RooStats Example with WH, H→bb

Lianliang Ma（马连良）
Shandong University

iSTEP 2014

Beijing, Aug. 25 2014

Outline

- ✧ Introduction
- ✧ Code explanation based on WH
  - Input histograms for each channel and data
  - Xml file to describe the analysis
  - Build the workspace
  - Run Roostats on the workspace to get stats. results
- ✧ Run the example step by step
- ✧ Tips to use Roostats

Brief Introduction to Roostats

http://pos.sissa.it/archive/conferences/093/057/ACAT2010_057.pdf

- ✓ A joint project between the LHC experiments and the ROOT team, upon the request in Higgs search studies
- ✓ Roostats based on RooFit package, a set of C++ classes to provide major statistical techniques with coherent interfaces
- ✓ Emphasis on discoveries, confidence intervals, and combined measurements

More information https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome

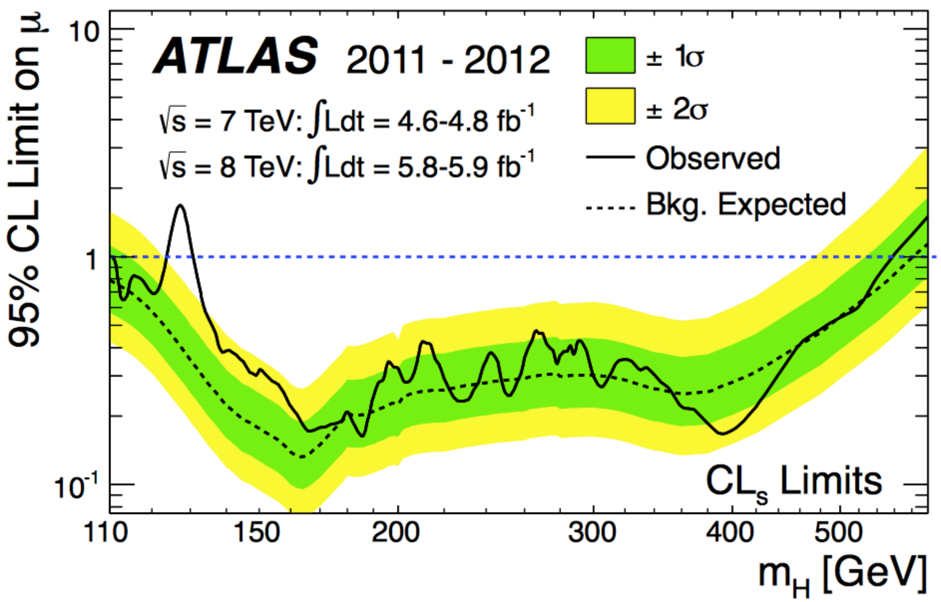
Introduction: Hypothesis Tests

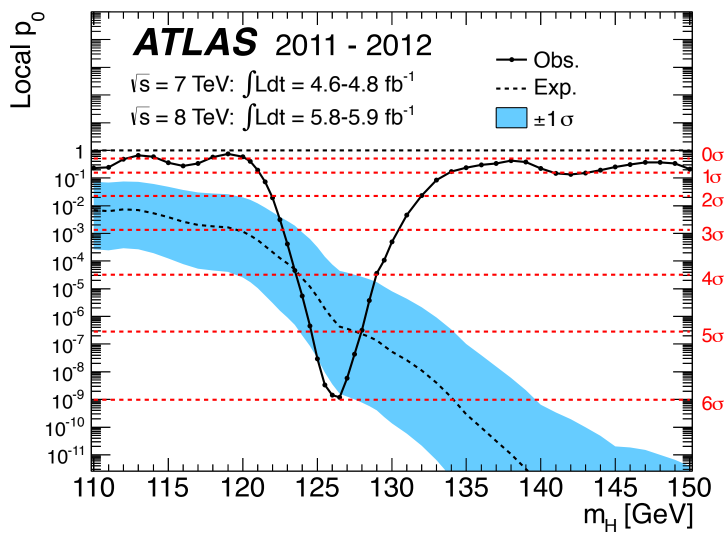
Hypothesis tests on physics measurements:

✓ Exclusion limit: exclude a signal of a given Xsec at X C.L.

✓ Significance (p0): the possibility of the background only to describe a measurement (3σ: evidence; 5σ: discovery)

