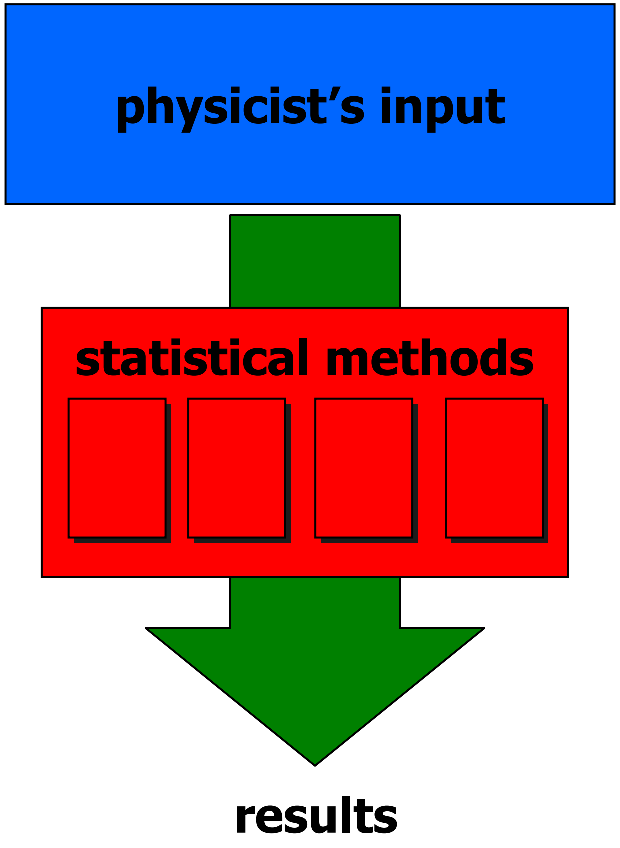
Validation benchmarks and tools

Mingshui Chen
University of Florida

Many thanks to all who contributed to the validation effort that we are discussing today, in particular (in alphabetical order)

Paolo Giacomelli
Andrey Korytov
Luca Lista
Giovanni Petrucciani
Gregory Schott

Practical point of view



Physicist's input, e.g.:

- made-up $H \to WW \to 2l2\nu$ at $L=1\ pb^{-1}$
- syst. errors: all assumed to be lognormal

Statistical methods

- exclusion limits
  - Bayesian
    - flat and 1/sqrt(r) priors
  - Frequentist and Modified Frequentist
    - three test statistics
  - PL approximation
- significance
  - PL approximation
  - From p-value
    - three test statistics

Software

- RooStats (toolkit being validated)
- LandS (reference software package)

Physicist's input (1): HWW

$H \rightarrow WW \rightarrow 2l2\upsilon$ benchmark points

- **made-up model:** numbers used are reasonable, but should not be assumed to represent the actual analysis status
- **4 channels (cut-and-count):** $\mu\mu$, ee, $e\mu$, μe
- **each channel has several separate backgrounds** assumed to be tracked separately either via data-driven measurements or MC
- **more than 30 independent sources of uncertainties** with the full table of correlations within and across channels

Physicist's input (1): HWW

- **HWW benchmark points at 1/fb**
  - m_H=160 GeV:
    - total signal ~ 36, total background ~ 22
    - most sensitive SM Higgs mass point with good S/B-ratio
    - expected exclusion r~0.3, expected significance ~5σ
  - m_H = 140 GeV:
    - total signal ~16, total background ~42
    - the role of systematic errors more pronounced
    - expected exclusion r~1.7, expected significance ~3σ
  - For each mass points we then take a few plausible "experimental outcomes"
    - **background-like:** "observed" event yield is approx. the expected background
    - **undershoot:** an outcome that can be loosely classified as a -2sigma fluctuation
    - **overshoot:** an outcome representing a +2sigma fluctuation,
    - **signal-like:** an outcome that would look like a signal.

Physicist's input (2): one-channel exp.

