Statistical Data Analysis, Assignment 6

Roel Deckers
Utrecht University

October 21, 2018

A Introduction

In this report we look for a Z-particle in a dataset in the range of 1 TeV to 3 TeV. By generating many psuedo experiments we determine the distributions of the log likelihood-ratio (LLR) for the case where we know that the mass $M_z = 2.1$ TeV and for the case where we do not known the real mass M_z and fit to the dataset instead.

The expected background follows an exponential distribution as a function of the invariant mass. In total, we expect to see 200 events. The background density is given by

$$\rho(m) = 200/(\exp(-1) - \exp(-3)) \times e^{-m}. \tag{1}$$

We expect to see 10 signal events in our data. Theory predicts that this invariant mass of the Z particle is 2.1 TeV. The detector resolution is guassian and known to be 50 GeV. We therefore expect a Guassian signal peak at 2.1 TeV, with a width of 50 GeV. We denote our expected background distrubtion by ρ_{H0} and our background + signal distribution by ρ_{H1}. We use ρ_H to refer to either of the two, depending on context.

B Psuedo-experiment generation.

In order to generate psuedo-experiments for our simulations we fill each bin μ_i according to a Poisson distribution with an expectation value equal to our model, weighted by the bin-width, evaluated at the center of the bin m_i, i.e. $\mu_i = \text{Poisson}(\text{binwidth} \times \rho_H(m_i))$. A pair of generated psuedo-experiments is shown in figure 1.

C Log likelihood-ratio

If we have a hypothesis H, and corresponding density function ρ_H then given histogram with bin centers $x_0, \ldots, x_n$ and corresponding counts $C_1, \ldots, C_n$, the likelihood $L_{M_{a,b}}$ of model $M_{a,b}(m)$ is given by

$$L_H = \prod_{i=0}^{n} \text{Poisson}(C_i; \mu = \hat{\rho}_H(x_i)). \tag{2}$$

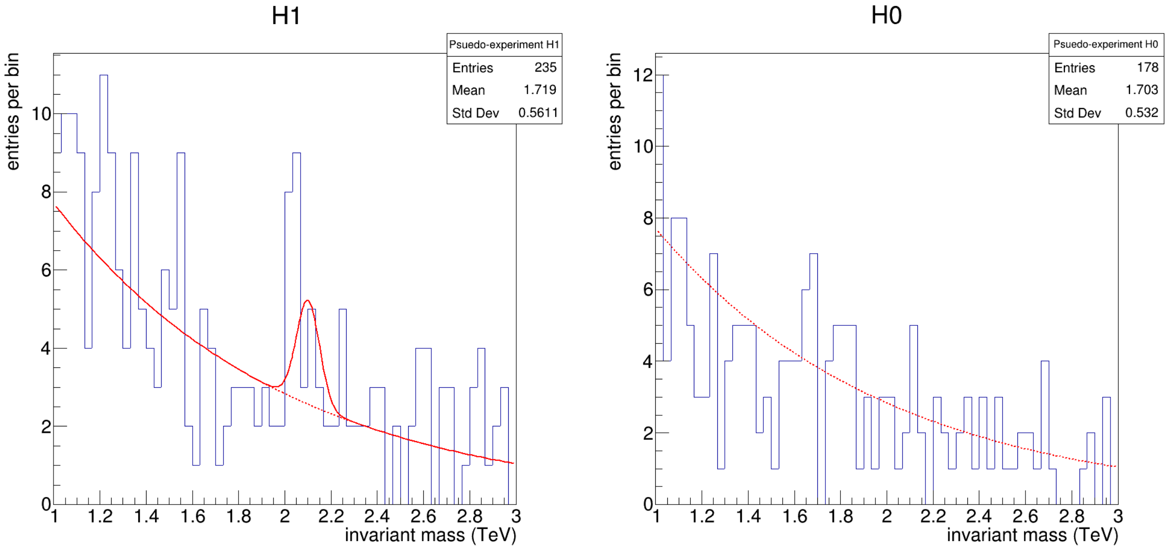

Figure 1: Psuedo experiments for H1 and H0, the expected distribution for H1 is shown in solid red, the expected distrubtion for H0 in dashed red. N.b. that the amount of entries s not fixed, but follows a Poisson distribution centered around 210 for H1 and 200 for H0.