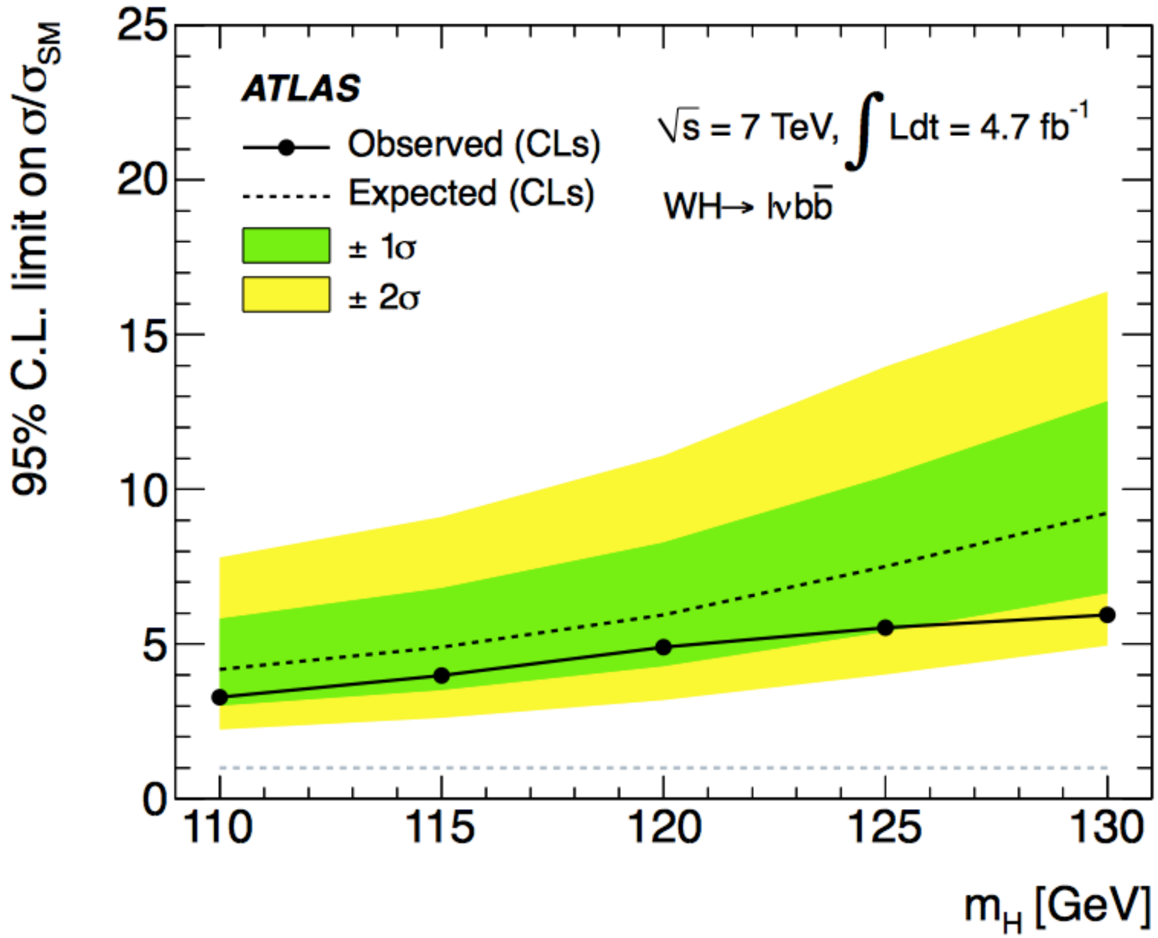
ATLAS discovery paper: Phys. Lett. B 716 (2012) 1-29





Example

ATLAS 7 TeV result: Phys. Lett. B 718 (2012) 369-390

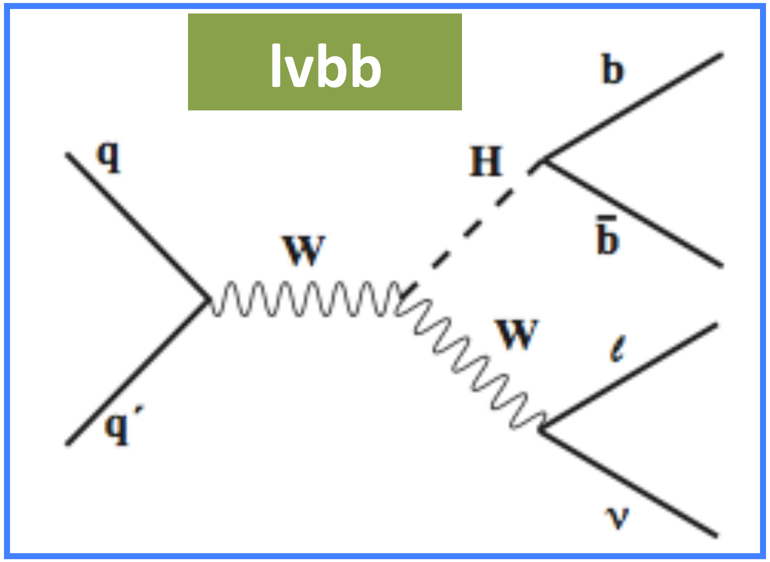


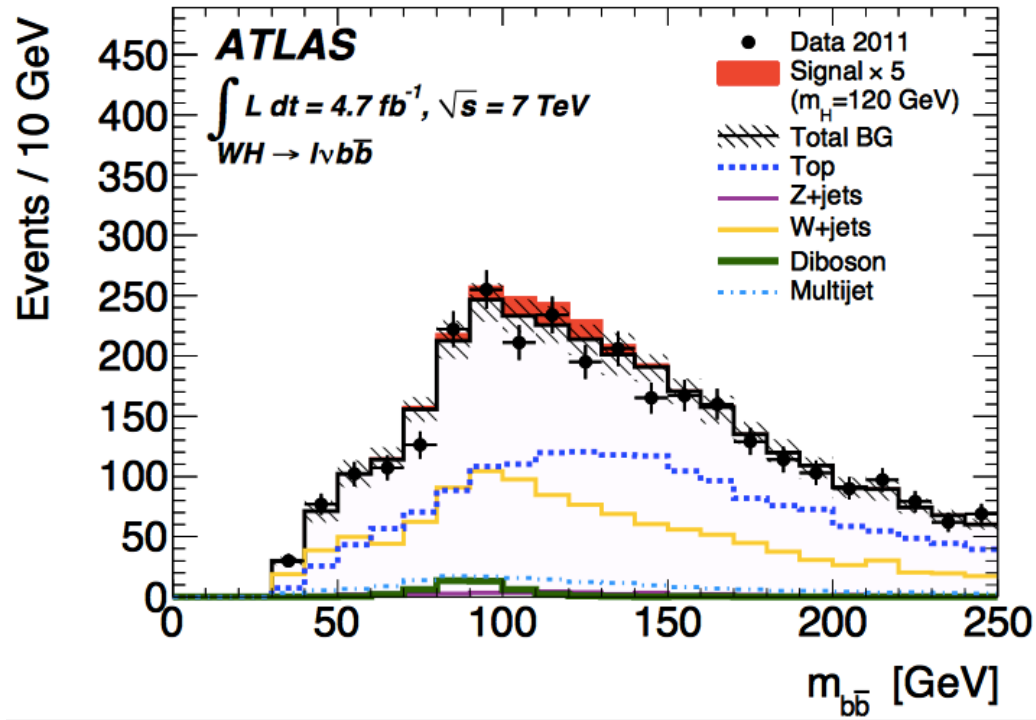
This tutorial tells you how to produce the upper limits for one mass point with WH, H->bb analysis

