Statistical issues and analysis methods in CEPC Higgs Physics

Yaquan Fang (IHEP), Zhang Kaili (IHEP), Jin Wang (Sydney)

4th CEPC physics and simulation workshop
06/27-06/29,2018

Statistics computations needed in CEPC

- Signal Significance : discovery
- Exclusion : limit setting (95% CL)
- Precision measurement

Simple way to do that :

s/sqrt(b), s/sqrt(s+b) sqrt(s+b)/s

Low statistic case:

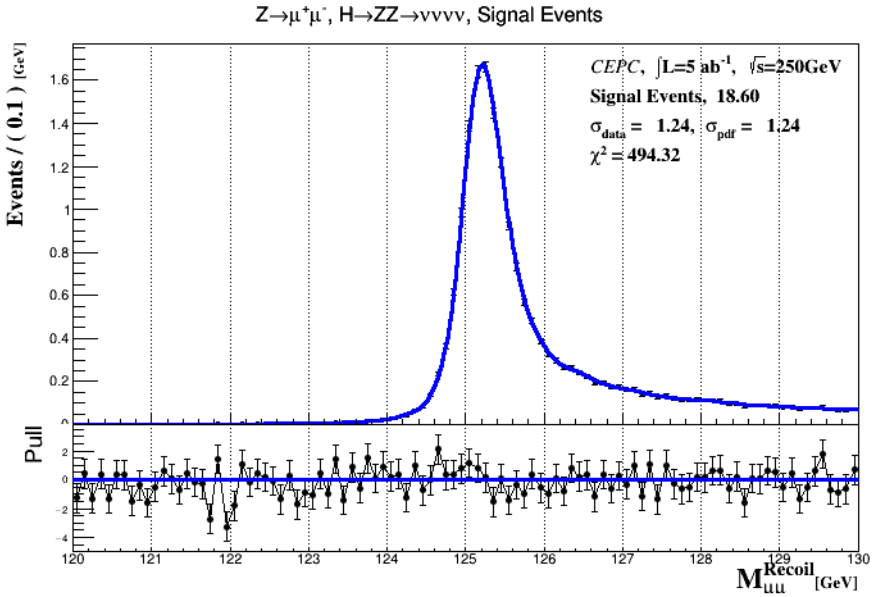
sqrt(2(s+b)*ln(1+s/b))

($5ab^{-1}$)	Pre_CDR	Current 2018.6	ILC 250	Fcc-ee
$\sigma(ZH)$	0.51%	**0.50%**	1.2%	0.40%
$\sigma(ZH) * \mathrm{Br}(H \to bb)$	0.28%	**0.28%**	0.6%	0.2%
$\sigma(ZH) * \mathrm{Br}(H \to cc)$	2.2%	**3.3%**	3.9%	1.2%
$\sigma(ZH) * \mathrm{Br}(H \to gg)$	1.6%	**1.3%**	3.3%	1.4%
$\sigma(ZH) * \mathrm{Br}(H \to WW)$	1.5%	**1.1%**	3.0%	0.9%
$\sigma(ZH) * \mathrm{Br}(H \to ZZ)$	4.3%	**5.1%**	8.4%	3.1%
$\sigma(ZH) * \mathrm{Br}(H \to \tau\tau)$	1.2%	**0.8%**	2.0%	0.7%
$\sigma(ZH) * \mathrm{Br}(H \to \gamma\gamma)$	9.0%	**8.2%**	16%	3.0%
$\sigma(ZH) * \mathrm{Br}(H \to \mu\mu)$	17%	**16%**	46.6%	13%
$\sigma(\nu\nu H) * \mathrm{Br}(H \to bb)$	2.8%	**3.1%**	11%	2.4%
$\mathrm{Br}_{\mathrm{upper}}(H \to \mathrm{inv.})$	0.28%	**0.42%**	0.4%	0.50%
$\sigma(ZH) * \mathrm{Br}(H \to Z\gamma)$	\	**4σ(21%)**		

Fancy ways :

- Fit method taking into account the shape information, correlation,
- Combined fit…..
- RooFit, Roostats …..

Number counting vs fitting method



Z->mm	signal	bkg	s/b	$\sqrt{s+b}/s$	fit
120-150	23.69	36540	0.0006	807%	242%
120-130	18.60	8802	0.0021	505%	252%
124-130	18.44	5644	0.0032	408%	253%
124-126	13.04	1793	0.0072	326%	241%

Establish the model and Likelihood ratio

1. **A likelihood L(Θ) is built :**

$$f(n_{cb}, a_p \mid \phi_p, \alpha_p, \gamma_b) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}} \text{Pois}(n_{cb}|\nu_{cb}) \cdot G(L_0|\lambda, \Delta_L) \cdot \prod_{p \in \mathbb{S}+\Gamma} f_p(a_p|\alpha_p)$$

Shape info. for the discriminating Variables considered.