That is, the product over all bins of the odds of seeing a measured bin count C_i, assuming measured bin counts are distributed according to a Poisson distribution with a mean given by the model prediction at the bin center.

Taking into account the definition of the Poisson distribution

$$\text{Poisson}(k; \mu) = \frac{\mu^k e^{-\mu}}{k!}, \tag{3}$$

and writing the factorial in terms of the Γ-function, the log-likelihood can be written as:

$$\log L_H = \sum_{i=0}^{n} \left(C_i \log(\hat{\rho}_H(x_i)) - \hat{\rho}_H(x_i) - \log\Gamma(C_i + 1) \right). \tag{4}$$

We use $\log\Gamma(k+1)$ because direct computation of $k!$ is computationaly infeasible for larger k and there exist[1] good premade approximations to $\log\Gamma$.

The log likelihood-ratio of two hypothesis H_0 and H_1 can be defined as

$$\lambda = \log\frac{L_{H_1}}{L_{H_0}} = \log L_{H_1} - \log L_{H_0} \tag{5}$$

D distribution of the Log likelihood-ratio

We have computed the distribution of λ using the just derived formula for the cases were the data follows H0, and the case were the data follows H1 using 500 000 pseudo-experiments for each. The results are shown in figure 2. As expected, we see a larger value of λ when the data follows H1 than when it follows H0 but that there still is a significant overlap between the two distributions. The expectation value of λ for H1 is 2.86 and -2.43 for H0.

[1] https://root.cern.ch/root/html524/TMath.html#TMath:LnGamma

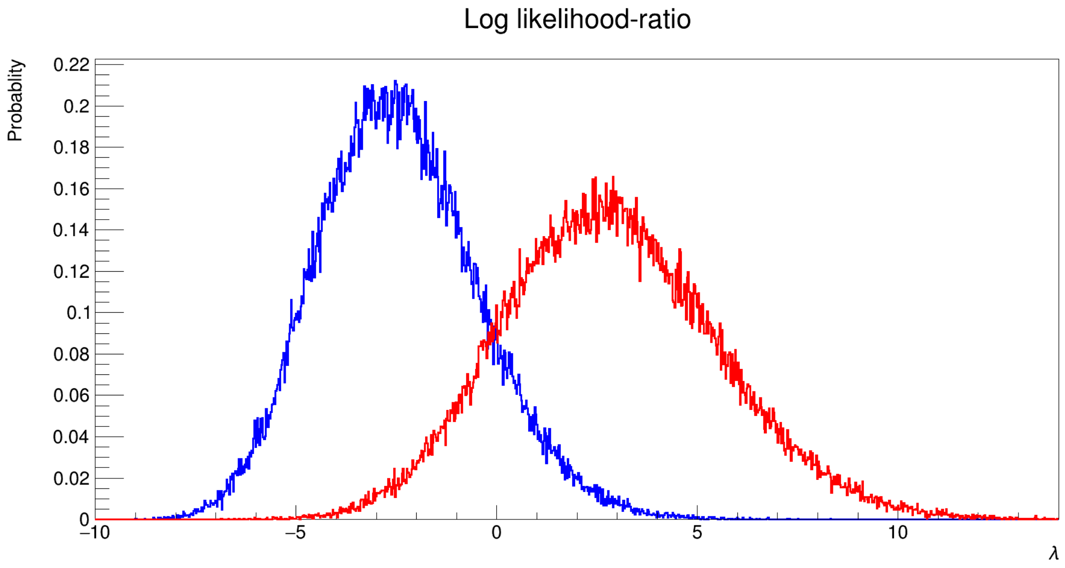

Figure 2: The probability distrubtion of λ for the case where the data behaves according to H1 (red) and H0 (blue).

