PROCEEDINGS OF SCIENCE

# Open Questions in Statistical Practice for Particle Physics

**Francisco Matorras**

*IFCA, Instituto de Física de Cantabria, Universidad de Cantabria-CSIC, Santander, Spain*

E-mail: francisco.matorras@cern.ch

Statistical methods play a crucial role in the data analysis and interpretation in particle physics experiments. As we deepen in the theories and build more complex experiments, new challenges arise in statistical practice. In this talk, we explore some of the open questions, new and old, and ongoing debates in statistical methodology as applied to particle physics research.

Hypothesis testing is one of the basic grounds in particle physics research: when claiming a discovery, when confirming a given theoretical model or maybe simply to convince ourselves that our detector simulation properly reproduces data. Despite being a rather basic and classical concept in statistics, common practices are not always fully aligned with statistical theory.  We are still asking ourselves why requiring "5-sigma" to present an excess as a discovery and what does it really imply. We started to get used to jargon like global and local significance or "look-elsewhere effect", but what do they really mean?

Optimal combination of results from different experiments is also a complex task when accounting for correlations and systematics. A crucial part of this problem lies in the amount of information that should be made publicly available to permit an optimal combination or a correct and efficient contrast with theory.

It becomes more and more frequent the use of unfolding to correct experimental results for detector effects and provide a direct way of comparison with theory, yet its application comes with inherent limitations and complexities. Challenges arise from the interplay of detector effects, statistical fluctuations, and model assumptions, leading to uncertainties in the unfolded results. Additionally, unfolding methods must contend with the non-trivial task of balancing between bias and variance, often requiring sophisticated algorithms and careful validation procedures.

The explosion of machine learning provides at the same time powerful new techniques but also poses new problems in its interaction with classical statistics. We are familiar with its use in optimal classification, but more and more applications are appearing ranging from alternatives to Monte Carlo simulation to detector effects unfolding, among others.

Through an examination of these open questions and challenges, this talk aims to stimulate discussion and bring awareness of the assumptions and approximation lying behind the different methods.

*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.

https://pos.sissa.it/

## 1. Introduction

Statistics forms an indispensable core of particle physics, having been fundamental for many years. It serves as the essential tool for processing often massive datasets to extract meaningful information. This includes determining crucial observables such as cross sections and particle masses, along with their associated uncertainties, ensuring proper statistical interpretation (e.g., whether a quoted error bar truly represents a 68% confidence interval). Statistics is also vital for setting limits on New Physics (NP) models, defining what a 90% exclusion means, and, most notably, for establishing discoveries based on the stringent 5σ criterion. The field is characterized by a constant evolution and application of increasingly powerful statistical techniques, driven by the desire to maximize the information extracted from experimental data. This pursuit can lead to significant benefits, such as potential cost savings through shorter experiment run times and the ability to probe further away physics. Furthermore, statistically-sound practices are crucial in avoiding embarrassing announcements based on spurious findings.

The statistical landscape in High Energy Physics (HEP) is rich and often presents non-trivial challenges. Currently, numerous new statistical techniques are being explored alongside a deeper understanding of established methodologies. It is noteworthy that physicists often tend to reinvent existing statistical methods and may lack awareness of recent (or even older but useful) advancements. Recognizing this, there is an ongoing effort to foster collaboration between physicists and statisticians. The PHYSTAT initiative [1], founded in 2000 by L. Lyons, serves as a key platform for seminars, conferences, and workshops dedicated to debating relevant statistical issues.

This paper will concentrate on a few highlighted areas within statistical practice relevant to particle physics discoveries, aiming to provide insights into the complexities behind discovery statements, data reinterpretation, and the interplay between machine learning and statistics.

## 2. Discovery, why 5σ?

It is well known that in HEP we have a (historical) convention that a discovery cannot be claimed unless an excess is found equivalent to at least 5σ. But do we really know what it means?

From a statistical perspective, a discovery typically arises from a hypothesis test where experimental data is confronted with a baseline model (the null hypothesis in statistician words, $H_0$) against an alternative hypothesis, $H_1$. The null hypothesis usually represents the Standard Model (SM), the SM without a specific process, or simply background-only expectations. Conversely, the alternative hypothesis often represents our model incorporating "new" physics, frequently depending on one or more parameters. Alternatively, we can reverse this logic, using a NP model as $H_0$ to set limits on its parameters.

