Statistical Analysis of Simulated Data for the CMS Experiment at the LHC

Lukas Kostal, Leonardo Rotondi, Theo Gatward, Laith Mohajer, Hugo Boîtel-Gill

***Abstract*—A program for the statistical analysis of simulated CMS experiment data is presented. First, the methods used in the data generation for the $H \rightarrow \gamma\gamma$ decay mode are introduced. The paper explains the methods used to carry out histogrammed data error estimation, analytical and numerical background and signal parameter estimation and finally, reduced chi-squared statistical tests. The results obtained show the significance of the measured signal to be 4.52σ and therefore agree with the value of 4.6σ presented in the official Higgs boson discovery paper [1]. Additional details and results regarding presented program are then discussed in the appendix.**

I. Introduction

The standard model of elementary particle physics originates from the unification of the fundamental forces responsible for the electromagnetic, weak and strong interactions whose mechanism include gauge bosons. In order to explain the origin of the mass property of these bosons a scalar field called the Higgs field has been proposed in 1964. In this model the values for the mass of the gauge bosons originate from the extent of their interaction with the Higgs field which occurs via a mechanism involving the Higgs scalar boson.

Current theory can not predict the Higgs boson mass m_H, however multiple experiments from CERN have placed upper and lower bounds on the possible mass $122.1 < m_H < 129.2$ GeV/c^2 at a 95% confidence level [1]. The CMS experiment searches for the evidence of the existence of the Higgs boson by observing the following decay modes: $H \rightarrow \gamma\gamma$, ZZ, W^+W^-, $\tau^+\tau^-$ and $b\bar{b}$. This analysis focuses on the $H \rightarrow \gamma\gamma$ decay mode with a mass distribution of $m_{\gamma\gamma}$ which has the greatest mass resolution [1].

The observed signal has a low amplitude compared to the background. A complex statistical analysis is therefore required to characterise the data parameters. A reduced chi-squared test (RCS) can then be used to determine the significance of the signal peak.

II. Data Generation

The program presented in this paper generates a set of simulated data consisting of the background observations combined with the signal observations.

The background is generated by randomly sampling an exponential distribution with the scaled probability density function (PDF) shown in (1). The background depends on 2 parameters which are the background amplitude A_B and the rate parameter λ.

$$\rho_B(m_{\gamma,\gamma}; A_B, \lambda) = A_B e^{-\frac{x}{\lambda}} \quad \text{for } x \geq 0 \tag{1}$$

The signal is also generated by random sampling this time of a Gaussian distribution with a scale PDF shown in (2) which depends on 3 parameters: amplitude A_S, mean mass μ and the standard deviation σ.

$$\rho_S(m_{\gamma,\gamma}; A_S, \mu, \sigma) = \frac{A_S}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

The parameter values, shown in Tab. I, have been chosen to represent the CMS measurements.

TABLE I: Parameter Values Used in Data Generation

Parameter	Value	Units
A_B	10^5	Unitless
λ	30	GeV
A_S	400	Unitless
μ	30	GeV/c^2
σ	1.5	GeV

III. Parameter Estimation

In the first part of the statistical analysis the generated data is organized into 30 bins with an equal width of 1.7 GeV. For the histogrammed data the uncertainty in $m_{\gamma\gamma}$ is just halve the bin width and the uncertainty in the number of events in each bin is taken to be the standard deviation of the data points in each bin. This follows from the assumption that the data points in each bin follow a Gaussian distribution. This assumption is justified by the central limit theorem (CLT) and the standard deviation (SD) is then simply found by using the Gaussian SD estimator with the Bessel correction.

The 2 parameters characterising the background are estimated by considering a limited set of data which only includes the first 10 bins up to 120 GeV and therefore excludes the signal.

First the analytical estimate $\hat{\lambda}$ of the parameter λ is found by using (3) where x_i and N represent the limited data and the total number of data points respectively.

The equation has been derived by applying the maximum likelihood method (equivalent to the minimum χ^2 method) to the scaled exponential PDF.

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3}$$

An estimate of A_B is found by ensuring the area under the exponential curve to be equal to the area of the limited histogrammed data. The area under the curve is found by numerical integration of (1) with $A_B = 1$.

The analytical estimates are then used as initial guesses for a 2D numerical parameter estimation. The estimation works by defining a discretised 2D plane of trial values for the A_B and λ parameters. A RCS ${\chi_r}^2$ is then calculated for every point to find a global minimum whose position gives the numerical parameter estimates $\hat{A}_B$ and $\hat{\lambda}$.

A similar method has been used to estimate the signal parameters $\hat{A}_S, \hat{\mu}, \hat{\sigma}$, as described in the appendix. The estimated parameters are then used to determine expectation values for the signal and the background as shown in Fig. 1.