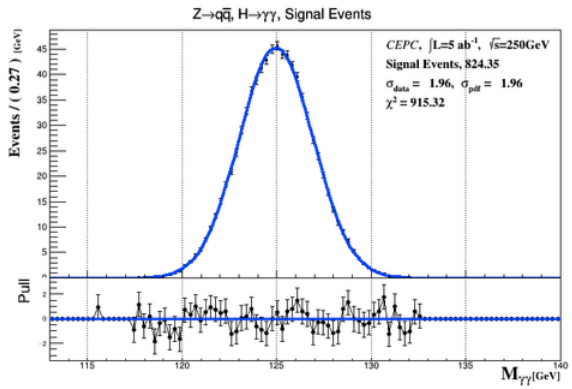
Likelihood as usual → $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta})$

2. **A profile likelihood ratio λ(μ) (μ signal strength :σ▪Br/(σ▪Br)$_{SM}$) is constructed to estimate the parameters of interest:**

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})} \qquad 0 \leq \lambda \leq 1$$

μ is the test hypothesis (μ=0,1 correspond to bkg /signal+bkg only hypothesis)

Sometimes, complicated functions or sum of functions has to choose to model the shape.



$$CB(m_{\gamma\gamma}) = N \times \begin{cases} e^{-t^2/2} & \text{if } -\alpha_{low} \leqslant t \leqslant \alpha_{high} \\ \dfrac{e^{-\frac{1}{2}\alpha_{low}^2}}{\left[\frac{1}{R_{low}}(R_{low} - \alpha_{low} - t)\right]^{n_{low}}} & \text{if } t < -\alpha_{low} \\ \dfrac{e^{-\frac{1}{2}\alpha_{high}^2}}{\left[\frac{1}{R_{high}}(R_{high} - \alpha_{high} - t)\right]^{n_{high}}} & \text{if } t > \alpha_{high} \end{cases}$$

$$t = (m_{\gamma\gamma} - \mu_{CB})/\sigma_{CB}$$

Fangyi Guo's talk

KEYS PDF:

For some more complicated function, Keys would be another option.

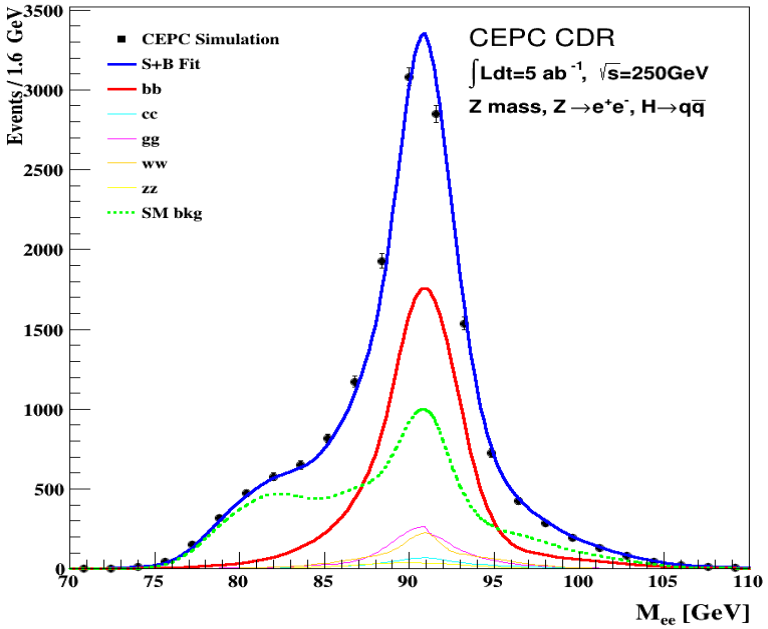
class RooKeysPdf: public RooAbsPdf

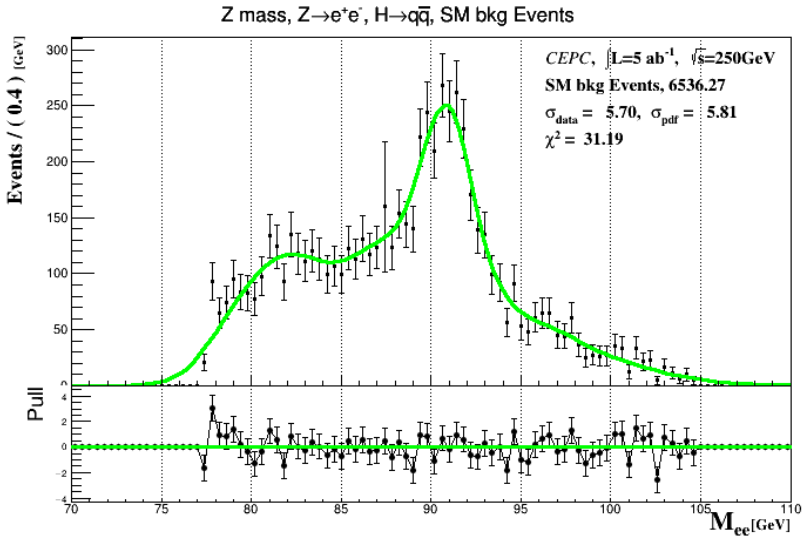
Class RooKeysPdf implements a one-dimensional kernel estimation p.d.f which model the distribution of an arbitrary input dataset as a superposition of Gaussian kernels, one for each data point, each contributing 1/N to the total integral of the p.d.f.