The usual technique for quantifying the hypothesis test is the likelihood ratio, which built a test statistic defined as the ratio $q = -2\log(\mathcal{L}(H_0)/\mathcal{L}(H_1))$ where the likelihoods are evaluated at their best-fit values under each hypothesis. A small value of q indicates that the data prefers $H_0$, while a large q suggests evidence against $H_0$. To assess the significance of the observed data, we calculate the p-value: the probability, given $H_0$ is true, of observing a fluctuation that yields a test

ISSN 1824-8039 . Published by SISSA Medialab.
https://pos.sissa.it/

statistic q greater than or equal to the value $q_0$ obtained from our data. This p-value is then commonly translated into a significance, or z-score, representing the equivalent number of standard deviations $\sigma$ in a Gaussian distribution.

In HEP, there is a well-established, albeit historical, convention that a discovery cannot be claimed unless the observed excess reaches a significance of at least 5σ (see for example the discussion in [2]). This threshold corresponds to a p-value of approximately 3 ·, meaning the probability of observing such an extreme event due to a background fluctuation is less than three in ten million. Interestingly, statisticians typically consider a significance of *only* 3σ (corresponding to a p-value of 0.001) sufficient to guard against random fluctuations. This raises the question of whether the intense effort required to push a 4.9σ result to 5σ is always justified. The author suggests it might be naïve to strive for such precision when systematic uncertainties and the validity of asymptotic approximations at that level cannot be confidently controlled.

The reasons behind the 5σ convention are likely historical. One primary motivation was to avoid embarrassing mistakes stemming from claimed discoveries with apparently large significance that subsequently faded away. Indeed, even some 5σ anomalies have been known to disappear over time. Another contributing factor might be a lack of complete confidence in the evaluation of systematic uncertainties, with the 5σ threshold providing a perceived safety margin (e.g., remaining above 3σ even if systematics were largely underestimated). However, this rationale becomes less compelling when an analysis is statistically dominated. Strictly speaking we want to avoid a false discovery arising from a background fluctuation, but the question arises whether reaching such an extremely low probability of $10^{-7}$ is always necessary. The 5σ criterion can also be seen as a safety margin to account for potentially missed systematic uncertainties.

L. Lyons [2] introduced the concept of "plausibility”, arguing that the same significance level should not be universally applied to all potential discoveries. For instance, the required significance for the Higgs discovery might differ from that needed for claiming evidence of a rare decay of the Z, faster-than-light neutrinos, or even an anomaly not predicted by any existing model. In some cases, 3σ might be sufficient, while in others, even 5σ might not be enough, warranting further scrutiny beyond purely statistical considerations. Ideally, one should define a case-dependent threshold for discovery, although implementing this in practice is a challenging task.

## 3.The Look Elsewhere Effect (LEE) and Local p-values

Another argument used to justify the 5σ threshold is related to what in HEP has been named as “look elsewhere effect” and in statistics is related to the concept of multiple testing.We can illustrate the problem with a simple example. Consider an analysis looking in a histogram for an excess of events over a background prediction. If a significant excess is observed in one particular bin, a *local* p-value can be calculated, indicating the probability of such a fluctuation occurring in that specific bin under the background-only hypothesis. However, one must consider that an equally surprising fluctuation could have happened in any other bin of the histogram. Therefore, the probability of observing one such a fluctuation *anywhere* in the spectrum is larger than the local p-value, in good approximation the sum of all the probabilities. This effect becomes even more pronounced if the properties of a potential signal (e.g., its width) are not precisely known beforehand. The excess could be reflected in smaller fluctuations in two or more nearby bins. One should account for the probability of having such a fluctuation for the range of bins compatible with the allowed width as defined by the physics model. Or even more complex patterns like oscillating excesses. The key point is that all relevant possibilities need to be accounted for when calculating the *global* p-value. The "elsewhere" can depend on the model parameters under investigation, accounting for various aspects like any bin in a spectrum, any possible width of a signal, or any location within the analyzed data.

