Improved Asymptotic Formulae for Statistical Interpretation Based on Likelihood-ratio Tests

Ligang Xia Yan Zhang

Nanjing University

Abstract

In this work, we try to improve the classic asymptotic formulae to describe the probability distribution of likelihood-ratio statistical tests. The idea is to split the probability distribution function into two parts. One part is universal and described by the asymptotic formulae. The other part is case-dependent and estimated explicitly using a 6-bin model proposed in this work. The latter is similar to doing toy simulations and hence is able to predict the discrete structures in the probability distributions.

Introduction

When we do physics interpretation in a measurement, the test-statistics with the biggest power is likelihood ratio.

$$t_\mu = -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}$$

The classical asymptotic formula [1] describing t_μ's probability distribution is based on Wald's theorem.

$$t_\mu = \frac{(\hat{\mu}-\mu)^2}{\sigma^2} + \mathcal{O}(\frac{1}{\sqrt{N}})$$

where $\hat{\mu}$ abides by a Gaussian distribution with a mean μ_H and standard deviation σ. In practise, the $\frac{1}{\sqrt{N}}$ term is neglected and it applies well in the large-statistics cases.
However, we still meet many cases where few events are expected, especially, in searching for rare physics signals. A natural idea to overcome the limitation above is to split the probability distribution function (PDF) of a likelihood-ratio based test statistics, T_μ, into two parts according to the expected number of events.
Here is our new asymptotic formula [2] for T_μ's PDF. It has two parts.

$$\begin{aligned} f(T_\mu|\mu_H) &= \sum_{n=0}^{+\infty} f(T_\mu|n,\mu_H)P(n|b+\mu_H s) \\ &= \sum_{n=0}^{n_{small}} f(T_\mu|n,\mu_H)P(n|b+\mu_H s) + \sum_{n>nsmall} f(T_\mu|n,\mu_H)P(n|b+\mu_H s) \\ &\approx \sum_{n=0}^{n_{small}} f_{SS}(T_\mu|n,\mu_H)P(n|b+\mu_H s) + (1-\sum_{n=0}^{n_{small}} P(n|b+\mu_H s))f_{LS}(T_\mu|n_{small},\mu_H) \end{aligned}$$

n_{small} is the threashold we split the PDF of $t\mu$. f_{SS} describes the small-statistics (SS) contribution. It will be calculated using a 6-bin model proposed here. f_{LS} describes the large-statistics (LS) contribution. It is universal and just the classical formula with a propoer correction.

An Example

We design a pseudo experiment. The prototype is searching for Higgs boson using the $H\to\gamma\gamma$ mode. The signal is described by a Gaussian function while the background is described by a smooth and slowly-dropping exponential function. In the signal-sensitive region ($123 < m_{\gamma\gamma} < 127$ GeV), the expected signal (background) yield is 0.91 (0.64).

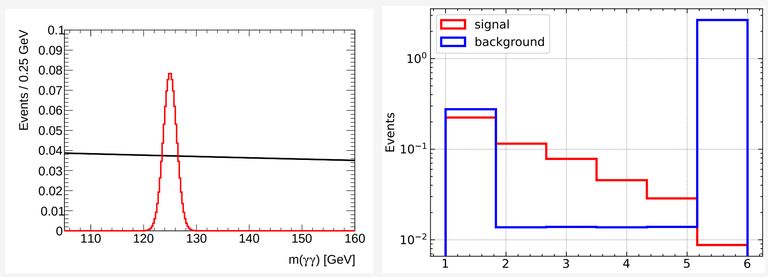
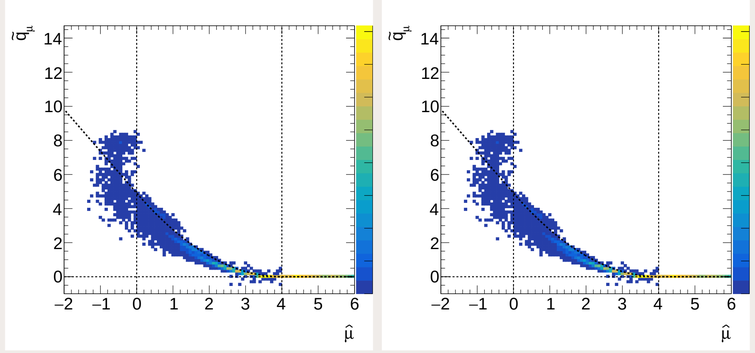


Figure 1. (L) Distribution of $m_{\gamma\gamma}$ in our pseudo experiment. (R) The 6-bin model used to calculate the small-statistics contribution.

Discrete Features in $\tilde{q}_\mu$ and $\hat{\mu}$

The scattering plot of $\tilde{q}_\mu : \hat{\mu}$ with $\mu = 3$ from the toy experiments under the hypothesis $\mu_H = 0$ (L) and $\mu_H = \mu = 3$ (R). The dashed curve shows the asymptotic formulae according to Wald's theorem. On the one hand, we can see that the asymptotic form still looks good even in these low-statisics cases. On the other hand, there are clear structures which reflect the discrete feature in the distribution of $\tilde{q}_\mu$ or $\hat{\mu}$.



Workflow

Here is the workflow to obtain the 6-bin model to calculate f_{SS}.

1. **Merge All Signal Region**, Merge the observable distributions in all signal regions into a fine-binning histogram for the signal and background component
2. **Re-order the bins with the decreasing significance**. Z_i represents each bin's significance. The defination of significance is
$$Z_i = 2\left[(b_i+\mu_H s_i)\ln\left(1+\frac{\mu_H s_i}{b_i}\right) - \mu_H s_i\right]$$
3. **Find the bin6** (denoted by i_6), the contribution of all the bins after which to the total significance is less than 0.1%. Define the signal and background yield summed over those bins as s6 and b6.
4. **Rebin into 6bins**, For the bins before i6 , we categorize them into 5 bins and the binning is determined by maximizing the significance.

