, with two or more parameters of interest. After the previous discussion one can see that the extrapolation is not trivial, because we would need to populate n-dimensional surfaces corresponding to combination of POI giving a similar significance as the observation we want to stress. But this discussion also gives us a handle to approach the problem.

Ideally a continuous admixture of pdf could be used, but it is impractical because it will imply integrals without a closed form except for trivial examples. However, it is feasible if instead we discretize and use the admixture of *some* points as described following.

The proposed solution is to perform a grid scan of the parameters $\vec{\alpha}$ obtaining the average q as a function of $\vec{\alpha}$. From the output of this scan, a *few* points $\vec{\alpha}_i$ in the POI space compatible with q_0 can be selected. As before, sampling pdf an admixture of these $\frac{1}{n}\sum_i \rho(\vec{x}^j, \vec{\alpha} = \vec{\alpha}_i)$ can be used. There is no general rule to define how many points are needed, but examples show the results are stable once a threshold is crossed.

The method is illustrated with the following example. A signal modeled with a gaussian with free normalization and mass is searched on top of an exponential background. Fig 6 (left) shows the grid, the values of μ and mass, Fig 6 (center) shows the contours that correspond to a given value of q and Fig 6 (right) an example of $\vec{\alpha}_i$ corresponding to the requested q_0. The sampling is then performed as before.

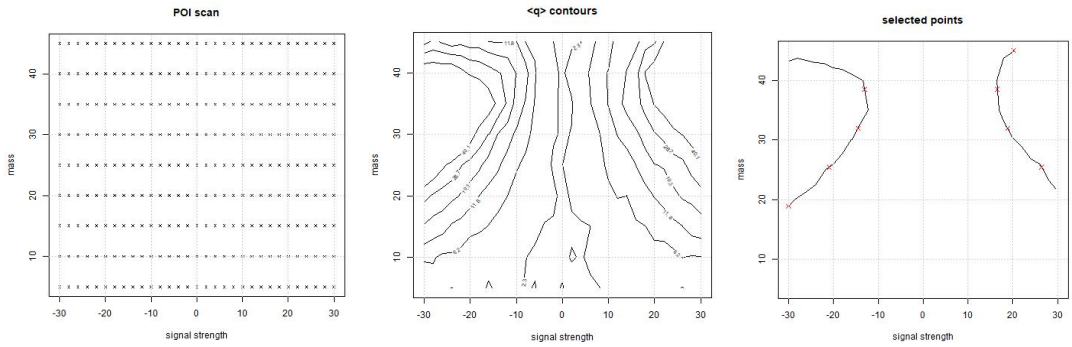

Figure 6. Example of application to a 2 POI problem: parameter scan (left), constant <q> contours (center) and selected points at q_0 (right).

Figure 7 summarizes the results of the p-value calculation for different selected numbers of points in which the contour is split. It can be seen that 5 points equally distributed seem to be sufficient. The method becomes less efficient when going to 2 dimensions but still can provide unbiased and relatively precise estimations with just 1000 simulations, even for small p-values.

An interesting conclusion from this particular example can be drawn from the comparison of these p-values with those predicted by the Wilks theorem assuming one degree of freedom, what is commonly used for the "local" p-value, or two degrees of freedom, assuming full application of the conditions of the theorem. The local p-value underestimates the true p-value by about a factor of ten and even using 2 dof underestimates by 50%. The method proposed here permits a correct calculation of the global p-value with attainable number of simulations.

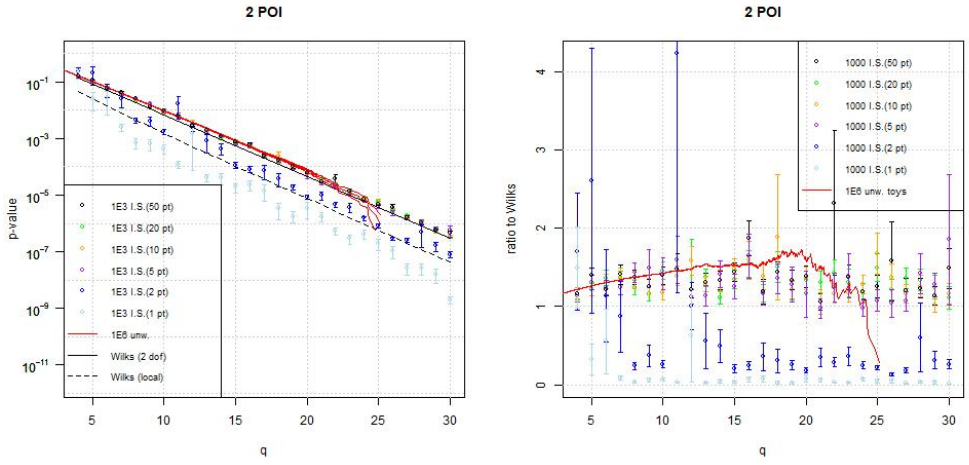

Figure 7: p-values obtained for the 2 POI example with different numbers of points along the contour compared with unweighted sampling estimation and asymptotic calculations for 1 dof (local) or 2 dof. Right-hand side figure shows the same information, but the p-value is shown relative to the Wilks-based calculation with 2 dof.

8. Conclusions

In this talk, a method was presented to estimate as small as desired p-values based on important sampling.

It is shown that for models with a single parameter of interest (POI) exhibiting a monotonic relationship with the test statistic, importance sampling from the signal+background model's pdf, with the parameter value set to its maximum likelihood estimate, significantly reduces the required number of pseudo-experiments to achieve a given precision, by several orders of magnitude for 5σ significance.

To address more complex scenarios, such as two-sided hypotheses or models with multiple POIs, a method based on mixtures of pdfs corresponding to different parameter values that yield similar test statistic values is proposed. While less efficient and requiring a scan of the parameter space, this approach provides a path to estimate global p-values and move beyond the limitations of local p-value approximation.

In summary, importance sampling presents a promising approach for calculating very small p-values when asymptotic methods are unreliable and direct Monte Carlo simulations are impractical. By strategically sampling from appropriately chosen pdfs and reweighting the events, this technique provides a more accurate and computationally feasible means to assess the significance of new physics signals. The methods discussed offer solutions for both simple and more complex models, paving the way for more robust statistical inference in high-energy physics analyses.

References

[1] Wilks, S. S. (1938). *The large-sample distribution of the likelihood ratio for testing composite hypotheses*. Annals of Mathematical Statistics, 9(1), 60–62.

[2] Behnke, O., Kröninger, K., Schott, G., & Schörner-Sadenius, T. (Eds.). (2013). *Data Analysis in High Energy Physics: A Practical Guide to Statistical Methods*. Wiley-VCH.

[3] Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo Method* (3rd ed.). Wiley.

[4] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. 54, No. 3 (Nov., 1943), pp. 426-482.

[5] Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). *Asymptotic formulae for likelihood-based tests of new physics.* European Physical Journal C, 71, 1554. DOI:10.1140/epjc/s10052-011-1554-0