The procedure at the Large Hadron Collider (LHC) often involves calculating a *local* p-value for a single free parameter (typically the signal strength) as a function of other parameters that are held fixed (e.g., mass and possibly width). The smallest p-value obtained from this scan across the fixed parameters is then quoted as the *local* p-value. Importantly, this *local* p-value is calculated assuming only one free parameter, even if the underlying model might have more. A good example can be found in the Higgs boson discovery by the Atlas [3] and CMS [4], as shown for CMS in Figure 1. The p-value was calculated for different hypotheses for the Higgs boson mass and given it went below 5σ at a mass near 125 GeV, a discovery was claimed.

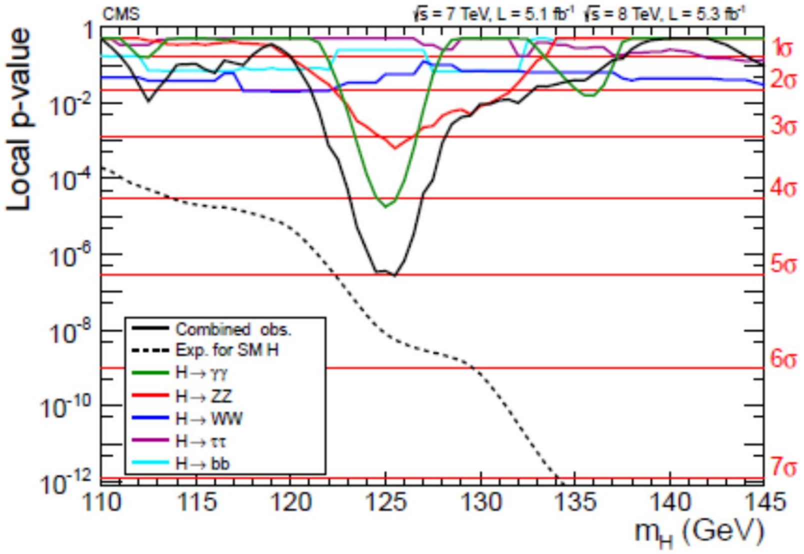

Figure 1: Local p-values as a function of the potential Higgs boson mass supporting its discovery by CMS [4]

A decision on whether an excess constitutes a discovery is then **often made based on this local p-value**. A global p-value, which attempts to account for the LEE, might be calculated eventually based on an estimation of the number of independent "elsewheres" considered.

However, this approach is argued to be unfair because not all cases exhibit the same LEE. As an extreme example, a 5σ local excess could potentially arise from a mere 1σ global fluctuation if one considers a very loose model with a large number of free parameters (e.g., 20).

Therefore, it is advocated that the field **should move towards the use of global p-values** to provide a more accurate quantification of the statistical significance. This shift towards global p-values should ideally be coupled with a re-evaluation of the strict 5σ discovery requirement. While experiments are understandably reluctant to lower their discovery threshold, it is argued that it is inconsistent to consider a result with a 5σ local and 4σ global significance a discovery while rejecting one with a 4.5σ global significance.

Beyond the studied spectrum, there will always be unpredictable "elsewhere” (different spectrum studied, different NP searched…), suggesting that some safety margin remains prudent. **The author's personal opinion is that global p-values should be adopted, and a 4σ global significance might be sufficient for discovery, with a greater emphasis placed on thoroughly scrutinizing systematic errors rather than solely focusing on achieving precisely a given significance.**

## 4. Discoveries from measurements, combinations and reinterpretation

Another path to discovery originates from measurements. It is common to encounter statements like, "A measurement yields $x = 5 \pm 1$, the SM predicts $x = 0$ therefore a 5σ deviation is observed, and a discovery was made". However, this direct interpretation is not entirely accurate. While such a result is certainly enticing, the quoted uncertainties are calculated based on underlying assumptions that may not hold true at the 5σ level. These assumptions often include Gaussian behavior of uncertainties, the validity of linear error propagation, and the adequate assessment of systematic uncertainties, particularly in the tails of the distributions. **Such a measurement should ideally be translated into a proper hypothesis test to rigorously assess the significance of the deviation from the SM prediction**.