If the 'adaptive mode' is enabled, the width of the Gaussian is adaptively calculated from the local density of events, i.e. narrow for regions with high event density to preserve details and wide for regions with log event density to promote smoothness. The details of the general algorithm are described in the following paper:

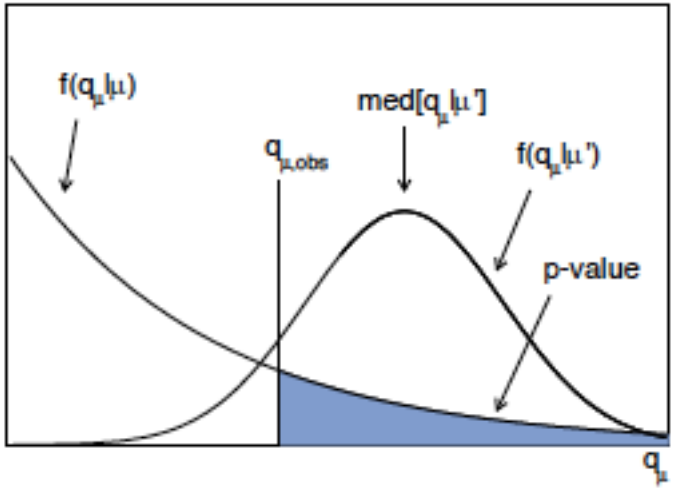
Cranmer KS, Kernel Estimation in High-Energy Physics. Computer Physics Communications 136:198-207,2001 - e-Print Archive: hep ex/0011057

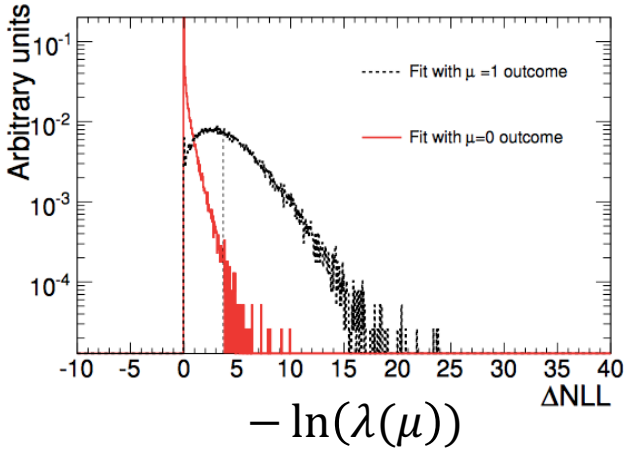
There is one parameter to decide how much fluctuation one wants to pick it up for the model.