This example here is only for illustration purpose (easy to read and run), and more advanced code can be provided upon request

Some plots are not plotted at the same luminosity

WH, H->bb Analysis





Invariant mass (mbb) used for statistical treatment after event selections for this talk (Other variables can also be used: BDT output)

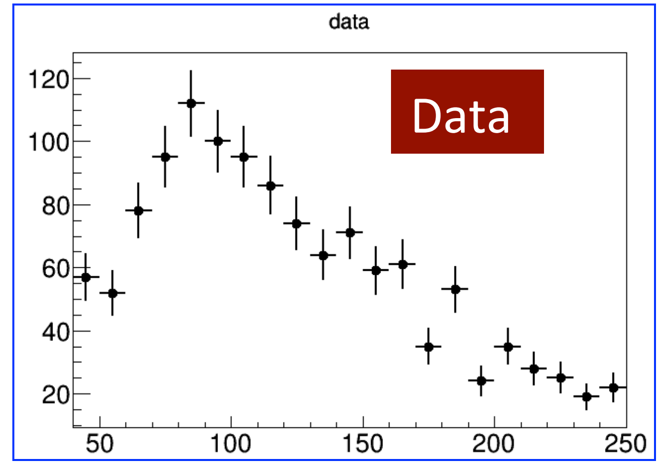
Input Histograms for WH

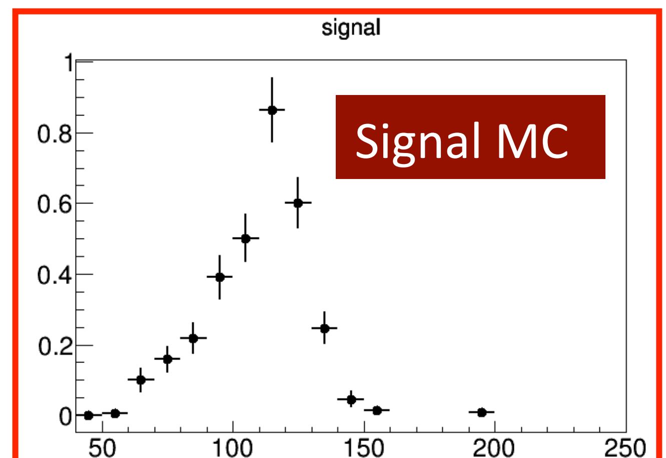
Data, signal, and each backgrounds at the nominal, or systematic variation cases: ./data/WH.120.root

```
root [0]
Attaching file WH.120.root as _file0...
root [1] .ls
TFile**         WH.120.root
 TFile*         WH.120.root
  KEY: TH1D     signal;1        signal
  KEY: TH1D     Z;1     Z
  KEY: TH1D     W;1     W
  KEY: TH1D     WAlt;1  WAlt
  KEY: TH1D     WZ;1    WZ
  KEY: TH1D     WW;1    WW
  KEY: TH1D     Ttbar;1 Ttbar
  KEY: TH1D     multijet;1      multijet
  KEY: TH1D     tot;1   tot
  KEY: TH1D     data;1  data
```

```
root [2]  data->Draw()
Info in <TCanvas::MakeDefCanvas>:
root [3] signal->Draw()
```

Mbb (GeV)