E Real Data

We now use data from a real experiment, shown in figure 3, and determine its p-value. The p-value of a (pseudo-)experiment can be calculated by taking it's value of λ, and then plugging that into the cumulative density function of the distribution of λ under H0 and subtracting this value from unity. That is:

$$p = 1 - \mathrm{CDF}_{\lambda_{H0}}(\lambda) \tag{6}$$

Doing so with the expectation value of λ under H1 (2.86) gives the expected p-value under H1 (i.e. if the Z particle exists), namely, 0.008.

The p-value corresponding to the data show in figure 3 is 0.003, or almost 3σ. The CDF of the log likelihood-ratio is shown in figure 4.

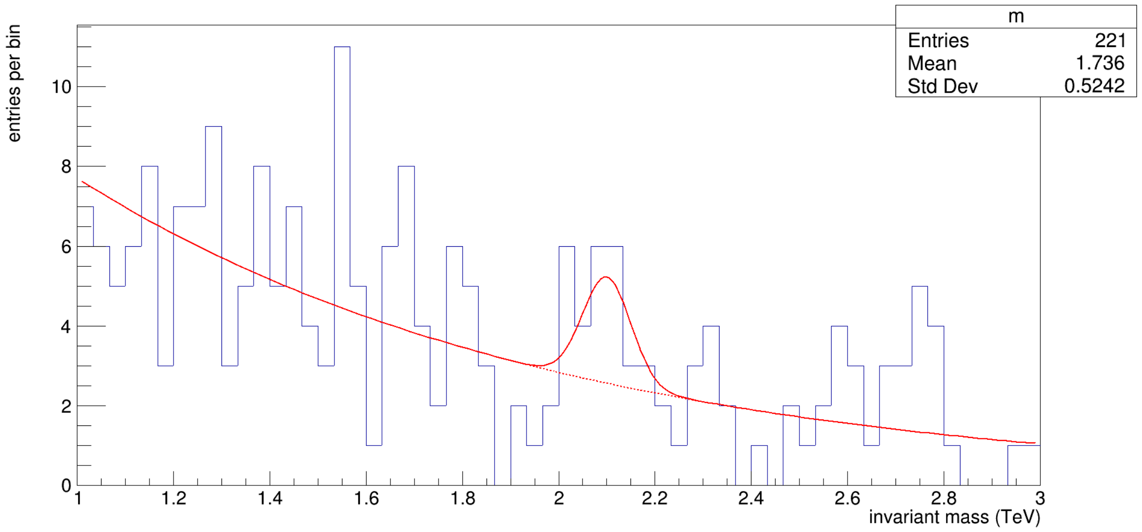

Figure 3: Distributed mass measurements with the expected distributions superimposed.