Discrete features in the small-statistics contribution

Here we use the toy MC simulations with observing 2 events to illustrate the discrete features in the small-statistics contribution. The curves with different colors represent the solutions of $\hat{\mu}$ predicted in a simplified model.

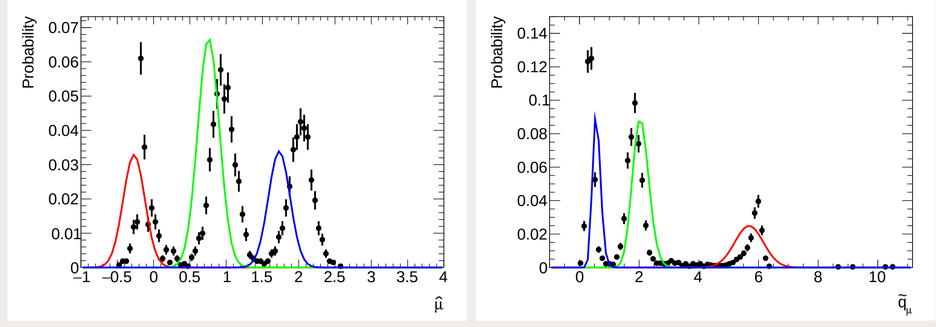


Figure 2. The distribution of $\hat{\mu}$ (L) and $\tilde{q}_\mu$ (R) from the toy experiments under the hypothesis $\mu_H = 3$.

6-Bin Model Approximation for Small Statistics

Once we obtain the 6-bin model, we can calculate the small-statistics part explicitly using the following formula.

$$\begin{aligned} f_{SS}(T_\mu|n,\mu_H) = &\sum_{k_0+k_1\ldots+k_5=n} \frac{n!}{k_0!k_1!\ldots k_5!}\prod_{i=0}^{5}(\frac{b_i+\mu_H s_i}{b+\mu_H s})^{k_i} \\ &\times f_{binned}(T_\mu|n_0=k_0,n_1=k_1,\cdots,n_5=k_5) \end{aligned}$$

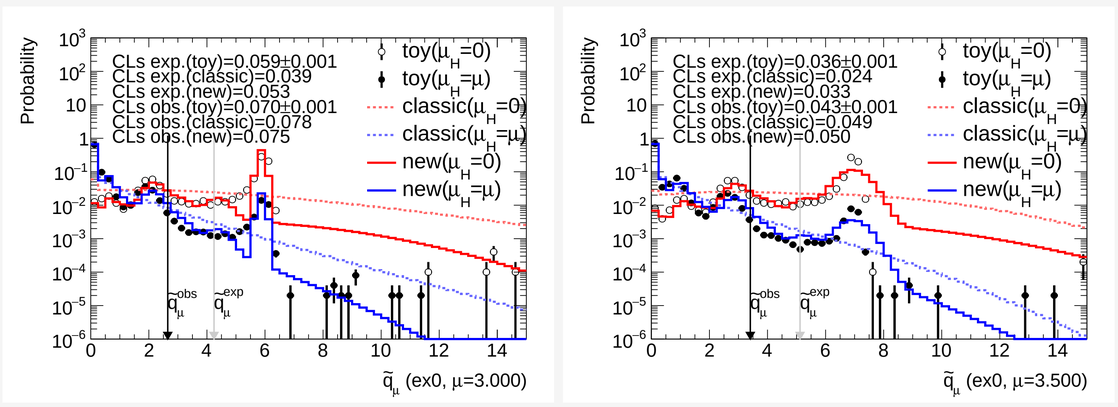
where f_{binned} is the PDF of T_μ for a given observation.

Results

Some examples of probability distributions of $\tilde{q}_\mu$ are shown below.

$$\tilde{q}_\mu(\mu) = \begin{cases} 0 & \hat{\mu} > \mu \\ -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & 0 \le \hat{\mu} \le \mu \\ -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \end{cases}$$

The black dots and open circles represent the toy MC results. The blue solid/dashed histograms represent the new asymptotic formulae in this work while the red solid/dashed histograms represent the classic asymptotic formulae from Wald's approximation. The black and gray arrows represent the observed and expected $\tilde{q}_\mu$, respectively.
We can see that new asymptotic formulae predict the bump structures in the PDF of $\tilde{q}_\mu$.



The table below summarizes the upper limits. We can see that our new formulae show better agreement with the toy MC results.

	Toy	Classic	New
Exp	3.19	2.80 (12.2%)	3.07 (3.5%)
Obs	3.38	3.48 (3.0%)	3.48 (3.0%)

Table 1. Summary of the upper limits at 95 % Confidence Level.

The numbers in the brackets indicate the relative difference with respective to the toy results.

Conclusion

In this work, we try to improve the classic asymptotic formulae to describe the probability distribution of the likelihood-ratio statistical tests which are commonly used in the field of high energy physics. The idea is to split the PDF into two parts. One is described by the classic formulae with proper corrections, and the other is calculated by mimicing the process of toy MC simulation. This idea successfully predict the discrete features in the small-statistics cases. Two examples with different sample sizes are presented and show that the new formulae have stable improvements on both the differential distribution of the test statistic and the upper limit calculation.

References

[1] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur. Phys. J. C 71 (2011) 1554, Eur. Phys. J. C 73 (2013) 2501 (Erratum), arXiv:1007.1727
[2] L.-G. Xia, Y. Zhang, arXiv:2101.06944