Software for statistical data analysis used in Higgs searches

Christian Gumpert\textsuperscript{1,a}, Lorenzo Moneta\textsuperscript{2,b}, Kyle Cranmer\textsuperscript{3}, Sven Kreiss\textsuperscript{3} and Wouter Verkerke\textsuperscript{4}

\textsuperscript{1} IKTP, TU Dresden, Germany
\textsuperscript{2} CERN, Geneva, Switzerland
\textsuperscript{3} New York University, USA
\textsuperscript{4} NIKHEF, Netherlands

E-mail: \textsuperscript{a}cgumpert@cern.ch, \textsuperscript{b}Lorenzo.Moneta@cern.ch

Abstract. The analysis and interpretation of data collected by the Large Hadron Collider (LHC) requires advanced statistical tools in order to quantify the agreement between observation and theoretical models. \textsc{RooStats} is a project providing a statistical framework for data analysis with the focus on discoveries, confidence intervals and combination of different measurements in both Bayesian and frequentist approaches. It employs the \textsc{RooFit} data modelling language where mathematical concepts such as variables, (probability density) functions and integrals are represented as C++ objects. \textsc{RooStats} and \textsc{RooFit} rely on the persistency technology of the ROOT framework. The usage of a common data format enables the concept of digital publishing of complicated likelihood functions. The statistical tools have been developed in close collaboration with the LHC experiments to ensure their applicability to real-life use cases. Numerous physics results have been produced using the \textsc{RooStats} tools, with the discovery of the Higgs boson by the ATLAS and CMS experiments being certainly the most popular among them. We will discuss tools currently used by LHC experiments to set exclusion limits, to derive confidence intervals and to estimate discovery significances based on frequentist statistics and the asymptotic behaviour of likelihood functions. Furthermore, new developments in \textsc{RooStats} and performance optimisation necessary to cope with complex models depending on more than 1000 variables will be reviewed.

1. Introduction
The tremendous amount of data delivered by the LHC imposes new challenges on statistical data analysis. Besides the technical ability to handle large data sets efficiently, tools for statistical interpretation should implement the most up to date concepts in frequentist and Bayesian statistics. Typical tasks in statistical inference comprise

- goodness of fit evaluation,
- parameter estimation/confidence interval calculation,
- hypothesis testing.

Notwithstanding the importance of the first two statistical techniques, this article will focus on the \textsc{RooStats} [1] framework in the context of hypothesis testing as it was used in the search for the Higgs boson. The search for new physics phenomena such as those predicted by theory of Supersymmetry, as well as the search for dark matter candidates and Higgs bosons, play a crucial...
role in the current high energy physics (HEP) programme. Having a common framework for statistical data analysis that implements the standards widely accepted in the HEP community helps establishing discoveries (e.g. the discovery of a neutral boson by ATLAS and CMS last summer [2, 3]), and is also simplifies the combination and/or comparison of results between different collaborations.

2. Concept of hypothesis testing
Common tasks in HEP include the evaluation of different theoretical models with respect to their ability of describing the data observed. In statistics, this task is called hypothesis testing where one evaluates the compatibility of the model under investigation (null hypothesis) which is accepted or rejected in favour of an alternative hypothesis at a certain confidence level. In the context of searches for new physics, it is sufficient to reject the null hypothesis at the desired confidence level.

A model is defined by a likelihood function \( L = L(\mu, \vec{\theta}) \) which depends on the parameter of interest \( \mu \) and a set of additional nuisance parameters \( \vec{\theta} \). The LHC experiments agreed on using the profile likelihood ratio, \( t^\mu \), as test statistic for searches. Such a variable is defined as

\[
t^\mu = \begin{cases} 
-2 \log \frac{L(\hat{\mu}, \hat{\vec{\theta}})}{L(\hat{\mu}, \hat{\vec{\theta}})} & \hat{\mu} > 0 \\
0 & \hat{\mu} \leq 0
\end{cases}
\]

with \( \hat{\mu} \) and \( \hat{\vec{\theta}} \) being the unconditional maximum likelihood estimators (MLE) while \( \hat{\vec{\theta}} = \hat{\vec{\theta}}(\mu) \) denotes the conditional MLE, as test statistic for searches. For a given (pseudo-)data set the \( p \)-value can be calculated from the sampling distribution\(^1\) \( f(t^\mu|\mu') \) according to

\[
p^\mu = \int_C f(t^\mu|\mu') dt^\mu
\]

where \( C \) is the set of possible observations representing less or equal compatibility with the tested model parameter \( \mu \) compared to the data observed. Sampling distributions \( f(t^\mu|\mu') \) can be obtained using parametric bootstrapping. Under certain conditions the distribution of test statistics is known in the asymptotic regime [4]. One regards a model as excluded at a confidence level \( 1 - \alpha \) if the corresponding \( p \)-value is found to be less than \( \alpha \). In order to prevent overly stringent exclusion limits from experiments with limited sensitivity, the \( p \)-value as defined in Equation 2 can be modified according to the \( CL_S \) technique. This prescription is commonly accepted in the HEP community for presenting exclusion limits and it is discussed in detail elsewhere [5].