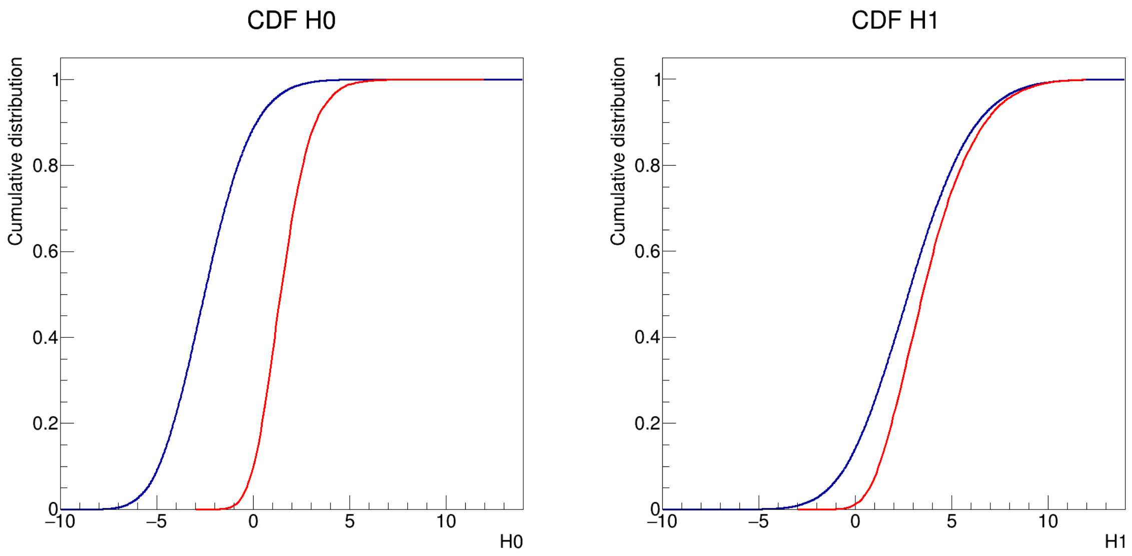

Figure 4: cumulative distribution functions for the log likelihood-ratio under H0 and H1. The blue lines correspond to the case where the mass of the Z is known and fixed at 2.1, whereas the red line corresponds to the case where we fit for the most likely mass (but our psuedo-experiments are still simulated at $M = 2.1$ TeV). We observe a significant shift and sharpening in the distribution under H0, making it harder to obtain good p-values.

F Unknown mass

We now consider the case where we do not know the true mass of the Z particle. That is, the mass of the Z particle is now a free parameter that we fit over in H1. That means we have to modify our definition of λ to

$$\lambda = \log \frac{L_{H_1}(\hat{M})}{L_{H_0}} = \log L_{H_1}(\hat{M}) - \log L_{H_0}, \tag{7}$$

where $\hat{M}$ is the the value of M which maximizes the likelihood of H1, determined by scanning over the range 1.2 TeV to 2.8 TeV with a step-size of ΔM =0.01 TeV.

In figure 5 we have recreated the log likelihood-ratio under H1 and H0, now using our adapted model for H1 which fits the Z mass $\hat{M}$ to the data. The corresponding CDF can be seen in figure 4. In figure 6 we have plotted the expected p-value as a function of the true Z mass, and in figure 7 we have plotted the distribution of the LLR under H1 as a function of the true Z mass.

Looking at these figures we note a few things.

First, looking at figure 5 and figure 4 we can see that even in the case where the true Z mass is still kept at 2.1 TeV it is still much harder to compare hypothesis H1 and H0 as there is a much larger overlap between the two LLR distributions. This is as expected, Statistical fluctuations of the background can now be confused for a signal event in a much wider range than before. Especially near the lower end of the mass spectrum, seeing a fluctuation on the order of 10 events is relatively likely, the variance of a Poisson distribution is equal to the expectation value after all.

Secondly, looking at figure 7 we see that as the true particle mass M_z increases, the LLR distribution of H1 shifts to the right and smears out. From this one would expect that the p-value decreases as the real mass increases, and this is confirmed by figure 6. This behaviour can be explained by looking at the background signal. As mentioned before the variance of a Poisson distribution is equal to its expectation value, and the expected number of entries in a bin follow a Poisson distribution with a mean determined by the model. The background model is an exponential curve meaning we expect much less background events at higher masses, which implies that there is much less variance on the background signal at higher masses. Meaning that we need much fewer signal events there to have a Statistically significant signal.

