Automatic Datacard Generation and Significance Estimation using the Punzi Criterion for Higgs Analyses

Gautham Anne[1], Peter Dong[1]
Illinois Mathematics and Science Academy[1]

IMSA Illinois Mathematics and Science Academy

ABSTRACT

In order to perform a statistical analysis of Higgs searches, datacards are used to encode the observed and expected events after event selection to discriminate between signal and background. We developed a framework to incorporate into our analysis group's framework to automatically generate datacards. In addition, we utilized the Higgs Combine Tool to test combining multiple generated datacards. In the past, our group's selection cuts were based on the standard significance = S/sqrt(B) formulation. However, this method is dependent on a-priori expectations of limit settings to assert that a discovery has been made. We instead used the Punzi criterion, developed by George Punzi, as a better estimate of the significance. Simple, approximate formulas were derived and used for Poisson counts with background, with the suggested values for parameters (a=5).

INTRODUCTION

Datacards are required for setting limits on the masses for our doubly charged Higgs searches and dark photon searches. Manual setting of datacards leads to human error and can be very time consuming. Thus, we set out to develop a framework that allows for efficient automatic datacard txt file generation.

A frequentist definition of sensitivity for search for new phenomena is discussed, that of which has several useful properties. It is known as the Punzi Criterion, and it is particularly suitable for optimization, thus allowing the determination of a single set of cuts optimal for both setting limits and for making a discovery. In fact, we derive simple, quick approximation for the common problem of Poisson counts with background.

METHODS: DATACARDS

Each datacard contain information about the background, signal, and systematic uncertainties.

- We first initialize a HiggsPlusPlusAnalysis object
- For every channel in the object
  - We create a txt file with the channel name
  - For each process in the channel
    - Include yields and systematics

The Higgs Combine Tool runs thousands of pseudoexperiments based off these numbers and returns how much signal we would see if the doubly charged Higgs does not exist. This is repeated for various mass values of the Higgs, and for electrons/muons.

```
imax 4 number of bins
jmax 5 number of processes minus 1
kmax 0 number of nuisance parameters
----------------------------------------------------------------------------------------------------------
bin          Lep2       Lep3       Lep4       Lep5
observation  43.4204533 43.4204533 43.4204533 43.4204533
----------------------------------------------------------------------------------------------------------
bin          Lep2         Lep2         Lep2         Lep2         Lep2         Lep2         Lep3         Lep3         Lep3         Lep3         Lep3         Lep3
process      Higgs        DYMuons      WZ3LNu       DYElectrons  ZZ           TTBar        Higgs        DYMuons      WZ3LNu       DYElectrons  ZZ           TTBar
process      0            1            2            3            4            5            0            1            2            3            4            5
rate         0.0052796301 2942.016539  4.519904887  9.726097434  1.076796211  609.65       0.1658915382 478.7322276  4.842755236  1.015842367  0.7487769504 0
----------------------------------------------------------------------------------------------------------
```

Figure 1: Datacard generated from framework.