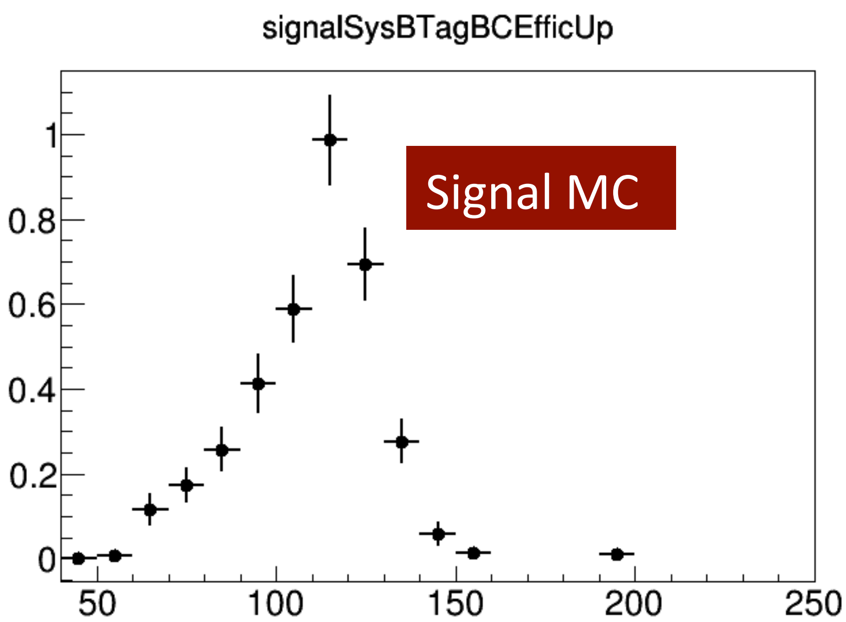
WH, H->bb: Systematic Uncertainties

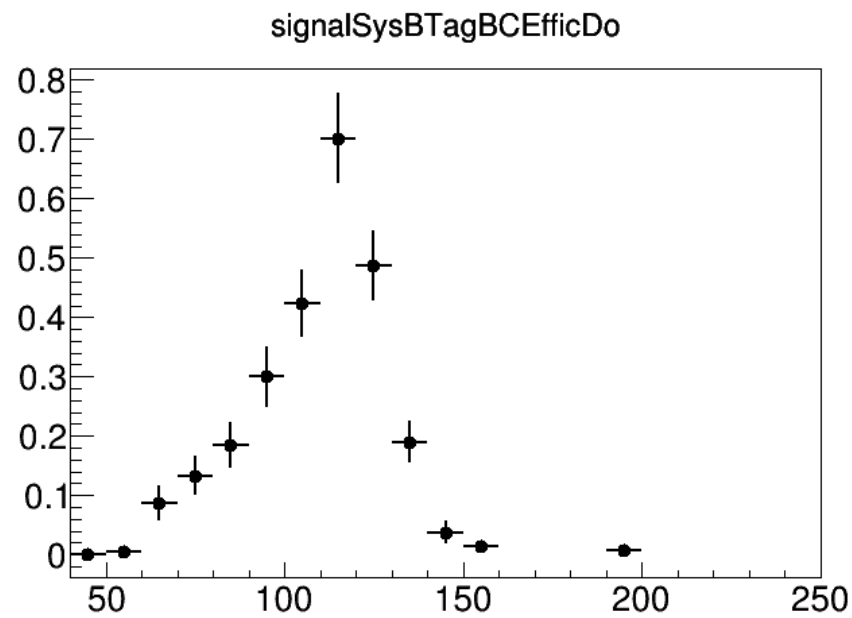
- ✧ Uncertainties on the simulated backgrounds
  - ✓ Main uncertainties: jet multiplicities, p_T^V distributions, flavor composition, and the m_{bb} distributions
- ✧ Experimental uncertainties
  - ✓ Trigger, object reconstruction and identification, energy and momentum calibration and resolution: JES and b-tagging efficiencies are the largest ones
  - ✓ Lumi: 1.8% for 7 TeV, and 2.8% for 8 TeV
- ✧ Theoretical uncertainties on the signal: PDFs, BR, renormalization and factorization and acceptance

All uncertainties considered as Normalization uncertainty (Overall), or shape uncertainty (HistoSys), or both of them

Input Histograms (Btag Uncert.)

```
KEY: TH1D     signalSysBTagBCEfficUp;1         signalSysBTagBCEfficUp
KEY: TH1D     TtbarSysBTagBCEfficUp;1 TtbarSysBTagBCEfficUp
KEY: TH1D     ZSysBTagBCEfficUp;1     ZSysBTagBCEfficUp
KEY: TH1D     WZSysBTagBCEfficUp;1    WZSysBTagBCEfficUp
KEY: TH1D     WWSysBTagBCEfficUp;1    WWSysBTagBCEfficUp
KEY: TH1D     signalSysBTagBCEfficDo;1         signalSysBTagBCEfficDo
KEY: TH1D     TtbarSysBTagBCEfficDo;1 TtbarSysBTagBCEfficDo
KEY: TH1D     ZSysBTagBCEfficDo;1     ZSysBTagBCEfficDo
KEY: TH1D     WZSysBTagBCEfficDo;1    WZSysBTagBCEfficDo
KEY: TH1D     WWSysBTagBCEfficDo;1    WWSysBTagBCEfficDo
```





Analysis Description---XML File (1)

```
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>
  <Channel Name="WH_120" InputFile="./data/WH.120.root" HistoPath="" >
    <Data HistoName="data"/>
    <Sample Name="signal" HistoName="signal" NormalizeByTheory="True" >
       <HistoSys Name="JER" HistoNameHigh="signalSysJetEResol" HistoNameLow="signal"/>
       <HistoSys Name="EES" HistoNameHigh="signalSysElecEUp" HistoNameLow="signalSysElecEDo"/>
       <HistoSys Name="M_EFF" HistoNameHigh="signalSysMuonEfficUp" HistoNameLow="signalSysMuonEfficDo"/>
       <HistoSys Name="M_RES_MS" HistoNameHigh="signalSysMuonEResolMSUp" HistoNameLow="signalSysMuonEResolMSDo"/>
       <HistoSys Name="E_EFF" HistoNameHigh="signalSysElecEfficUp" HistoNameLow="signalSysElecEfficDo"/>
       <HistoSys Name="E_RES" HistoNameHigh="signalSysElecEResolUp" HistoNameLow="signalSysElecEResolDo"/>
       <HistoSys Name="M_RES_ID" HistoNameHigh="signalSysMuonEResolIDUp" HistoNameLow="signalSysMuonEResolIDDo"/>
       <HistoSys Name="MES" HistoNameHigh="signalSysMETEUp" HistoNameLow="signalSysMETEDo"/>
       <OverallSys Name="signorm" Low="0.95" High="1.05"/>
       <NormFactor Name="SigXsecOverSM" Val="1" Low="0." High="50." Const="True" />
    </Sample>
```