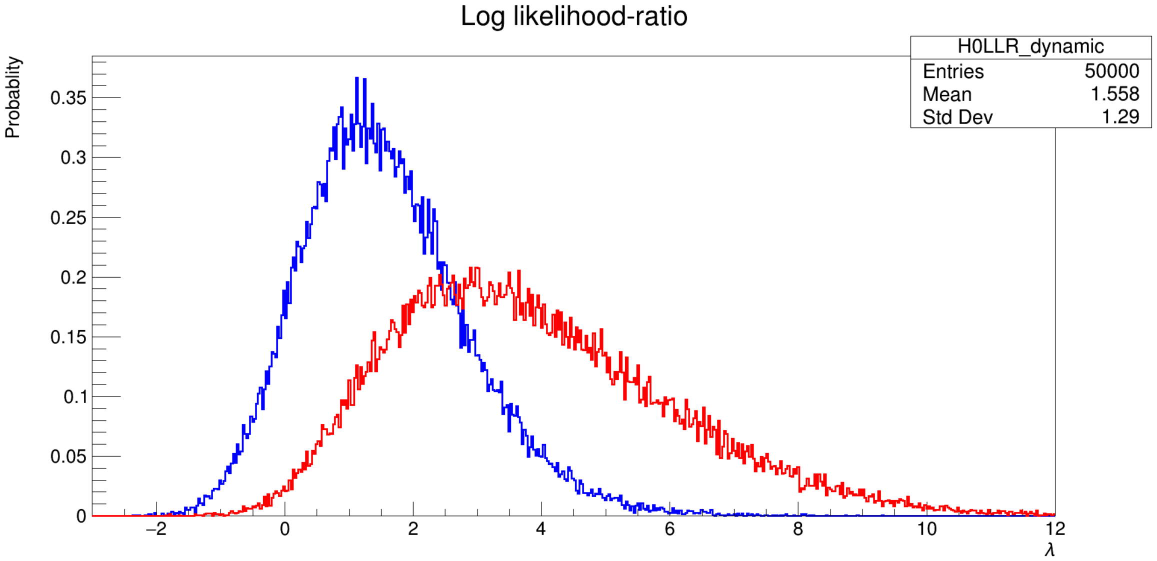

Figure 5: The probability distribution of λ for the case where the data behaves according to H1 (blue) and H0 (red) with the mass of the Z1 fitted by scanning over possible values and the real mass set to 2.1 TeV.

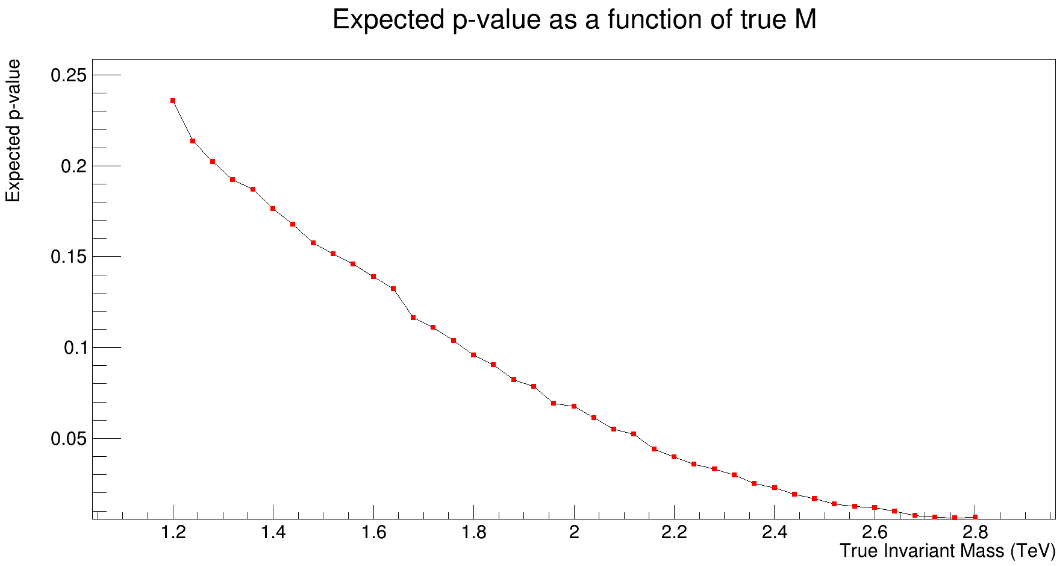

Figure 6: The expected p-value determined by running many psuedo-experiments as a function of the true particle mass M.