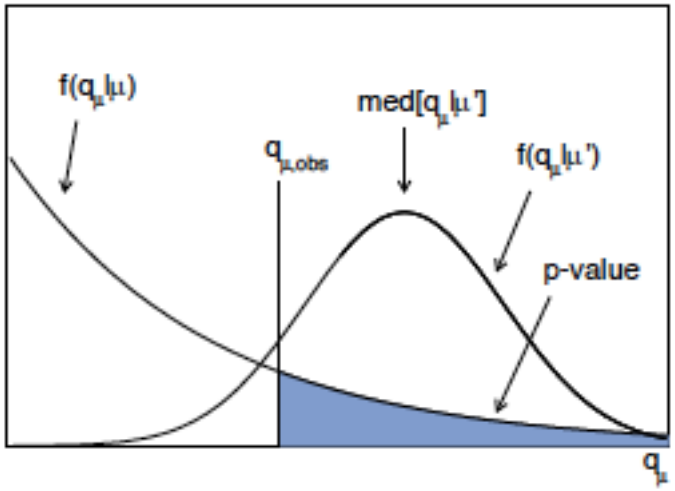
Discovery

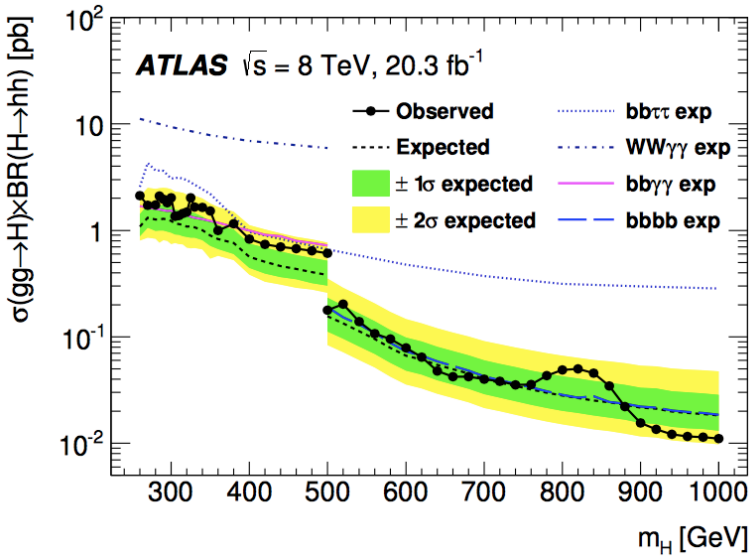




- Exclude the Background-only hypothesis.
  - Construct the test statistics
    - $\lambda(\mu)$, in practice $q(\mu) = -2\ln(\lambda(\mu))$
  - For discovery: exclude signal strength μ=0
  - Need to know how far away $-2\ln(\lambda(1))$ from $-2\ln(\lambda(0))$
  - Throw toys to do that.
  - For expected one, use medium value of $-2\ln(\lambda(1))$ to compute.
  - Integrate the tail of bkg-only hypothesis curve (p-value) and convert it into significance.

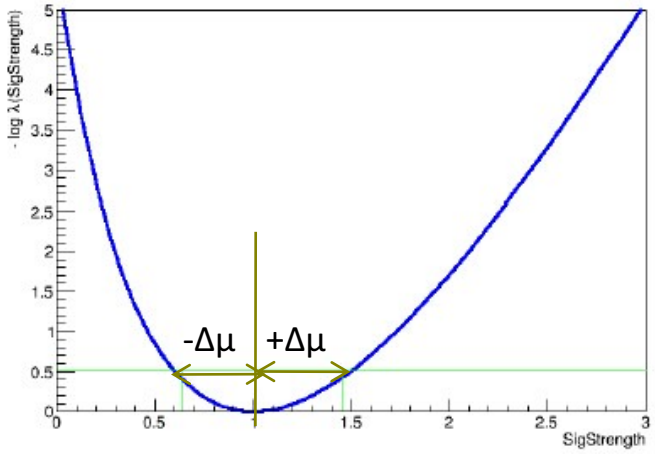
Exclusion





- Exclude the Signal+Background hypothesis.
  - Construct the test statistics
    - $q(\mu) = -2\ln(\lambda(\mu))$
  - For exclusion: exclude signal strength μ=X
  - Need to know how far away $-2\ln(\lambda(0))$ from $-2\ln(\lambda(X))$
  - Throw toys to do that.
  - For expected one, use medium value of $-2\ln(\lambda(0))$ to compute.
  - Integrate the tail of Signal+bkg hypothesis curve and convert it into 95% CL.