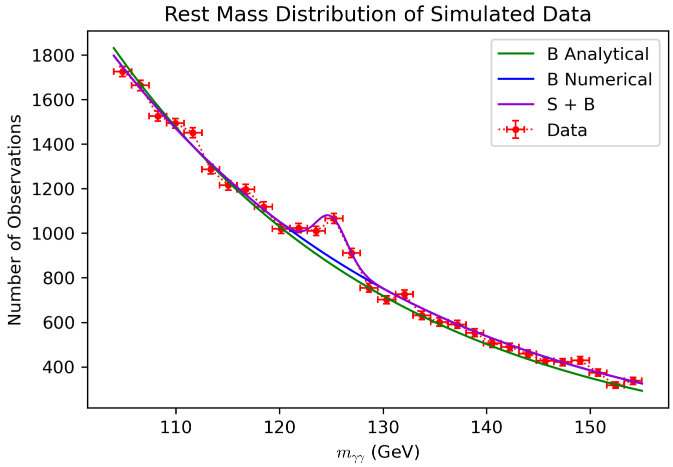

Fig. 1: Histogrammed rest mass spectra for $m_{\gamma\gamma}$.

IV. Hypothesis Testing

The reduced chi-squared hypothesis (RCS) tests are carried out by calculating a reduced chi-squared statistic ${\chi_r}^2$ as shown in (4) [2], where x_i represents the N data points considered in the test, ρ is the scaled PDF given by the null hypothesis H_0 and ν is the number of degrees of freedom.

$${\chi_r}^2 = \frac{1}{\nu}\sum_{i=1}^{N}\frac{(x_i - \rho(x_i))^2}{\rho(x_i)} \tag{4}$$

The determined RCS value is then used to calculate an alpha value α which represents the probability of the given ${\chi_r}^2$ value being obtained due to random fluctuations. The alpha value is calculated using the integral in eq:5 where $\rho_\chi(x;\nu)$ is the PDF of the reduced chi-squared distribution and x is a dummy variable.

$$\alpha({\chi_r}^2, \nu) = \int_{{\chi_r}^2}^{+\infty} \rho_\chi(x;\nu)dx \tag{5}$$

The calculated value of α is then compared to a chosen significance level given by $_C$, and if $\alpha < \alpha_C$ the null hypothesis can be rejected.

First the goodness of fit of the estimated parameters is tested using a background only null hypothesis $H_0 : \rho = \rho_B(x_i; \hat{A}_B, \hat{\lambda})$ and the limited dataset which excludes the signal.

Another RCS test is carried out to determine the significance of the signal present in the entire dataset including the signal, using the same background-only null hypothesis.

A final RCS test is performed to determine the goodness of fit of both the background and the signal parameter estimates using the entire dataset. This time the combined signal and background null hypothesis is used $H_0 : \rho = \rho_B(x_i; \hat{A}_B, \hat{\lambda}) + \rho_S(x_i; \hat{A}_S, \hat{\mu}, \hat{\sigma})$.

V. Results and Discussion

The background parameter estimates found using the 2D numerical estimation are $\hat{A_B} = 5.87 \times 10^4$ (3sf) and $\hat{\lambda} = 29.8$ GeV (3sf). The corresponding test statistic is ${\chi_r}^2 = 0.789$ (3sf) with an alpha value of 0.612 (3sf) which shows a good fit since ${\chi_r}^2 \sim 1$ and the H_0 is not rejected at $\alpha_C = 0.05$.

The second hypothesis test returns a value of ${\chi_r}^2 = 2.70$ (3sf) with a corresponding alpha value of $\alpha = 3.04 \times 10^{-6}$ (3sf). This leads to a rejection of the signal-only null hypothesis and shows the signal to be significant at 4.52σ (3sf).

The final hypothesis test for the signal and the background gives a value of ${\chi_r}^2 = 0.981$ (3sf) with an alpha value of $\alpha = 0.489$ 3sf indicating that both the background and the signal parameter estimates are close to the true values.

VI. Conclusion

In conclusion the search for the Higgs boson at the CMS experiment has been motivated. A program for simulated data generation as well as the subsequent statistical analysis has been presented and explained. RCS tests have shown the estimated parameters to fit the background as well as the signal data well. Additionally the observed signal has been show to be significant at 4.52σ which agrees with the value of 4.6σ presented in the official discovery paper [1]. The discovery has therefore been confirmed using the presented statistical analysis program.

References

[1] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan *et al.*, "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012.

[2] J. Taylor, *Introduction to error analysis, the study of uncertainties in physical measurements*, 1997.

[3] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2

[4] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[5] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 2015, pp. 1–6.

Appendix A
Analysis Program Details

The program presented in this paper can be found at the following GitHub repository, along with detailed comments explaining the functioning of the program.