The situation becomes even more complex and potentially problematic when results from multiple measurements are combined. It is strongly cautioned against confronting a theoretical model with an ad-hoc combination of existing experimental measurements. This is because experimental results are often correlated, both within a single experiment and between different experiments. A naïve combination that fails to account for these correlations will almost invariably lead to an underestimation of the true uncertainties. Even the combinations provided by the Particle Data Group (PDG), while significantly better, may still be incomplete for the purpose of precise theory testing. Best Linear Unbiased Estimator (BLUE) [5] combinations represent a more sophisticated approach, as they account for correlations to some extent. However, even BLUE combinations typically rely on Gaussian and linear assumptions.

Recognizing these limitations, LHC experiments and others are increasingly adopting a more rigorous approach to combining data. The current trend is to combine the underlying data from different analyses rather than simply combining the final results. This involves constructing a global likelihood function that incorporates all relevant datasets, includes systematic uncertainties (often referred to as "nuisances"), and accounts for their known correlations to the best of the experiments' knowledge. Performing a single global fit to this combined likelihood yields the most precise overall measurement. Combining results across different experiments presents even greater challenges. This becomes particularly difficult for theorists who wish to interpret experimental data in the context of their own physics models. A key challenge is how to test a theoretical model against a published cross-section measurement properly, taking into account all the relevant systematic uncertainties and their correlations.

The LHC experiments are aware of the limitations associated with combining results and the need for more transparent data presentation. This has led to initiatives aimed at facilitating the reinterpretation of their data for a wider range of theoretical models. HEPDATA serves as a crucial resource in this regard, publishing the most relevant information from particle physics experiments in the form of over 14000 data tables.

Furthermore, there have been attempts to make simplified versions of the likelihood functions publicly available. The goal of these simplified likelihoods is to allow anyone to easily plug in their own theoretical predictions or additional datasets for comparison. A more ambitious approach involves publishing the complete statistical model used by the experiment, including the full set of data, the nuisance parameters representing systematic uncertainties, the underlying statistical model, and the associated software tools. Some results are already being made available in this comprehensive format.

These efforts enable several important activities in data reinterpretation [6-11]:

- Updating existing analyses using more precise theoretical calculations, improved experimental calibrations, or different probability models.
- Kinematic reinterpretation of data by considering different physical processes with potentially different phase space distributions and detector efficiencies.
- Combinations of analyses or datasets within model surveys or for global averages of parameters.
- Reuse of datasets for other studies, such as the determination of parton distribution functions.

## 5. Unfolding

Unfolding is an important technique widely used in particle physics to correct experimental results for detector-induced effects like efficiency and resolution, allowing for a more direct comparison with theoretical predictions. The underlying problem is an inverse problem: to recover the original physical quantities ("truth") before they are distorted by the detector's resolution and efficiency effects as sketched in Fig. 2.

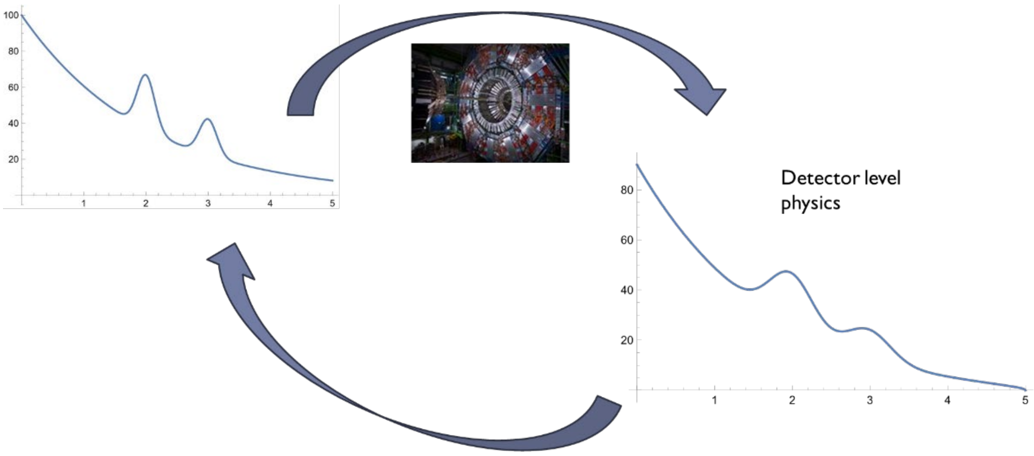

Figure 2: Schematic representation of unfolding: we have a physics process with a certain spectrum (top left), but the detection process distorts it (bottom right) and we look for a procedure to recover the original one.