- L1: dtd file defines how to read the xml file for a give root version
- L2: channel name should be different for different channels
- L3: observed data
- L4: start to define the signal sample
- L5: shape uncertainty considered for JER
- 3rd last line: normalization uncertainty from cross section
- 2nd last line: parameter of interest (mu)

Analysis Description---XML File (2)

```xml
    <Sample Name="multijet" HistoName="multijet" NormalizeByTheory="False" >
       <OverallSys Name="mjet2enorm" Low="0.5" High="1.5"/>
    </Sample>
    <Sample Name="W" HistoName="W" NormalizeByTheory="False" >
       <HistoSys Name="WAlt" HistoNameHigh="WAlt" HistoNameLow="W"/>
       <OverallSys Name="wnorm" Low="0.79" High="1.21"/>
    </Sample>
  </Channel>
```

- Multijet, and W+jets backgrounds: data-driven estimation, only normalization uncertainty
- How to code the xml file depends on:
  - How each background components affect the results
  - Real effect on the final result for a give syst. Uncert.
  - Physics meaning of each systematic uncertainty

Analysis Description: Driver XML

```
<Combination OutputFilePrefix="./results/WH_combined_weight_H120" >

 <Input>./xmlFiles/roostat_input_WH120_674_weight.xml</Input>

 <Measurement Name="AllSYS" Lumi="1.0" LumiRelErr="0.045" >
  <POI>SigXsecOverSM</POI>
 </Measurement>

</Combination>
```

- L1: define the prefix of the name of output, including workspace
- L2: channel to include to build workspace (>=1 channel)
- L3: Luminosity uncertainty
- L4: Parameter of Interest

Now we can build the workspace, and then run two sets of code to get the exclusion limit and p0

Start to run jobs

- ✓ Histograms
- ✓ XML files
- ✓ Workspace
- ✓ Limits
- ✓ p0

Run Jobs: Preparation

- ssh atlas02@202.122.38.63 with password: tutorial
- Make your own work directory: mkdir lma (any name you prefer)
- Go to you work directory: cd lma
- cp -r /home/atlas01/tutorial/Roostats-Example ./ (copy the code to your work directory)
- cd Roostats-Example
- cat README (vi README)

Text in red is commands to excute

- source /home/atlas01/bin/root/bin/thisroot.sh
- root (root is ready to use if you can see the message as below)

```
[atlas01@sl62-vm Roostats-Example]$ root -l
*** DISPLAY not set, setting it to 222.173.43.54:0.0
root [0]
```

Run Jobs: XML File (1)

```
[atlas01@sl62-vm Roostats-Example]$ cat makeXML.sh
#!/bin/bash

# for mass in 110 115 120 125 130 140
 for mass in 120
 do

    echo ${mass}

    source Template_XML_WH.sh WH ${mass} 674

 done
```

Higgs mass points

Template of xml file

674: Useless now, used for overall uncertainties

./makeXML.sh (to get the file xmlFiles/roostat_input_WH120_674_weight.xml)

Now let's take a look at Template_XML_WH.sh

Run Jobs: XML file (2)

```
#!/bin/bash

channel=$1
mass=$2
lumi=$3
shift 3

cat > xmlFiles/roostat_input_${channel}${mass}_${lumi}_weight.xml << EOF
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>
  <Channel Name="${channel}_${mass}" InputFile="./data/WH.${mass}.root" HistoPath="" >
    <Data HistoName="data"/>
    <Sample Name="signal" HistoName="signal" NormalizeByTheory="True" >
       <HistoSys Name="JER" HistoNameHigh="signalSysJetEResol" HistoNameLow="signal"/>
       <HistoSys Name="EES" HistoNameHigh="signalSysElecEUp" HistoNameLow="signalSysElecEDo"/>
       <HistoSys Name="M_EFF" HistoNameHigh="signalSysMuonEfficUp" HistoNameLow="signalSysMuonEfficDo"/>
       <HistoSys Name="M_RES_MS" HistoNameHigh="signalSysMuonEResolMSUp" HistoNameLow="signalSysMuonEResolMSDo"/>
       <HistoSys Name="E_EFF" HistoNameHigh="signalSysElecEfficUp" HistoNameLow="signalSysElecEfficDo"/>
       <HistoSys Name="E_RES" HistoNameHigh="signalSysElecEResolUp" HistoNameLow="signalSysElecEResolDo"/>
       <HistoSys Name="M_RES_ID" HistoNameHigh="signalSysMuonEResolIDUp" HistoNameLow="signalSysMuonEResolIDDo"/>
       <HistoSys Name="MES" HistoNameHigh="signalSysMETEUp" HistoNameLow="signalSysMETEDo"/>
       <OverallSys Name="signorm" Low="0.95" High="1.05"/>
       <NormFactor Name="SigXsecOverSM" Val="1" Low="0." High="50." Const="True" />
    </Sample>
```