In practice the random sampling of the exponential and Gaussian distributions during the simulated data generation is implemented using the np.random.exponential and np.random.normal functions from the NumPy library [3]. The calculation of the alpha values as shown in (5) has been carried out using the scipy.stats.chi2.sf implementation of the survival function from the SciPy library [4]. The alpha values are converted into multiples of σ using the stats.norm.ppf implementation of the percent point function which is essentially the inverse cumulative distribution function (CDF).

Since the program involves many numerical calculations which require a large number of iterations, the initial version of the program took at least several hours to process the full dataset. In order to reduce the run time of the program, it has been written in terms of NumPy arrays instead of Python lists which take longer to perform operations on. The Numba JIT Python compiler library [5] has been used to further optimise all of the loops within the code and to parallelise; the RCS distribution calculation, the critical signal amplitude estimation and also the μ numerical parameter estimation. In the end the program performed the entire statistical analysis in $t = 2903$ s or 48.4 min(3sf), taking 162 s for the 2D numerical parameter estimation, 1342 s for the iterated calculation of reduced chi-squared values and optionally 3194 s for the numerical estimation of the critical signal amplitude using mean chi-squared values. It should be noted that these times heavily depend on the numbers of iterations set in the program.

Appendix B
Signal Parameter Estimation

The program uses multiple different methods for the signal parameter estimation all of which rely on the isolated signal peak shown in Fig. 2.

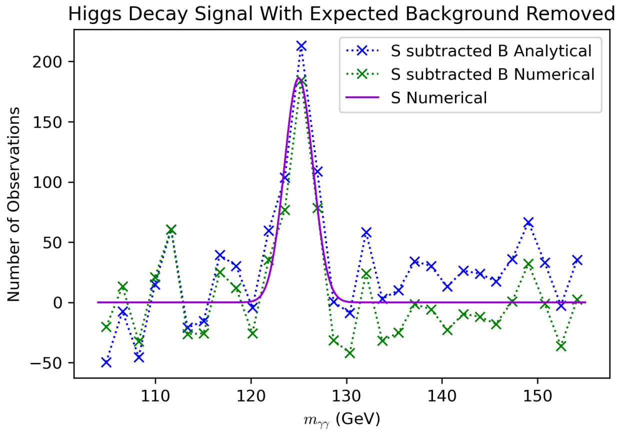

Fig. 2: Isolated signal peak obtained by subtracting the estimated background.

The simplest method of estimating the mass μ is to determine the center position of the highest bin in the isolated signal shown in Fig. 2. This method returns a value of $\hat{\mu} = 125.3 \pm 1.7$ GeV (4sf) with the uncertainty estimated from the bind width.

The most accurate method to estimate μ is a 1D numerical parameter estimation in which RCS values are calculated for a set of trial μ values. The parameter estimate $\hat{\mu}$ is then found as trial value which gives the global minimum of the RCS value as shown of Fig. 3 with the corresponding alpha values shown on Fig. 3. This estimation method returns a value of $\hat{\mu} = 125$ GeV exactly.

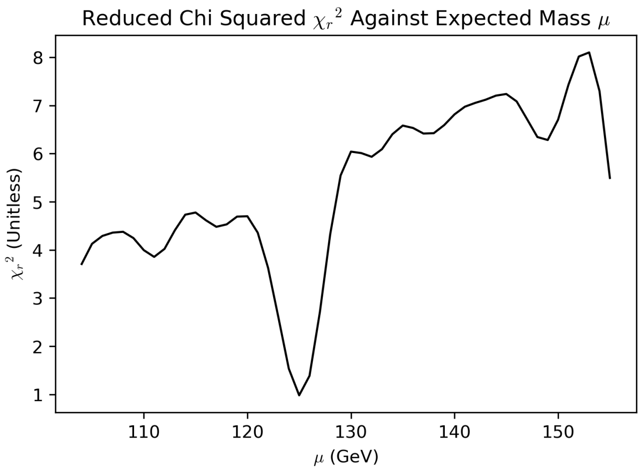

Fig. 3: Reduced chi squared value resulting from varying the expected mass μ.

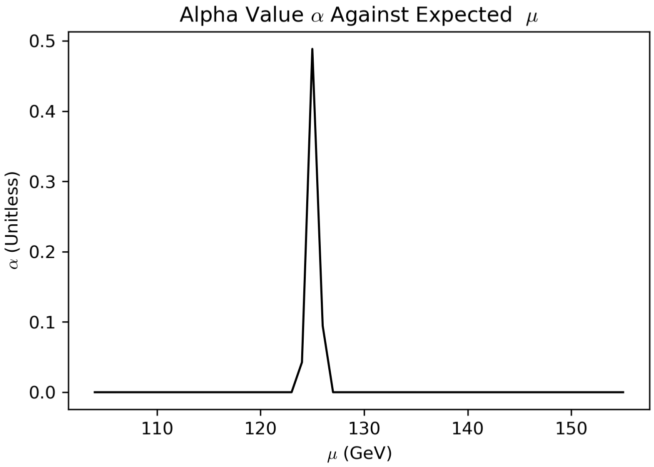

Fig. 4: Corresponding alpha values resulting from varying the expected mass μ.