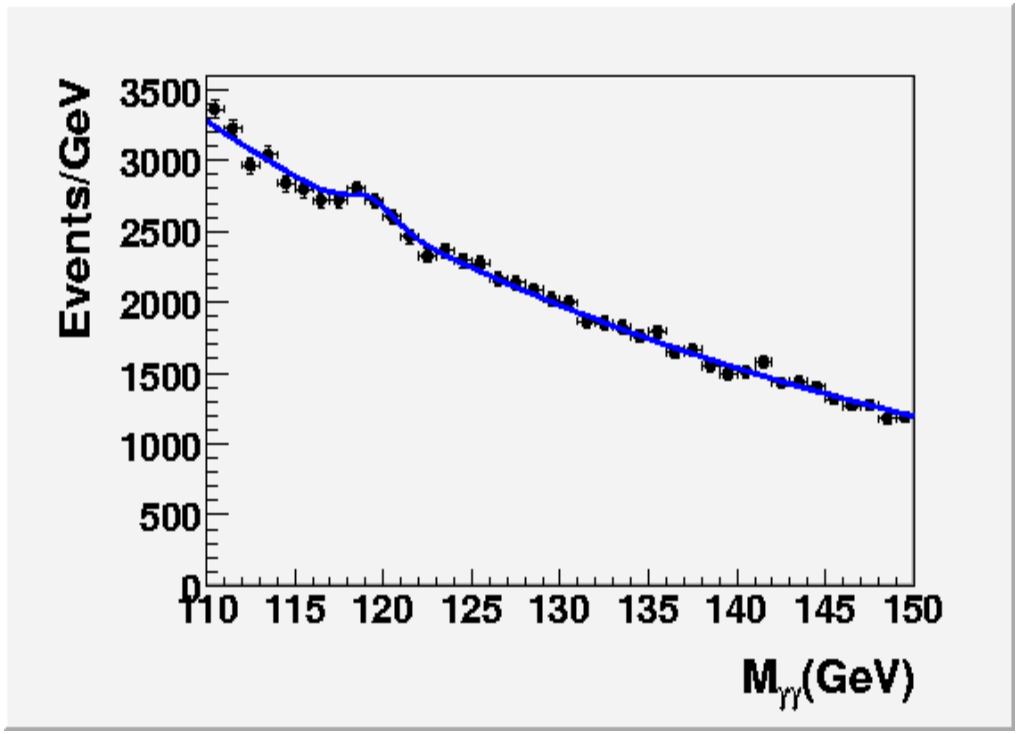
Precision measurement



- One can scan μ (Minuit) and the error of μ is the distance of x-axis between μ=1 and the point on the curve corresponding to –log(λ)= 0.5.
- The uncertainty can be incorporated into the fit.
  - ✓ Currently the luminosity uncertainty, xsection from direct measurement are considered.
- Technically, just one S+B fit on data to extract Δμ from Minos

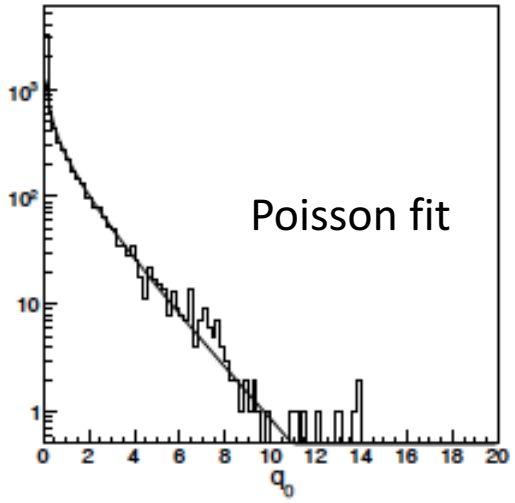
MC toys

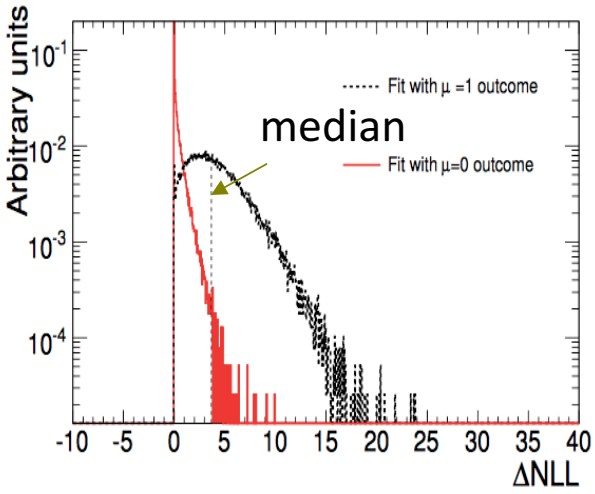
- As one obtain the models for signal+backgrounds, one can throw toys and fit it (maybe many times) to test the statistics.
- Most reliable, however CPU intensive. For example, one needs 1000 CPU X one month to reach 5σ for inclusive ATLAS H->gamma gamma analysis.
  - Use Histogram data
  - Asimov data

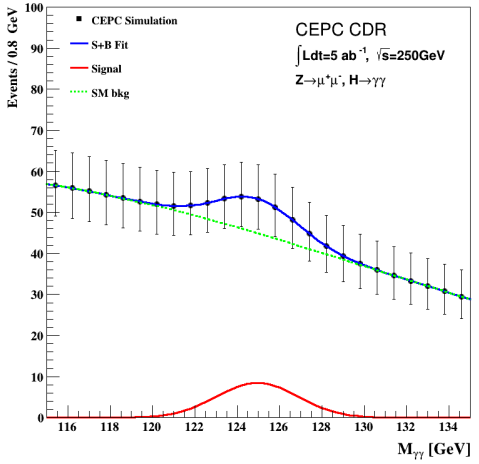


Toys vs asymptotical way

- In the case the hypothesis zero follows a Poisson distribution, one can compute the significance based on hypothesis 1.
  - Don’t need to throw huge toys to obtain a reasonable tail.
  - Significance = sqrt($-2\ln(\lambda(\mu))$ for the median one.
- One can even be more aggressive to generate one Asimov data to serve as one MC to obtain the significance of the median value.