Most cases one sys. has an overall uncert. if considered as a shape uncert.

```
    <Sample Name="multijet" HistoName="multijet" NormalizeByTheory="False" >
       <OverallSys Name="mjet2enorm" Low="0.5" High="1.5"/>
    </Sample>
    <Sample Name="W" HistoName="W" NormalizeByTheory="False" >
       <HistoSys Name="WAlt" HistoNameHigh="WAlt" HistoNameLow="W"/>
       <OverallSys Name="wnorm" Low="0.79" High="1.21"/>
    </Sample>
  </Channel>

EOF
```

Run Jobs: Workspace (1)

```
[atlas01@sl62-vm Roostats-Example]$ cat MakeWorkspace.sh
#!/bin/bash

  ####### combined for WH
  thisCombinedXML=WH_comb_weight.xml

#  for mass in 110 115 120 125 130 140
  for mass in 120

  do

  sed -e "s/H120/H${mass}/g" ./config/${thisCombinedXML} >./config/WH_MA${mass}.xml
  hist2workspace  ./config/WH_MA${mass}.xml

  done
```

./config/WH_comb_weight.xml is template for each mass point
./MakeWorkspace.sh to generate the workspace under ./results/

Run Jobs: Workspace (2)

If it works successfully, you will see
./results/WH_combined_weight_H120_combined_AllSYS_model.root

root results/WH_combined_weight_H120_combined_AllSYS_model.root

```
root [0]
Attaching file results/WH_combined_weight_H120_combined_AllSYS_model.root as _file0...

RooFit v3.60 -- Developed by Wouter Verkerke and David Kirkby
                Copyright (C) 2000-2013 NIKHEF, University of California & Stanford Univ
                All rights reserved, please read http://roofit.sourceforge.net/license.t

root [1] .ls
TFile**         results/WH_combined_weight_H120_combined_AllSYS_model.root
 TFile*         results/WH_combined_weight_H120_combined_AllSYS_model.root
  KEY: RooWorkspace     combined;1      combined
  KEY: TProcessID       ProcessID0;1    7dfb3c52-2619-11e4-9717-1ec36acabeef
  KEY: TDirectoryFile   WH_120_hists;1  WH_120_hists
  KEY: RooStats::HistFactory::Measurement       AllSYS;1
root [2]
```

Run Jobs: Workspace (3)

```
root [2] combined->Print()

RooWorkspace(combined) combined contents

variables
---------
(Lumi,SigXsecOverSM,alpha_EES,alpha_E_EFF,alpha_E_RES,alpha_JER,alpha_MES,alpha_M_EFF,alpha_M_
lpha_mjet2enorm,alpha_signorm,alpha_topnorm,alpha_wnorm,alpha_wznorm,alpha_znorm,binWidth_obs_
,binWidth_obs_x_WH_120_2,binWidth_obs_x_WH_120_3,binWidth_obs_x_WH_120_4,binWidth_obs_x_WH_120
Cat,nom_alpha_EES,nom_alpha_E_EFF,nom_alpha_E_RES,nom_alpha_JER,nom_alpha_MES,nom_alpha_M_EFF,
_MS,nom_alpha_WAlt,nom_alpha_mjet2enorm,nom_alpha_signorm,nom_alpha_topnorm,nom_alpha_wnorm,no
inalLumi,obs_x_WH_120,weightVar)

p.d.f.s
-------
```

```
datasets
--------
RooDataSet::asimovData(obs_x_WH_120,weightVar,channelCat)
RooDataSet::obsData(channelCat,obs_x_WH_120)
```

```
generic objects
---------------
RooStats::ModelConfig::ModelConfig

root [3]
```

Start to Run Jobs: dojob.sh

vi Asymptotic/dojob.sh

```
#!/bin/bash

for ich in WH
do

   for imass in 120
   do

        echo $imass  $imode $iread

       ./runAsymptoticsCLs ${imass} ${ich}  |tee out_${ich}_${imass}

       done

   done
```

./dojob.sh

- ✓ out_WH_120: log file for fit and calculation information
- ✓ root-files/WH/120.root: limits results saved into 1-D histogram
- ✓ out.root: correlation matrix between nuisance parameters (sys.)
- ✓ UpperLimit_CLs_120_WH.dat: limits in text format

Run Jobs: runAsymptoticsCLs.C

```
int main(int argc, char* argv[]){
  //
  RooMsgService::instance().setGlobalKillBelow(RooFit::FATAL); // lowest message level
  //
  //Argument variables
  //
  imass = (argc>1) ? atoi(argv[1]) : 120;
  if(argc>2) channel=argv[2];
  if(channel == "ZH") chout="ZH";
  if(channel == "WH") chout="WH";
  if(channel == "ZvvH") chout="ZvvH";
  if(channel == "VH3") chout="VH3";

  TString fname="/home/atlas02/lma/Roostats-Example/results/"+channel+Form("_combined_weight_H%i_combined
llSYS_model.root", imass);
```

- ✓ When you want to change the run condition, open the file runAsymptoticsCLs.C, and change the variables in the main() function as above
- ✓ After the change done, re-compile by "make" with root available