A simplified approach involves discretizing both the true underlying physics, and the detector-level measurements into histograms, $t_j$, $d_i$. Simulations are then used to calculate a transfer (or migration) matrix $R_{ij}$, which describes the probability that a true event in bin $j$ will be reconstructed in detector-level bin $i$. Ideally, one would expect the measured data to be related to the true distribution by $\vec{d} \leftarrow R\vec{t}$, and one might naively attempt to invert this relationship to obtain the true distribution $\vec{t} = R^{-1}\vec{d}$ or more generally obtain the best $\vec{t}$ with a Maximum Likelihood Estimation.

However, this naïve maximum likelihood approach is generally ill-posed and prone to instabilities that introduce high-frequency artifacts in the unfolded distribution. To overcome this, several advanced techniques are employed, including iterative methods, Tikhonov regularization, neural networks, and wide/narrow binning adjustment. These methods often involve incorporating regularization, which implies introducing additional information to suppress unwanted fluctuations.

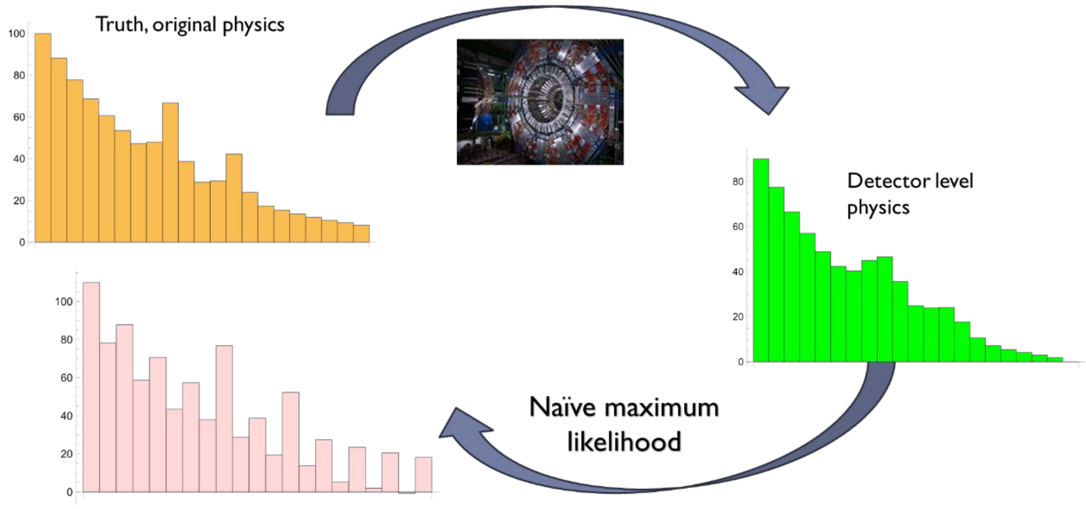

Figure 3: Schematic view of naïve maximum-likelihood approach to unfolding of discrete data (histograms), high frequency terms appear not present in the "truth".

The challenge lies in optimizing the trade-off between the bias induced by the regularization and the uncertainty introduced by the inversion. Users of unfolded results must be aware that a bias will exist (an additional systematic uncertainty) and that there can be significant correlations between the unfolded data points. These potential correlations also need to be considered when combining unfolded results with other measurements.

More details can be found in Mikael Kuusela´s talk [12].

## 6. Machine Learning

Machine Learning (ML) techniques have become increasingly present in HEP over the last two decades, experiencing a significant surge in applications recently, similar to many other scientific fields. Initially, ML was primarily used for classification tasks and, to a lesser extent, for regression. Now, its applications have expanded dramatically to include particle identification and calibration, anomaly detection, unfolding and other inverse problems, simulation, density estimation, detector optimization, reweighting Monte Carlo samples, and even in theoretical calculations like parameter tuning and lattice QCD. A wide array of ML techniques is employed, such as Deep Neural Networks (DNN), Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), and Graph Neural Networks (GNN). This represents a whole new active area of research within particle physics whose description is beyond the scope of this talk.