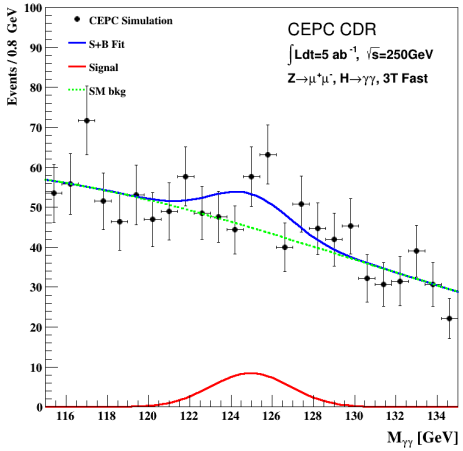


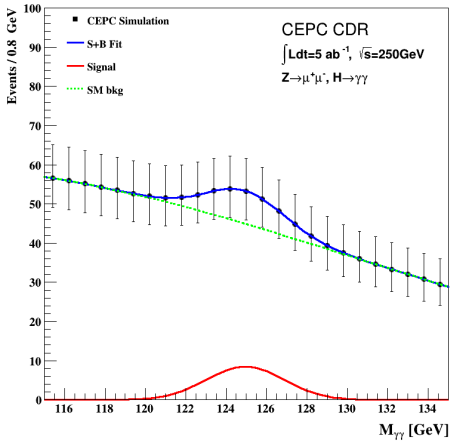




Motivation of Asimov Data

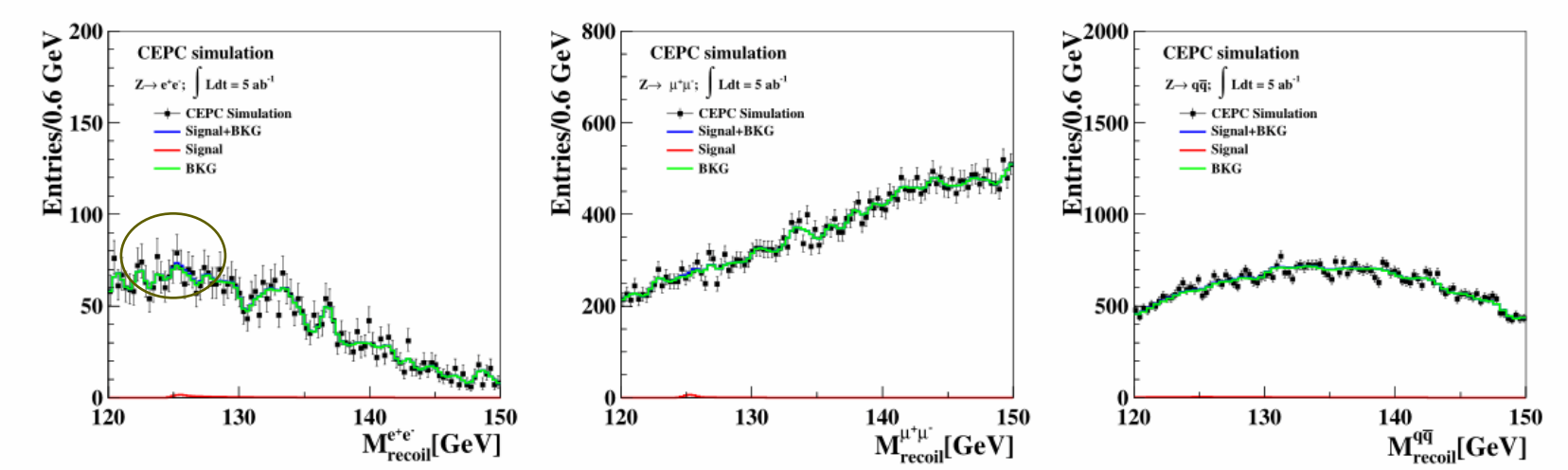
- For the simulation study, one will probably run into the following situation:
  - Only MC for one experiment has been generated (samples with Luminosity 5 ab^{-1} in our case)
  - Sometimes fluctuation due to the limit MC could be picked up when one estimates the expected sensitivities.
  - Asimov data is equivalent to what is produced by unlimited MC normalized to expected luminosity.
  - If we cannot reach the limit of statistics, we can extract the precision using Asimov data (avoid the issue of the limited MC sample having some fluctuation).





Example: H->invisible

- fit range all 120-150GeV



- Consider shape; but using generated data to fit;
- Can be regarded as “one experiment” measurement.
  - Huge bkg-> large fluctuations could ”create” some bump while it is not reflected in bkg model.
  - Central value not 1;

120-150	signal	bkg	s/b	$\sqrt{s+b}/s$	Measured μ
Z->ee	12.86	4205	0.003	505%	$3.30 \pm 481\%$
Z->mm	23.69	36540	0.0006	807%	$3.30 \pm 273\%$
Z->qq	224.41	426540	0.0005	290%	$0.88 \pm 141\%$

Solution:

- Build signal and background models based on the MC samples.
- Use these models to generate Asimov data
- Fit the models on the Asimov data to obtain the expected measured precision.
  - One can also try toys….
- Similar issues happened in H->bb,cc,gg analyses.