Start to Run Jobs: log information

```
vi out_WH_120
```

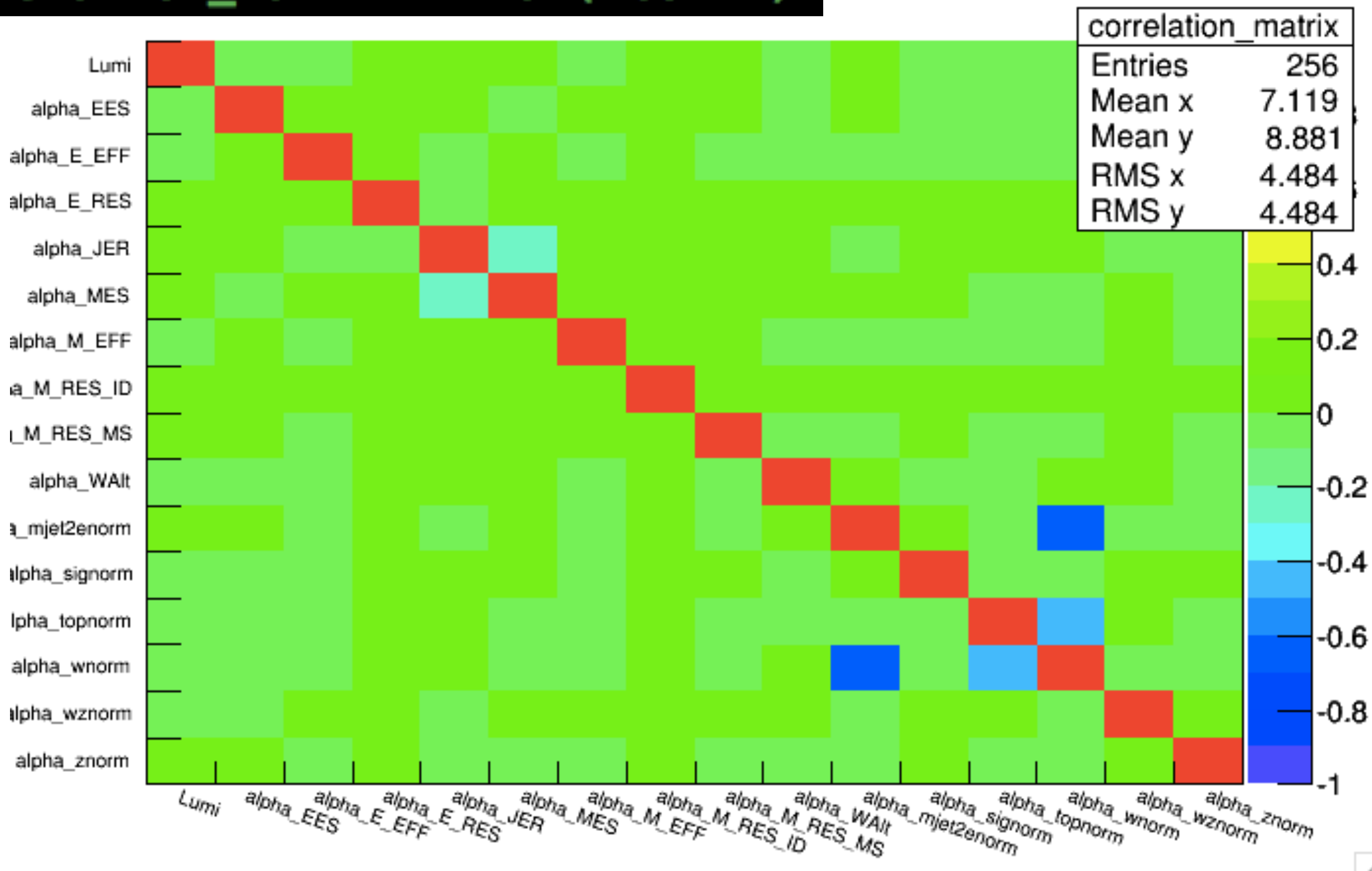
A good fit has to be: CONVERGED, tiny EDM, and ERROR MATRIX ACCURATE

```
MIGRAD MINIMIZATION HAS CONVERGED.
MIGRAD WILL VERIFY CONVERGENCE AND ERROR MATRIX.
FCN=12.5793 FROM MIGRAD    STATUS=CONVERGED     286 CALLS         287 TOTAL
                    EDM=9.4364e-06    STRATEGY= 1       ERROR MATRIX ACCURATE
 EXT PARAMETER                                   STEP         FIRST
 NO.   NAME       VALUE           ERROR          SIZE       DERIVATIVE
  1  Lumi         1.00109e+00   4.49890e-02   3.74247e-05   1.54622e-04
  2  alpha_EES    1.75565e-02   9.00035e-01   4.51804e-04   7.90493e-05
  3  alpha_E_EFF  -3.47326e-02   9.56009e-01   4.75254e-04  -1.02005e-04
  4  alpha_E_RES   5.08975e-02   9.22738e-01   4.62733e-04  -1.15077e-04
  5  alpha_JER    9.49765e-01   5.84757e-01   2.88784e-04  -1.56028e-03
  6  alpha_MES    7.88589e-01   6.84171e-01   3.35757e-04  -9.43437e-04
  7  alpha_M_EFF  -1.03017e-02   9.57437e-01   4.80648e-04   5.44073e-05
  8  alpha_M_RES_ID    9.06848e-03   1.06181e+00   5.34361e-04   2.90441e-06
  9  alpha_M_RES_MS   -3.17478e-02   1.06251e+00   5.34331e-04   4.38587e-06
 10  alpha_WAlt   4.30014e-01   1.06451e+00   5.30581e-04   2.98018e-04
 11  alpha_mjet2enorm  -4.39036e-02   7.94556e-01   2.38160e-04   1.08254e-02
 12  alpha_signorm   0.00000e+00   9.93347e-01   4.99468e-04   0.00000e+00
 13  alpha_topnorm  -4.57727e-01   7.71071e-01   2.63655e-04   6.21784e-04
 14  alpha_wnorm  -5.85347e-02   5.83723e-01   1.49330e-04  -1.38499e-02
 15  alpha_wznorm   5.88429e-02   9.89430e-01   4.96798e-04   2.69978e-05
 16  alpha_znorm   1.31898e-01   9.81142e-01   4.89562e-04   2.27965e-04
```