- **Simplified counting experiment benchmark points**
  - to help understand the differences
  - and trace down any possible issues

<table>
<tr><th>N_{bkg}</th><th>N_{obs}</th><th>reasoning</th><th colspan="5">Systematic errors</th></tr>
<tr><td rowspan="4">5.5</td><td>6</td><td>Observation ~background only</td><td rowspan="4">none</td><td rowspan="4">δb/b ~30%</td><td rowspan="4">δs/s ~30%</td><td rowspan="4">δb/b ~30%
δs/s ~30%
no correl</td><td rowspan="4">δb/b ~30%
δs/s ~30%
100% correl</td></tr>
<tr><td>1</td><td>Downward fluctuation</td></tr>
<tr><td>11</td><td>Upward fluctuation</td></tr>
<tr><td>20</td><td>Significant excess</td></tr>
</table>

Physicist's input (3): uniform input

- **Same "data cards" as an input to RooStats and LandS**
- **Complete map of correlations between errors within and across different channels**
- **Lognormal pdf's for all systematics** (may try more later)

- **Conceptual form is as follows:**

	Bin 1 (channel 1)				Bin i (channel i)			
events observed in experiment ==>	n_1				n_i			
	Signal	Bkgd 1	…	Bkgd j	Signal	Bkgd 1	…	Bkgd j
MC or DataControlSample events ==>	N(0,1)	N(1,1)		N(j,1)	N(0,i)	N(1,i)		N(j,i)
overall scale factor ==>	α(0,1)	α(1,1)		α(j,1)	α(0,i)	α(1,i)		α(j,i)

Systematic Error Sources and Parameters

No.	Uncertainty Source description	pdf typ	Parameters							
			parameters	parameters	parameters	parameters	parameters	parameters	parameters	parameters
1	Luminosity	lnN	1.05	1.05	1.05	-	1.05	1.05	1.05	-
2	Signal cross section x acceptance	lnN	1.10				1.10			
3	Bkgd 1 cross section	lnN		1.30				1.30		
..	…	lnN								
..	Bkgd j (ch1) data-driven from control region: dw/w	lnN				1.10				
..	Bkgd j (ch2) data-driven from control region: dw/w	lnN								1.20
..	…									
..	muon Reconstruction Efficiency (2%)	lnN	1.04	1.04	1.04	1.04	1.02	1.02	1.02	1.02
..	electron Reconstruction Efficiency (2%)	lnN					1.04	1.04	1.04	1.04
..	…									

Statistical methods: limits

Method	Options
Bayesian*	flat prior on signal strength ***r***
	1/sqrt(***r***) prior
Modified Frequentist (CL_s)*	no "fitting" in test statistics
	with "fitting" for syst. errors
	with "fitting" for syst. errors and signal strength
Frequentist (CL_{s+b})*	no "fitting" in test statistics
	with "fitting" for syst. errors
	with "fitting" for syst. errors and signal strength
Profile Likelihood	

*** Description of these methods are in back-up slides**

Statistical methods: significance

Method	Options
Hybrid Bayesian-Frequentist (CL_b)	no "fitting" in test statistics
	with "fitting" for syst. errors
	with "fitting" for syst. errors and signal strength
Profile Likelihood	

Validation tool: LandS

LandS: Limits and Significance

- **Source and instructions:** https://cern.ch/mschen/lands/

- **Standalone package: desn't depend on ROOT, except for minuit library and final plotting**

Can handle all statistical methods from the previous two slides.

Being fast and accurate, it has been extensively used in the CMS Higgs group over the last year...

What we compare: RooStats vs LandS

Results:

- any systematic shifts?
- computational (stat) precision

Performance:

- computational time (CPU consumption)
- instabilities, memory leaks, …
- ability to insulate a user from internal technicalities

Example

m_H=140 GeV with the "observed" events **consistent** with the expected background-only rate

Technique	Test statistic or Prior	RooStats 5.27.06 (HiggsAnalysis/CombinedLimits V00-03-01)			LandS			Comments
		Limit (r ± δr)	Toys, etc.	timing (CPU GHz)	Limit (r ± δr)	Toys, etc.	timing (CPU GHz)	
Bayesian	flat prior on *r*	MCMC: 1.66 ± 0.10[A] MCMC*: 1.746 ± 0.013[A2] BAT: 1.64 ± ???	100k 200x20k 5*(20+4)k[BAT1]	28min (2.1GHz) 24min (2.3GHz) 20 min (2.4 GHz)	1.709±0.001[MA]	100k	0.3 min (2.6GHz)	
	flat prior, no syst.	MCMC*: 1.589 ± 0.004	25x20k	0.1min	1.5867	1	<1s	
	alternative prior (*1/sqrt(r)*)	MCMS: 1.52	100k	24min	1.534±0.001[MA]	100k	0.3 min (2.6GHz)	
CL_s	no profiling	1.613 ± 0.044[B]	C	57min (2.1GHz)	1.64± 0.02	100K(x5)	1.3 min (2.6GHz)	
	profile syst errors	1.962 ± 0.044	-	270min (2.1GHz)	1.67± 0.03	10k(x4)	~11.5h (2.6GHz)	
	profile syst. and *r*	failed	???	???	1.70±0.03	10k (x4)	~11.5h (2.6GHz)	
CL_{s+b}	no profiling	1.613 ± 0.044[B]	-	70min (2.1GHz)	1.62± 0.02	100K(x4)	1.2 min (2.6GHz)	
	profile syst errors	2.147 ± 0.044	-	218min (2.1GHz)	1.64 ± 0.03	10K(x4)	~14h (2.6GHz)	
	profile syst. and *r*	failed	???	???	1.69 ± 0.04	10k(x4)	11h (2.6)	
PL approx.	n/a	1.861	n/a	<1s	1.860	n/a	1s	