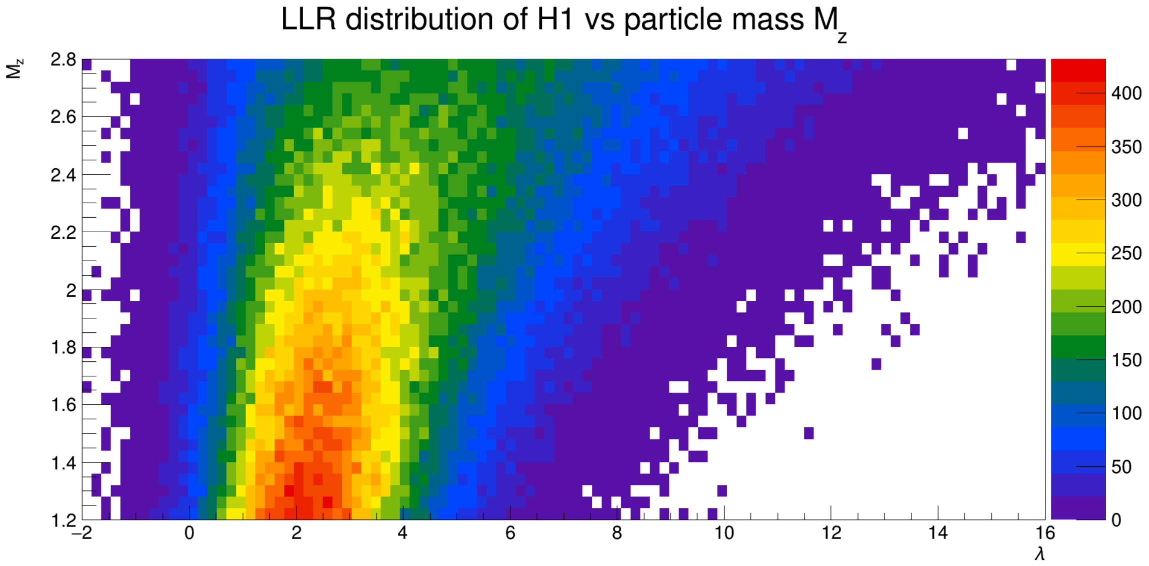

Figure 7: The probability distribution of λ for the case where the data behaves according to H1 as function of the true mass of the Z. Note how the mean increases and the distribution spreads out as the true mass increases. This can be attributed to the fact that we expect fewer background at higher invariant masses, leading to a more significant impact on the log-likelihood if the there exists a signal in that region.

G Results of real data

We now turn our attention back to the our real experiment data and fit to determine the most likely mass $\hat{M}$ of the data. When we step over all possible values of M in the range 1.2 TeV to 2.8 TeV with a resolution of 0.01 TeV we find that

$$\hat{M} = 2.74\text{TeV}. \tag{8}$$

If we take this to be the true mass of the Z particle, we would expect a p-value of 0.008. We observe a p-value of 0.013. While this would be Statistically significant in everyday use, this is much too large for a particle physics experiment to draw any conclusion (given the sheer volume of data produced in particle physics experiments false observations occur much more often, in absolute number, than many other fields). The result of this fit is plotted in figure 8, the LLR distribution corresponding to $\hat{M} = 2.74$ TeV is shown in figure 9.