Start to Run Jobs: Set limit(1)

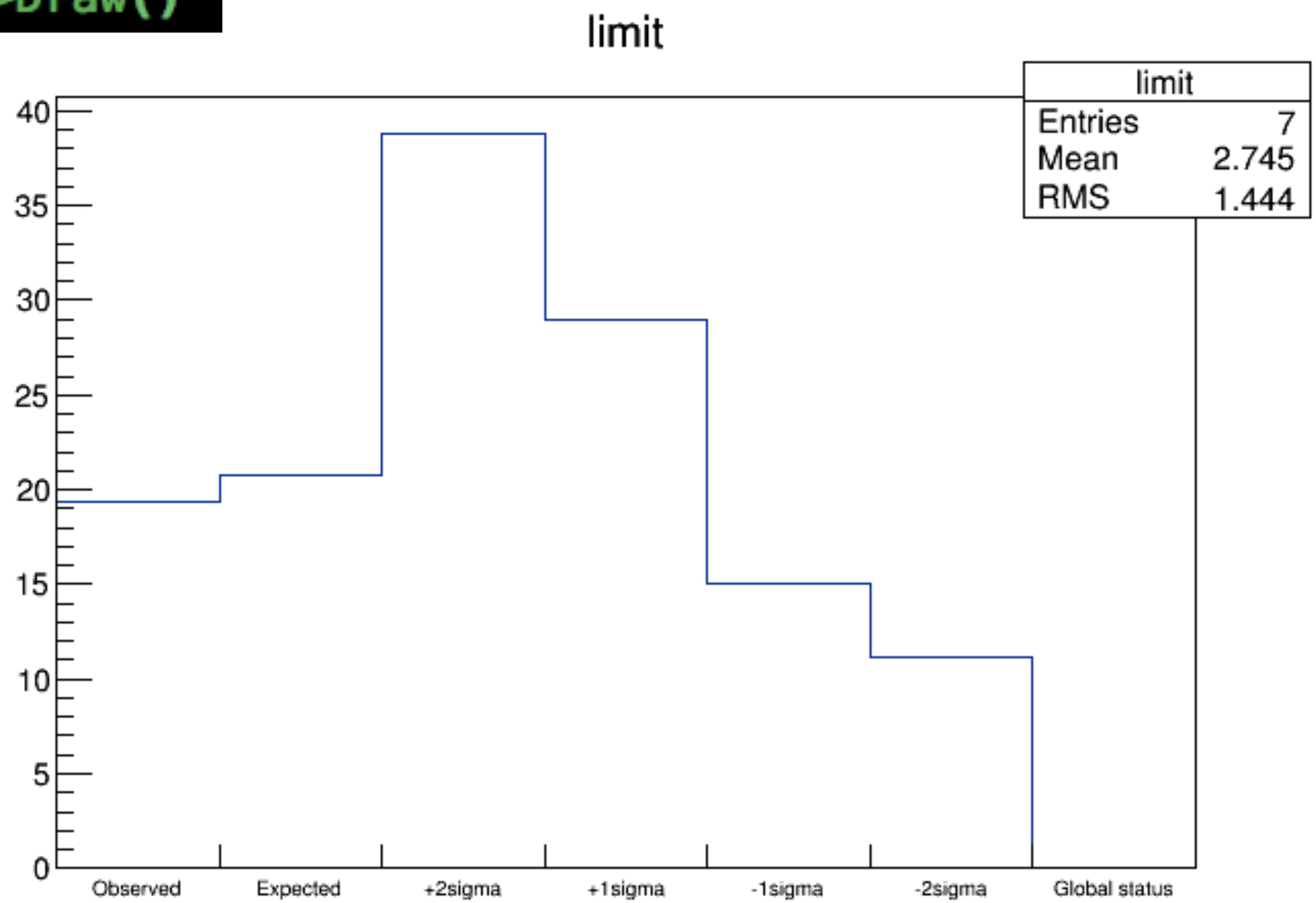
```
[atlas02@sl62-vm Asymptotic]$ root out.root
```

```
root [3] correlation_matrix->Draw("colz")
```



Start to Run Jobs: Set limit(1)

```
root root-files/WH/120.root
root [2] limit->Draw()
```



Tips on Roostats Usage

- ✧ A histogram for shape uncertainty must be normalized to the same entry as the nominal histogram
- ✧ HistFactorySchema.dtd better to use the one under the same root version
- ✧ Not all uncertainties need to be considered, especially to save time (check the effect on the final result, ignore it if its effect is tiny)
- ✧ Check the fit status before you plot the result (not reliable if the fit quality is poor)
- ✧ Well understand the correlation of a sys. Between channels, and the pull of sys. uncertainties

Thanks!

Home work

- ✧ Normalize histograms in WH.120.root for shape uncertainties (not correct usage in this tutorial even the effect is tiny)
- ✧ Check the sys. histograms to figure out the most important shape and overall (normalization uncertainty)
- ✧ Update the xml file with the changes in last two step
- ✧ Re-calculate the upper limit and tell the difference

- ✧ Advanced: make a complete analysis by keep all normalization uncertainty (>1%), and all shape uncertainty (variation >1% for either bin)