RooStats validation conclusions are in the next talk

What we do not compare

Results obtained by different methods…

We leave this subject for discussions over the next few weeks together with the stat forum gurus

Summary

RooStats validation:

- **performed in comparison to LandS**
- **using a few plausible "experimental outcomes"**

The complete digested summary of our findings is in Giovanni's talk…

Back up

CL_s: simple likelihood ratio Q

Discriminator: simple likelihood ratio (Q)

- n_i **number of observed events in channel** i
- s_i **our best estimate of the expected signal events in channel** i
- b_i **our best estimate of the expected background events in channel** I
- r ***signal strength modifier (common for all channels)***

$$Q = \frac{p(observation \mid b+s)}{p(observation \mid b)} = \frac{\prod_{channels} \frac{(b_i + r \cdot s_i)^{n_i}}{n_i!} e^{-b_i - r \cdot s_i}}{\prod_{channels} \frac{b_i^{n_i}}{n_i!} e^{-b_i}} = e^{-r \cdot S_{TOT}} \cdot \prod_{channels} \left(1 + r \frac{s_i}{b_i}\right)^{n_i}$$

Log-Likelihood Ratio

$$-2 \ln Q = 2 r S_{TOT} - 2 n_i \sum_{channels} \ln\left(1 + r \frac{s_i}{b_i}\right)$$

The other two test statistics:

Ratio of profiled likelihoods (with "fitting" for syst. errors)

$$Q_{TEV} = L_{s+b}(\mu = 1, \hat{\hat{\nu}}) / L_b(\mu = 0, \hat{\hat{\nu}}')$$

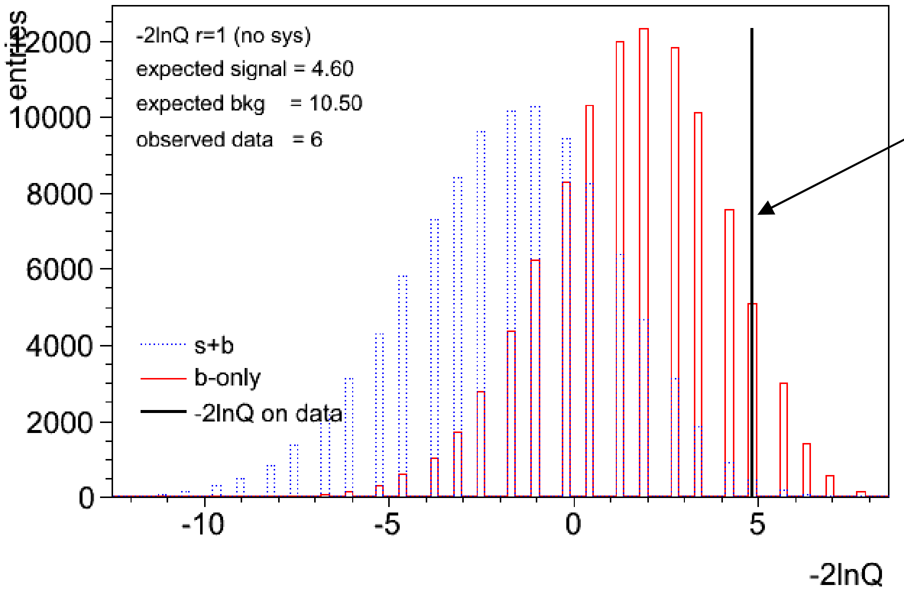
Profile likelihood ratio (with "fitting" for syst. errors and signal strength)

$$\lambda(\mu) = L_{s+b}(\mu, \hat{\hat{\nu}}) / L_{s+b}(\hat{\mu}, \hat{\nu})$$

CL_s: -2lnQ → CL_s

1. Throw **10^5** pseudo-experiments according to **background-only** hypothesis
2. Throw **10^5** pseudo-experiments according to **signal+background** hypothesis
3. Build -2lnQ distributions

Example for single channel: s=4.6, b=10.5, observed n=6



$-2\ln Q_d$ for n=6

$CL_b = P(-2\ln Q \geq -2\ln Q_d)$ cumulative probability in bkgd-only distribution

$CL_{sb} = P(-2\ln Q \geq -2\ln Q_d)$ cumulative probability in signal+bkgd distribution

$CL_s = CL_{sb}/CL_b = \alpha$ ratio of two probabilities from above (make your bets!)