3. The RooStats framework
RooStats is a joint C++ project between the LHC experiments ATLAS and CMS and it is based on ROOT [6] and RooFit [7]. Basic functionality like I/O operations and plotting are inherited from the ROOT package which is the most widely used software for data analysis in HEP. The RooFit framework provides all necessary functionality to construct and work with likelihood/probability density functions (PDF) as well as a convenient interface to the MINUIT minimisation package. Additionally, the C++ class RooWorkspace of RooFit can store PDFs, data sets and variables along with all necessary information about a model. These objects can

---

1 The theoretical formulation can easily be extended to problems with multiple parameters of interests.

2 A sampling distribution \( f(t^\mu|\mu') \) describes the probability of observing a value \( t^\mu \) for the test statistic assuming that the true model parameter is given by \( \mu' \).
be written to persistent data storage. Hence, the RooFit toolkit defines a standard data format for exchanging self-contained information between experiments.

While the data modelling infrastructure is provided by RooFit, RooStats adds another layer of high-level information only relevant to statistical interpretation. A common ingredient to all statistical methods is a PDF or likelihood function which, from the technical point of view, is a real-valued function of a certain number of variables. However, in statistics one distinguishes between the following types of parameters:\(^3\):

- **observables** describe quantities measured by an experiment,
- **parameters of interest** are the parameters of the model one is interested in,
- **nuisance parameters** are the set of all other model parameters.

When applying statistical tools, the roles of various parameters must be defined. In RooStats all interface classes representing statistical techniques can be configured by passing an instance of the `ModelConfig` class. A `ModelConfig` object can also hold information about prior PDFs which are needed for Bayesian approaches.

A versatile framework for statistical tools is needed to deal with simple scenarios (e.g. event counting analyses), as well as to handle complicated measurements incorporating the shape of observables together with all related uncertainties.

The design of RooStats is shown in Figure 1. It is driven by the goal to disentangle statistical concepts from the explicit model under investigation. Each statistical technique is mirrored by an interface to accomplish this task. This philosophy allows for applying the same tools to models of different experiments or even combined models of several analyses.

The `HypoTestCalculator` class provides the abstract interface for performing statistical hypothesis tests. After specifying the model for the null hypothesis (and the alternative) by passing an instance of the `ModelConfig` class and after providing the data observed, the method `HypoTestCalculator::GetHypoTest` can be invoked. The `HypoTestResult` object returned contains information about this hypothesis test, most importantly the \(p\)-values for the null and alternative hypothesis as well as the sampling distributions for the tested values of \(\mu\).

\[^3\] The distinction between observables and parameters is already important for the normalisation of PDFs. In RooFit the role of variables is inferred from the context in which the PDF is called/evaluated.
Confidence interval calculations are carried out by classes derived from IntervalCalculator. After initialising a concrete implementation of this class with a ModelConfig object and the data observed, the method IntervalCalculator::GetInterval returns a ConfInterval object. As the exact type of the calculated confidence interval (e.g. dimensionality, connected or disconnected parameter space) depends on the specific model, data and statistical method used, this abstract interface implements the general method ConfInterval::IsInInterval which checks whether a given point is included in the confidence interval.

Another important feature of RooStats is the support for parallelisation using the PROOF infrastructure. Especially, the creation of sampling distributions $f(t_{\mu} | \mu')$ requires the generation and evaluation of a large number of pseudo-data sets at each tested value $\mu'$. As the pseudo-data sets are statistically independent, the execution time can be reduced significantly by distributing this task over many CPU nodes. Furthermore, the class HypoTestResult supports merging which allows the user to add more test points/statistics at a later stage.

4. Conclusion and discussion
The RooStats project is motivated by the goal of providing a set of generic tools for statistical data analysis. The coherent class design facilitates the evaluation of complex models. Together with the data modelling capabilities of RooFit, it is possible to combine measurements at the analysis level by constructing a global PDF which allows for a correct treatment of correlated uncertainties. Providing a link to the PROOF infrastructure makes it possible to perform hypothesis tests on complicated models based on MC toy generation. The possibility of using parallelisation becomes crucial when the expected p-values are very small as was the case for the discovery of the Higgs boson in summer 2012.

In the context of the combination of ATLAS and CMS results for the Higgs boson search in summer 2011 [9], the RooStats toolkit was tested extensively and validated against independent software for specific analyses. By now, RooStats has been scrutinised by many analysis groups and was used to produce a large number of public results. Improving the interfaces and consolidating the source code are the focus of the current development activities, based on the experience of users and their valuable feedback.

In the early stages, the development of RooStats was driven by the needs of analyses searching for a Higgs boson which concentrate on models with one parameter of interest. With the increased amount of accumulated data, the focus will shift to models with more parameters of interest (e.g. simultaneous coupling/cross section measurements). While the generic interfaces are independent of the number of parameters of interest, some limitations may exist when visualising/interpreting higher-dimensional confidence intervals. These issues will be addressed in the future to meet the evolving needs of the HEP community.

Acknowledgments
CG is thankful to the LHC Physics Centre at CERN for the funding of his stay at CERN.

References
[1] Moneta L et al. 2010 PoS ACAT2010 (Jaipur, India)
[2] ATLAS Collaboration 2012 Phys. Lett. B 716 1–29
[3] CMS Collaboration 2012 Phys. Lett. B 716 30–61
[4] Cowan G, Cranmer K, Gross E and Vitells O 2011 Eur. Phys. J C 71 1554
[5] Read A L 2002 J. Phys. G 28 2693
[6] Antcheva I et al. 2009 CPC 180 2499–2512
[7] Verkerke W 2007 PHYSTAT LHC Workshop (CERN)
[8] Ballintijn M, Brun R, Rademakers F and Roland G 2003 CHEP (La Jolla, USA)
[9] ATLAS and CMS Collaborations 2011 ATL-PHYS-PUB-2011-011, CMS-NOTE-2011-005