	Precision
Z->ee	1.0+/-350%
Z->mm	1.0+/-242%
Z->qq	1.0+/-226%
Combined	1.0+/-148%

From Roofit to Roostats

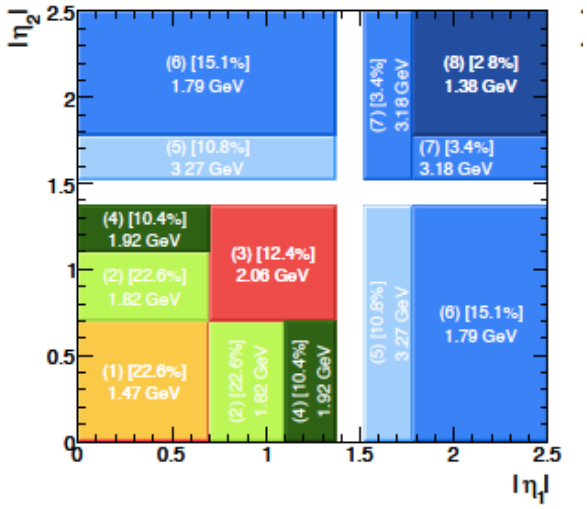
- Based on Root, Roofit provides easy-adapted code to do unbinned/binned fit, toy generation, model production etc…
- Based on Roofit, Roostats provides friendly framework dedicated for the statistic tests.
  - Users follow the examples to prepare the input files and do minimal coding.
  - Widely used in ATLAS/CMS experiments for the statistical study.
  - Some codes are integrated in the package, the users may not have chances to dig into the black box if there is no need.
- If people working on individual channels choose Roostats and prepare the workshops, it is very convenient to do the combination.

Some Analyses strategies in ATLAS

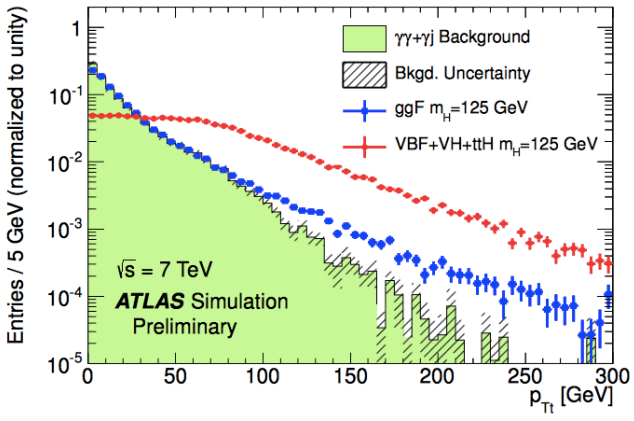
H->gamma gamma analysis

- The inclusive analysis is very simple :
  - Photon ID, Isolation, Kinematic cuts on leading/subleading photon.
- Is there anything else that we can explore?

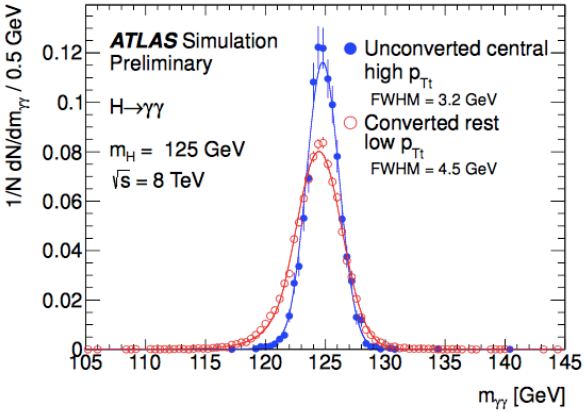
Divide different eta regions for two photons



P_T of Higgs (P_{Tt} is perpendicular to the thrust direction of two photon)



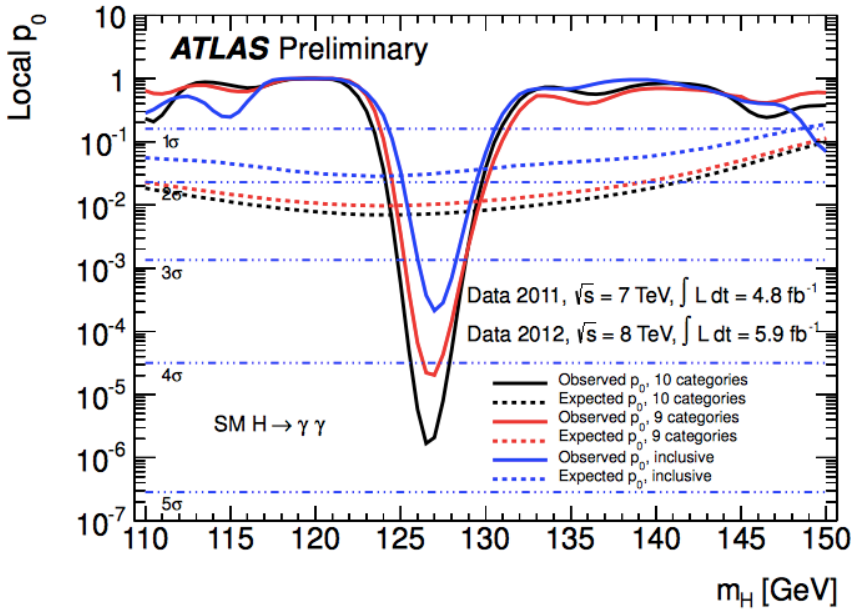
Conversion of the photons



How to use these information?

