This page documents the tutorial for the "LPC Stats II Hands On Tutorial Event", focusing on the usage of the CMS combination tool to extract the signal strength, statistical significance and upper limits in a physics model.

General Introduction

In this tutorial, we use the H->WW->lnln analysis as an example to illustrate how to build a physics analysis based on both cut-n-count and shape method and calculate the upper limits, significance. For simplification, we use ideal MC sample without reconstruction and include only one background (continuum WW) in the search.

Disclaimer: Numbers in this tutorial are motivated by the HWW analysis, but do not have one-to-one correspondance to the HWW results. Please do not be alarmed that your result is not the same as in the official note or AN.

Setting up the environment for executable combine

note: as for the rest of CMSSW software, this requires scientific linux 5, so you should log into lxplus5.cern.ch and not lxplus.cern.ch

```
setenv SCRAM_ARCH slc5_amd64_gcc472 # use export SCRAM_ARCH=slc5_amd64_gcc472 for bash
cmsrel  CMSSW_6_1_1
cd CMSSW_6_1_1/src
cmsenv
addpkg HiggsAnalysis/CombinedLimit V03-01-12
scramv1 b
```

After this, you have an excutable called "combine" that can be used to do tons of statistical calculations. Use `combine help` to see the ultimate list of options.

Building analysis inputs

Cut and count

In a cut and count analysis, we need to know what is the number of signal and background events expected, the number of events observed in data and associated systematic uncertainties. Each source of the systematic uncertainty is labelled as a nusiance. It is important to notice that even the MC statistic error for a given background is assumed as a nusiance.

Datacard is the input to run the statistical tools in the Higgs analysis in CMS. This can be adapted for other search program as well. There are two different softwares in CMS that does this computation, one is in LandS and other is in HiggsLimit/combination. The common data card for a cut-based analysis is given below.

Here is an example of the card called hww_20fb_cut.txt. In this example we consider 20% systematic uncertainty for the signal and 10% for the background.

```
imax 1 number of bins
jmax 1 number of processes minus 1
kmax 2 number of nuisance parameters
--------------------------------------------------
bin          of0j
observation  505.0
--------------------------------------------------
bin                             of0j      of0j
process                         ggH       qqWW
process                         0         1
```

```
rate                                  90.0000   430.0000
-------------------------------------------------------
uncert_HWW              lnN     1.2       1.0
uncert_qqWW             lnN     1.0       1.1
```

- This card can be combined with other channels for combination. Suppose we have another channel with input hww_20fb_1j_cut.txt. You can combine the two channels together as follows

```
combineCards.py of0j=hww_20fb_cut.txt of1j=hww_20fb_1j_cut.txt > hww_20fb_cut_comb.txt
```

Shape analysis

In addition to use only the number of events, shape analysis exploits the kinematic shapes as well. This is equivalent to sub-dividing the analysis into more categories according to the kinematic shape.

Here is an example card

```
imax 1 number of bins
jmax 1 number of processes minus 1
kmax 3 number of nuisance parameters
-------------------------------------------------------------------------------------------------------
shapes *         ofj0      hww_20fb_shape.input.root histo_$PROCESS histo_$PROCESS_$SYSTEMATIC
shapes data_obs  ofj0      hww_20fb_shape.input.root histo_Data
-------------------------------------------------------------------------------------------------------
bin          ofj0
observation  5729.0
-------------------------------------------------------------------------------------------------------
bin                                          ofj0       ofj0
process                                      ggH        qqWW
process                                      0          1
rate                                         228.1320   3981.6820
-------------------------------------------------------------------------------------------------------
CMS_hww_0j_WW_8TeV_SHAPE   lnN               -          1.1
CMS_hww_MVAWWBounding   shape               -          1.0
uncert_HWW              lnN                 1.2        1.0
```

Compare to the cut-based analysis you would notice the following

- You need an input file of the signal and background shapes, currently called "hww_20fb_shape.input.root". You can find this at `/afs/cern.ch/user/y/yygao/public/hww_20fb_shape.input.root`
- Inside the input root file, you can see the histograms with specific naming as declared at the beginning of the data card

```
  KEY: TH1D   histo_ggH;1   histo_ggH
  KEY: TH1D   histo_Data;1   histo_Data
  KEY: TH1D   histo_qqWW;1   histo_qqWW
  KEY: TH1D   histo_qqWW_CMS_hww_MVAWWBoundingUp;1   histo_qqWW_CMS_hww_MVAWWBoundingUp
  KEY: TH1D   histo_qqWW_CMS_hww_MVAWWBoundingDown;1   histo_qqWW_CMS_hww_MVAWWBoundingDown
```

- The number of signal and background events are much larger than in cut-based analysis. This is because we apply much looser analysis cut in the shape analysis.

Statistical calculations

Upper limit on the signal strength

```
combine -d hww_20fb_cut_comb.txt -M Asymptotic
```

```
Output:
 -- Asymptotic --
Observed Limit: r <1.8221
Expected  2.5%: r  <0.6019
Expected 16.0%: r  <0.8016
Expected 50.0%: r  <1.1289
Expected 84.0%: r  <1.6194
Expected 97.5%: r  <2.2782
```

This indicates that the 95% upperlimit observed is 1.8, while the median expected is 1.1 with 1sigma band [0.8, 1.6] and 2sigma band [0.6, 2.2]. This result tells us that we observed an excess in data at the level of 1-2 sigma.

Expected significance

```
combine -d hww_20fb_cut_comb.txt  -M ProfileLikelihood -v 1 --significance --expectSignal=1 -t -1

Output:
 -- Profile Likelihood --
Significance: 1.80964
       (p-value = 0.035176)
```

Observed significance

```
combine -d hww_20fb_cut_comb.txt  -M ProfileLikelihood -v 1 --significance  -m 125
Output:
 -- Profile Likelihood --
Significance: 1.52713
       (p-value = 0.063364)
```

Best fit signal strength with uncertainty

A maximum likelihood scan is performed to get the +/- 1 sigma error.

```
combine -d hww_20fb_cut_comb.txt   -M MaxLikelihoodFit
Output:
 --- MaxLikelihoodFit ---
Best fit r: 0.832828  -0.535222/+0.559076  (68% CL)
nll S+B -> -6.19048  nll B -> -5.02441
```

- Note, The fit result is also saved in the file called "mlfit.root", which contains all the results, including the best fit nusiances and the signal strength, correlations between the nuisances etc.

Useful Links

- Above tutorial is a super slimmed version of the complete tool, see the documentation here, Documentation of the RooStats-based statistics tools CMS

This topic: Sandbox > LPCStatsHandsOnTutorial
Topic revision: r4 - 2013-06-26 - YanyanGao