However, the application of ML in HEP also presents several statistics-related challenges:

- Overfitting and Generalization: Ensuring that ML models learn genuine patterns in the data rather than noise, and that they generalize well to unseen data. This also raises the question of whether the ML model itself introduces a systematic uncertainty.

- Modeling Uncertainty Quantification: Understanding and quantifying potential biases introduced by ML algorithms
- Bias and Systematic Errors: how to properly incorporate systematic errors. Traditional methods often calculate systematics one at a time, while ML derives its power from the combination of several variables. Promising methods are proposed to mitigate systematic errors.
- Interpretability and Explainability: Understanding the decision-making process of complex ML models is crucial for identifying potential issues and for understanding the origin of any discovered signals.
- Handling Imbalanced Datasets: Developing effective techniques for scenarios where the signal being searched for is very rare compared to the background.
- …

## 7. Summary and conclusions

In summary, statistics plays a vital and evolving role in particle physics. It is essential for extracting meaningful results from experimental data and for making claims of new discoveries.

The historical 5σ convention for discovery is a subject of ongoing debate, with arguments being made for its revision. The author advocates for a shift towards the decision based on global p-values replacing the local p-values together with a potential relaxation of the 5σ requirement, accompanied by a stronger focus on the rigorous evaluation of the effect of systematic uncertainties.

A significant effort is underway within the LHC community to publish comprehensive datasets, and the full statistical models employed in analyses. This transparency is crucial for enabling optimal and correct public (re)interpretation of experimental results and for maximizing the scientific output of these large-scale experiments.

Finally, machine learning has become a rapidly expanding area within particle physics, offering powerful new tools for data analysis but also presenting significant statistical challenges that need to be carefully addressed.

## References

[1] PHYSTAT Conference and seminars https://phystat.github.io/Website/

[2] L. Lyons. *Five sigma revisited.* CERN Courier, 23 May 2013. Available at: https://cerncourier.com/a/five-sigma-revisited/.

[3] ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Phys. Lett. B 716, 1-29 (2012), DOI: 10.1016/j.physletb.2012.08.020, arXiv:1207.7214

[4] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B 716, 30-61 (2012), DOI: 10.1016/j.physletb.2012.08.021, arXiv:1207.7235

[5] Lyons, L., Gibaut, D., & Clifford, P. (1988). *How to combine correlated estimates of a single physical quantity.* Nuclear Instruments and Methods in Physics Research Section A, 270(1), 110–117. https://doi.org/10.1016/0168-9002(88)90018-6

[6] LHC Physics Working Group. *Interpreting LHC Results*. Available at: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/InterpretingLHCresults.

[7] L. Heinrich, G. Kasieczka, D. Shih, et al. *Les Houches Guide to Reusable Machine Learning Models in LHC Analyses*. arXiv:2312.14575 [hep-ph], 2023. Available at: https://arxiv.org/abs/2312.14575.

[8] S. Kraml, T. Andeen, J. Behnke, et al. *Snowmass White Paper on Data and Analysis Preservation, Recasting, and Reinterpretation*. arXiv:2203.10057 [hep-ph], 2022. Available at: https://arxiv.org/abs/2203.10057.

[9] K. Cranmer, G. Heinrich, G. Perez, et al. *White Paper on Publishing Statistical Models: Getting the Most Out of Particle Physics Experiments*. arXiv:2109.04981 [hep-ph], 2021. Available at: https://arxiv.org/abs/2109.04981.

[10] A. Abdallah, J. Andrea, S. Banerjee, et al. *Reinterpretation of LHC Results for New Physics: Status and Recommendations after Run 2*. SciPost Phys. **9**, 022 (2020). DOI: 10.21468/SciPostPhys.9.2.022.

[11] S. Kraml, T. Andeen, J. Behnke, et al. *Snowmass White Paper on Data and Analysis Preservation, Recasting, and Reinterpretation.* arXiv:2203.10057 [hep-ph], 2022. Available at: https://arxiv.org/abs/2203.10057.

[12] M. Kuusela, *Response Matrix Estimation in Unfolding Differential Cross Sections*, this conference.