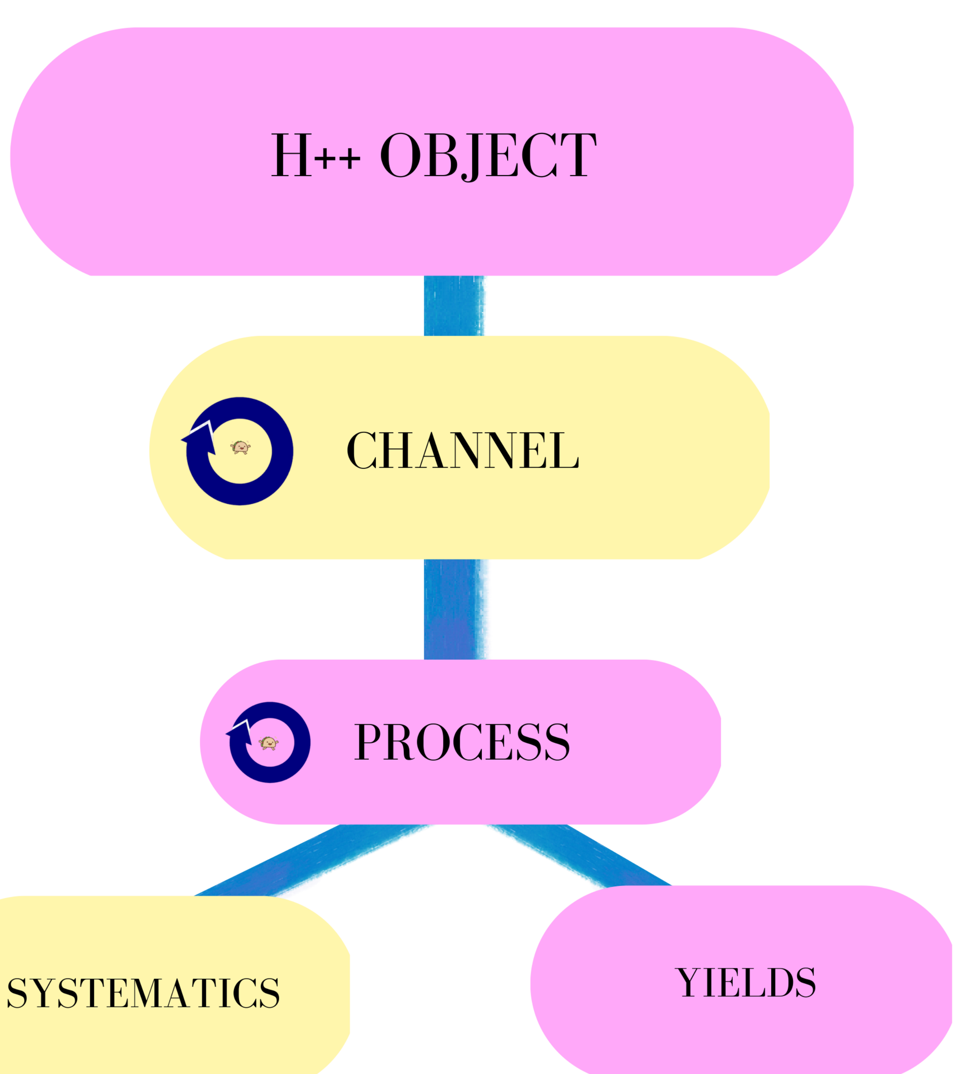

Figure 2: Diagram representing the process of creating datacards. Cycles symbolize for each loops.

```cpp
    HiggsPlusPlusAnalysis analysisObj;
    auto channels = analysisObj.getChannels();

  for(auto channel : channels)
    {
        channel->makeDatacard(channel);
    }
    return 0;
{
    auto processName=Utility::substitute(process->getName()," ", "_");
    datacard<< std::setw(20) << std::left << processName;
}

datacard<<"\n";
datacard<< std::setw(20) << std::left << "rate";

for (auto process : channel->getProcesses())
{
    auto yield=std::to_string(process->getYield("Invariant Mass"));
    datacard<< std::setw(20) << std::left <<yield;
}
datacard<<"\n";
```

```cpp
void Channel::makeDatacard(std::shared_ptr<Channel> channel)
{
    std::string channelName = channel->getName();
    std::string filename=channelName+".txt";
    std::ofstream datacard (filename);
    if(!datacard)
    {
        throw std::runtime_error("Unable to create file");
    }
    datacard << "\nimax    1 number of bins\n";
    datacard << "jmax    ______ number of processes minus 1\n";
    datacard << "-----------------------------------------------
    datacard <<"bin";

    for (auto i = 0; i < 5; i++)
    {
        datacard <<std::setw(20) << std::left << channelName;
    }

    datacard<< std::setw(20) << std::left << "process"<< std::left;
```

Figure 3: C++ implementation of Datacard Generator

METHODS: PUNZI CRITERION

In our experiments, we can define χ to be the observed variable, and the hypothesis H_0 is defined as the distribution of χ being a Poisson with the mean equal to the number of background events B. The hypothesis H_m is the Poisson distribution with a larger mean equal to $B + S_m$, where S_m is the new signal, and m the unknown free parameter of the theory.

We test H_0 by selecting a range of χ that helps us reject H_0. This range is known as the critical region, and the significance level of the test, α, is the probability of rejecting H_0 when it is indeed true. We currently abide by the standard physics convention of α corresponding to the 5σ tail of a gaussian distribution.

The new definition of sensitivity has to be independent of the choice of metric (both in the observable and parameter spaces) and should not require a choice of priors. It should be independent of the expectation of signal, discouraging the presence of bias in optimization. The following distributions represent the probability of a discrete observable n given H_0 and also H_m to be true:

$$P(n|H_0) = e^{-B} \cdot B^{n/n!} \quad \text{and} \quad P(n|H_m) = e^{-B-S_m} \cdot (B+S_m)^{n/n!}$$

We can assume that S_m is non-zero, and impose the condition that $n > n_{min}$, and similarly that $S_m > S_{min}$. Using a Gaussian approximation of the Poisson, we can state that

$$S_{min} = a\sqrt{B} + b\sqrt{B + S_{min}} \Longrightarrow S_{min} = \frac{b^2}{2} + a\sqrt{B} + (\frac{b}{2})(b^2 + 4a\sqrt{B} + 4B)$$

If we assume that the efficiency of the signal cuts are independent of m, that is, $S_m(t) = e(t) \cdot L \cdot \sigma_m$

and we aim to maximize the sensitivity by minimizing σ_m, the sensitivity equals.

$$e(t) \cdot \frac{L}{\frac{b^2}{2} + a\sqrt{B}} + (\frac{b}{2})(b^2 + (b^2 + 4a\sqrt{B} + 4B(t)$$

If we set a = b, then the expression becomes simplified to

$$\frac{e(t)}{\frac{a}{2} + \sqrt{B(t)}} \tag{1}$$

Consider the common significance expressions used for various optimization purposes:

$$\frac{S}{S+B} \quad \text{and} \quad \frac{S}{\sqrt{B}}$$

The first expression cannot be calculated unless the cross section of the signal is known. Also, it is not directly related to the significance of the search, but rather is related to the uncertainties in measurements of yields of a new process. The second expression is favorable as it is linear in S, and is independent of the cross section of the new process, but it breaks down at small values of B.