An alternative method to estimate μ and σ is to use Gaussian parameter estimators on the isolated signal peak. This method gives values $\hat{\mu} = 125.0 \text{ GeV } (4\text{sf})$ and $\hat{\sigma} = 0.899 \text{ GeV } (3\text{sf})$.

The signal amplitude can be determined by summing the frequencies of the bins under the signal peak. This method gives an estimated value of $\hat{A}_S = 482 \ (3\text{sf})$. Preferably the signal amplitude would be estimated similarly to the background amplitude by using numerical integration to determine the area under the scaled Gaussian distribution.

Appendix C
Effects of Random Fluctuations

In reality the random fluctuations in the CMS experiment, which have been simulated by the previously mentioned random sampling, cause the RCS values to follow a reduced chi-squared distribution given by $\rho_\chi(x;\nu)$ [2]. In the program, this behaviour has been analysed for the background fluctuations by regenerating the dataset 10000 times with the number of signal observations set to 0. The set of reduced chi-squared values are then histogrammed to show the distribution of $\chi_r{}^2$ caused by the random fluctuations.

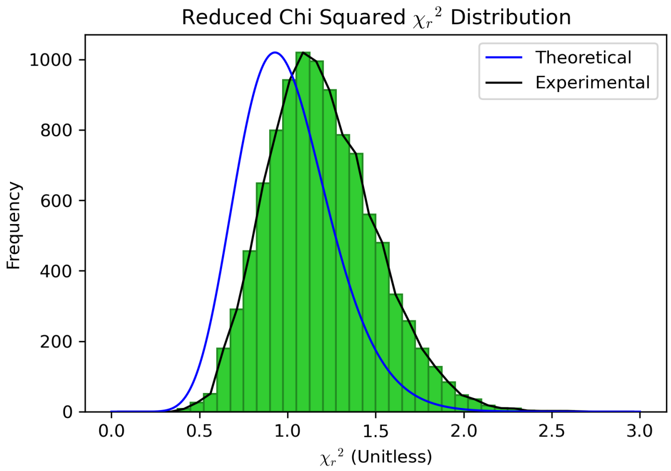

Fig. 5: Distribution of $\chi_r{}^2$ values from 10000 background-only datasets.

Fig. 5 also shows the theoretical $\rho_\chi(x;\nu)$ scaled PDF which appears to be slightly lower than the one obtained from the simulated data. This is to be expected since the two distributions would match only if the estimated parameter values were equal to the true parameter values. Since the parameter estimation is imperfect the $\chi_r{}^2$ values are slightly higher causing the distribution to shift right.

Appendix D
Critical Signal Amplitude Estimation

AS the amplitude of the signal is increased by increasing the number of observations in the generated dataset, the mean reduced chi-squared value for a background-only hypothesis increases which indicates an increasing significance of the signal. This behaviour can be seen in Fig. 6 shown below.

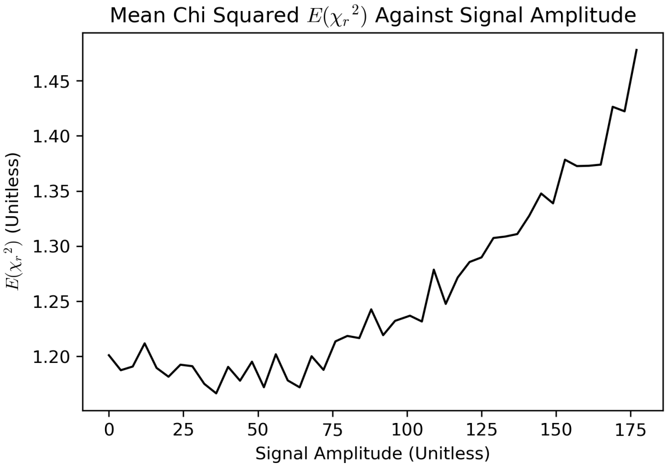

Fig. 6: Mean ${\chi_r}^2$ value for 1000 iterations against trial signal amplitudes

An inverse survival function with a critical value of $\alpha_C = 0.05$ is then used to determine the critical RCS value to be ${\chi_r}^2 = 1.48$ (3sf). The corresponding critical signal amplitude is found to be $A_S = 177$, so any number of signal observations greater than this value will lead to the rejection of the background-only null hypothesis at a 95% significance level.