When α is small, say that the signal is excluded with 1-α confidence level

(this is known to be on a conservative side from the true coverage)

CL_s: Tune r for the 95% C.L. exclusion

Measure CL_s for the first trial value r

If CL_s is far from the desired 0.05 (we use a $\pm$0.001 tolerance band),

- **modify r and repeat an exercise of 10^5 pseudo-experiments (previous two slides)**
- **keep doing this until we get CL_s = 0.05 within the tolerance band**

r obtained at the end of the loop is the r excluded at 95% C.L.

Same tuning technique applied to Frequentist approach (CL_{sb})

CL_s: Including systematic errors

- **Assign systematic errors to each b_i and s_i**
  (this implies a particular pdf; we now use the log-normal pdf)
- **Assign correlations of errors**

- **Before throwing each of the intended 10^5 pseudo-experiments, modify b_i and s_i according to the assigned errors and their correlations. Use modified b_i and s_i to generate pseudo-data n_i**

- **For each of pseudo-experiments, calculate -2lnQ as before, i.e. using un-modified b_i and s_i** (these are our best estimates)

- **All the rest is exactly the same as before**

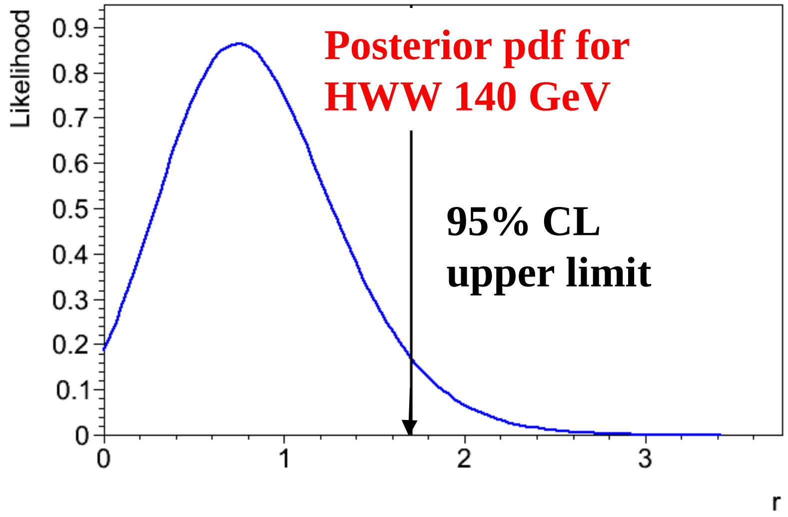
Bayesian: likelihood function

- **Assume the prior on *r* is flat $\pi(r)$=const and build the likelihood function as**

$$L(r) = \frac{p(\vec{n}\,|\,\vec{b}+r\vec{s})\cdot\pi(r)}{\int_0^{+\infty} p(\vec{n}\,|\,\vec{b}+r\vec{s})\cdot\pi(r)\cdot dr} = \frac{p(\vec{n}\,|\,\vec{b}+r\vec{s})}{\int_0^{+\infty} p(\vec{n}\,|\,\vec{b}+r\vec{s})\cdot dr}, \quad \text{where } p(\vec{n}\,|\,\vec{b}+r\vec{s}) = \prod_{channels} \frac{(b_i+rs_i)^{n_i}}{n_i\,!} e^{-rs_i}$$

- **Exclusion limit is obtained from**

$$\int_r^{+\infty} L(r)dr = \alpha \quad \text{(e.g. for 95\%CL } \alpha\text{=0.05)}$$



Bayesian: Including systematic errors

- **Assign systematic errors to each b_i and s_i**
  (this implies a particular pdf; we now use the log-normal pdf)
- **Assign correlations of errors**

- **Throw 10^5 set of b_i and s_i according to the assigned errors and their correlations.**

- **At each value of r evaluated, doing 10^5 integrations and average over them**

- **Tuning r and repeat the previous step to get exclusion limit**