If the efficiencies are not known, $\frac{S}{\frac{a}{2} + B(t)}$ is a good approximation of sensitivity and does not have the flaws of the standard expression.

The method outlined above is just to help the optimization of an analysis at early-stages of selection. It assumes a counting experiment and neglects systematics. The method cannot replace a sensitivity calculated with a full likelihood ratio approach.

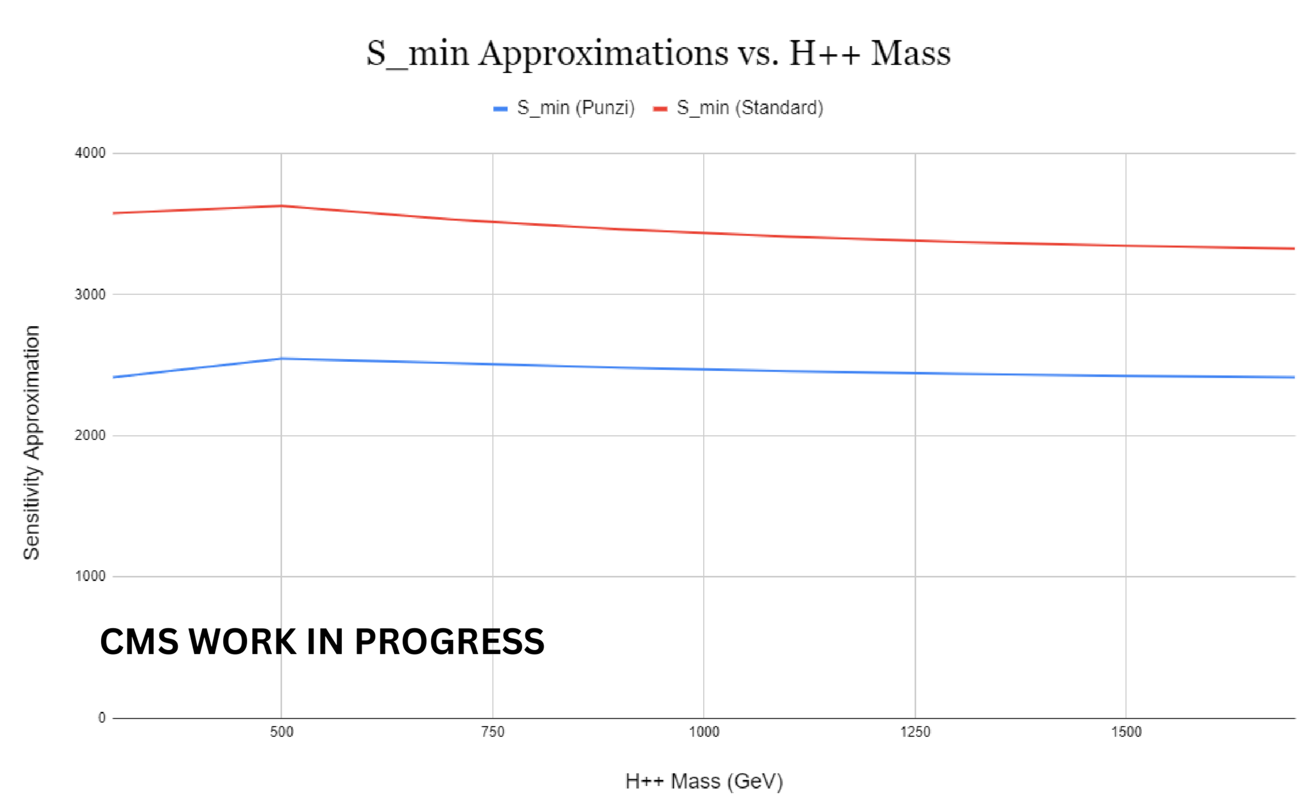

Figure 4: S_min as approximated by Punzi Criterion (blue) and S_min from standard approximation (S/sqrt(B)) (red). In this case, the cuts would be employed at 500 GeV regardless of the approximation, but that can be attributed to lower # of bins.

CONCLUSION

- Developed framework to automatically generate datacards
- Integrated with Higgs Combine Tool
  - Only compatible with CMSSW version 10.2.13, while current version is CMSSW 12.4.3
- Need to include systematics and global systematics
- Determined simple approximations of sensitivity for poisson counts with background
  - Cannot replace sensitivity search that accounts for systematic uncertainties calculated with likelihood ratios
  - Used in early stages of selection cuts
  - Independent from a-prior expectations

REFERENCES

Punzi, G. (2003). Sensitivity of searches for new signals and its optimization. Proceedings of PHYSTAT2003. doi.org/10.48550/arXiv.physics/0308063