Category	σ_{CB} [GeV]	FWHM [GeV]	Observed [N_{evt}]	S [N_{evt}]	B [N_{evt}]
Inclusive	1.63	3.87	3693	100.4	3635
Unconverted central, low p_{Tt}	1.45	3.42	235	13.0	215
Unconverted central, high p_{Tt}	1.37	3.23	15	2.3	14
Unconverted rest, low p_{Tt}	1.57	3.72	1131	28.3	1133
Unconverted rest, high p_{Tt}	1.51	3.55	75	4.8	68
Converted central, low p_{Tt}	1.67	3.94	208	8.2	193
Converted central, high p_{Tt}	1.50	3.54	13	1.5	10
Converted rest, low p_{Tt}	1.93	4.54	1350	24.6	1346
Converted rest, high p_{Tt}	1.68	3.96	69	4.1	72
Converted transition	2.65	6.24	880	11.7	845
2-jets	1.57	3.70	18	2.6	12

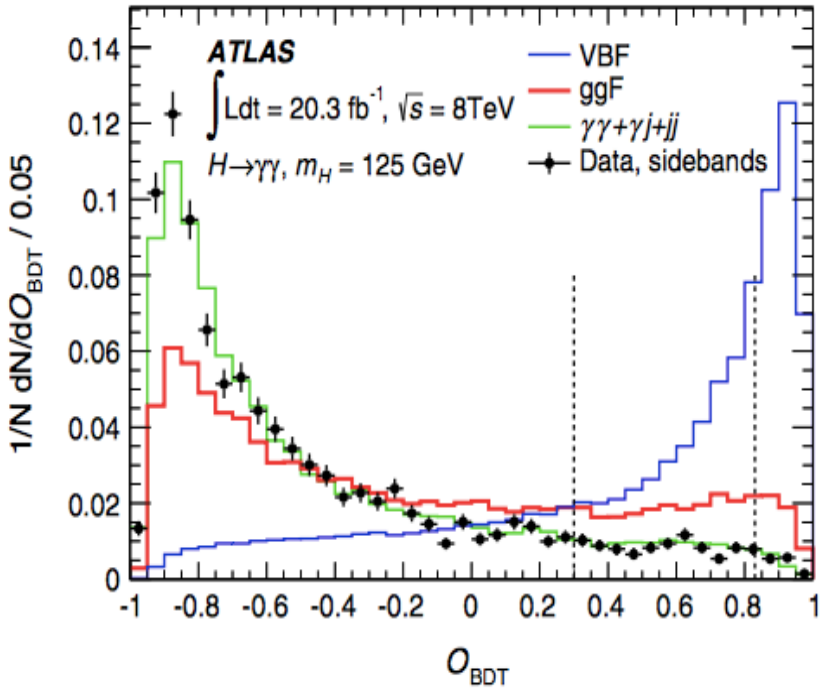
Results contributed to Higgs discovery



- Instead of throwing away events, we divide them into different categories according different S/B:
  - PTt
  - Conversions
  - Resolution regions
  - jets
- The improvement of the significance w.rt. to inclusive one is obvious.

Further usage of the BDT outputs

- Again, one example from ATLAS VBF H->gammagamma analysis:
- Instead of implementing one cut on the BDT output, we divide them into different regions, optimizing with combined significance.
  - S/sqrt(B) are different for 2 or 3 BDT output regions.



optimization normalized to 4 fb^{-1}

	MVA tight	MVA loose
VBF	1.64	2.17
ggF	0.51	1.90
background	2.42	17.71
VBF purity	0.76	0.53
significance	0.88	0.47
combined significance	1.00	

Categorization of events according to their final states

ttH in multi-lepton: Analysis Overview

- Signals mainly from H→WW* and H→ττ, small from H→ZZ*
- Signature: 2-4 leptons (τ_{had}), >= 2-jets and >= 1 b-jet
- Main backgrounds: ttW/ttZ from MC but validated to data, non-prompt bkg. (mainly ttbar) is data-driven
- Dominant syst.: estimations for fake lepton and non-prompt bkg.

Lianliang's talk "全国高能物理大会"