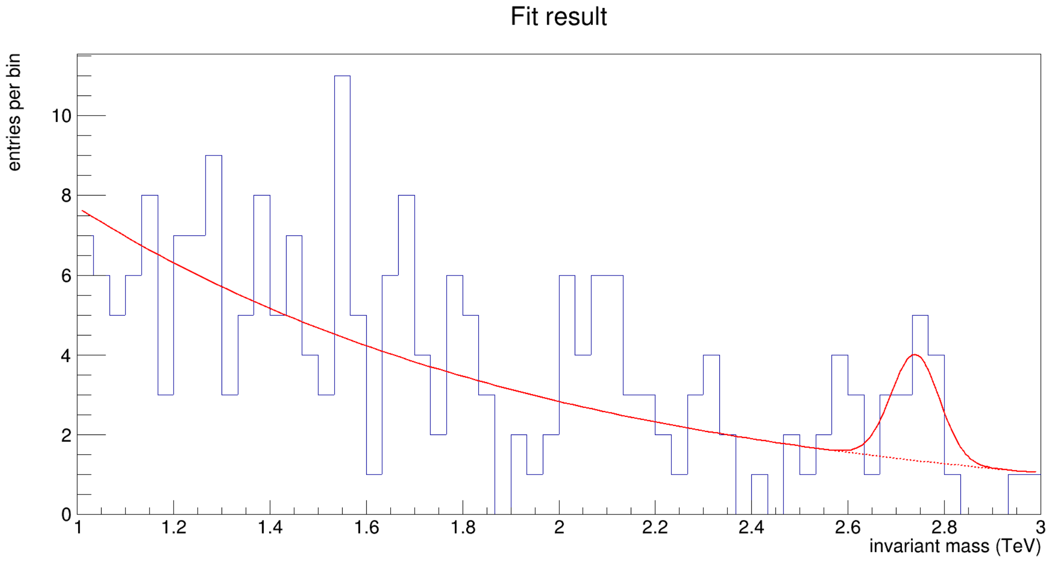

Figure 8: Experimental data with a fitted model superimposed where $M = 2.74$ TeV. The p-value of this fit and data is $p = 0.013$.

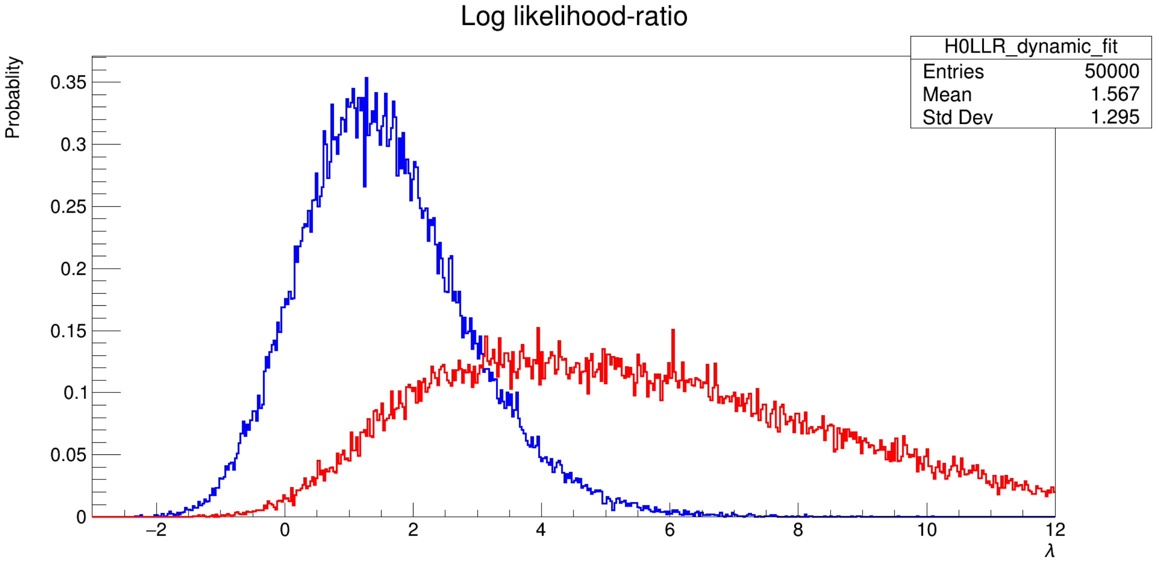

Figure 9: The log likelihood-ratio under H0 and H1 with a fitted mass, for a true mass of $M = 2.74$ TeV.

H conclusion

If the theoreticians are correct, and the mass of the Z particle is 2.1 TeV, we can say we have an observation of the Z particle in our dataset (i.e. we have a 3-σ measurement). However, if we consider the mass of the Z particle to be unknown to us except for restricting it to the 1.2 TeV to 2.8 TeV range, than we can not speak of an observation of the Z particle as we have a much larger p-value than acceptable.

This stark difference in statistical significance can be attributed to the simple fact that, if you look at enough bins of our dataset you will eventually always be able to find a bump simply due to the statistical fluctuations of the background signal. When we know the mass of the Z particle in advance we look at a much smaller set of data for our signal and therfore are less likely to suffer from what is known as the Texas-Sharpshooter fallacy.