Combine Tutorial

Terascale Statistics School

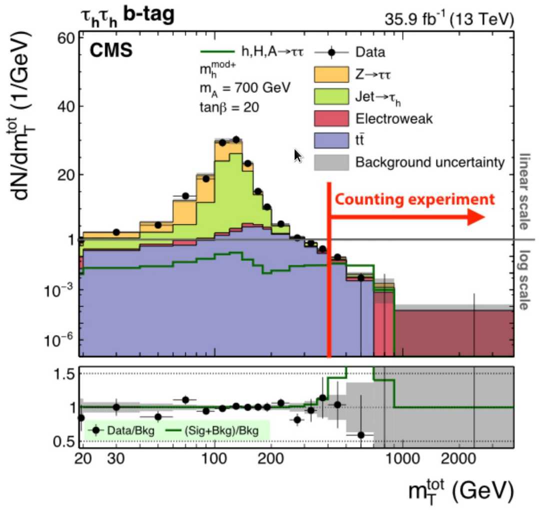
Kyle Cormier, Aliya Namigova, Nick Wardle

Introduction

Overview

Search for a heavy neutral higgs,
Which decays to $\tau\tau$

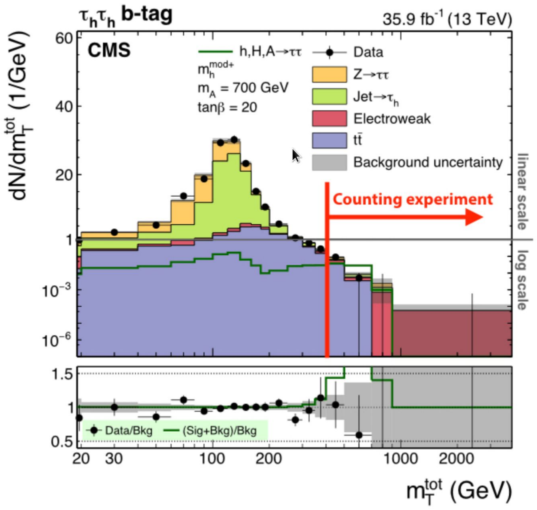
1. Simple Counting Experiment
2. Shape Based Analysis
3. Adding Control Regions
4. Physics models – beyond a single Signal strength parameter



Overview

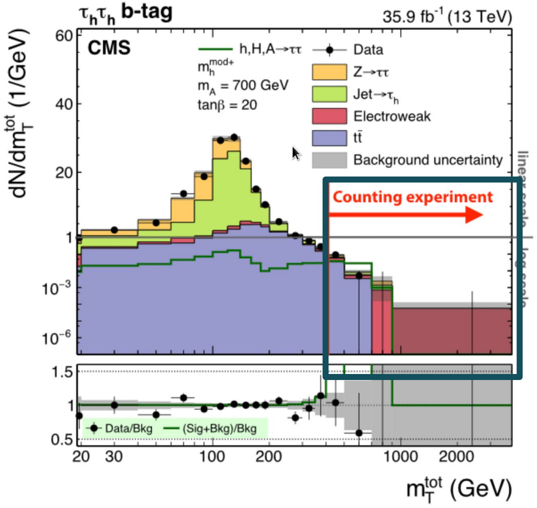
Search for a heavy neutral higgs,
Which decays to $\tau\tau$

- Limit Setting
- Significance Testing
- Asymptotic Calculations
- Toy-based Calculations
- Parameter Extraction
- Fit Debugging
- Fit Plotting
- Nuisance Parameter Checks
- Multi-Dimensional Likelihood Scans



Part 1 – Counting Experiment

Counting Experiment



Count number of events in high m_T region

```
imax     1 number of bins
jmax     4 number of processes minus 1
kmax     * number of nuisance parameters
--------------------------------------------------------------------------------
--------------------------------------------------------------------------------
bin              signal_region
observation      10.0
--------------------------------------------------------------------------------
bin                          signal_region  signal_region  signal_region  signal_region  signal_reg
process                      ttbar          diboson        Ztautau        jetFakes       bbHtautau
process                      1              2              3              4              0
rate                         4.43803        3.18309        3.7804         1.63396        0.711064
--------------------------------------------------------------------------------
CMS_eff_b            lnN     1.02           1.02           1.02           -              1.02
CMS_eff_t            lnN     1.12           1.12           1.12           -              1.12
CMS_eff_t_highpt     lnN     1.1            1.1            1.1            -              1.1
acceptance_Ztautau   lnN     -              -              1.08           -              -
acceptance_bbH       lnN     -              -              -              -              1.05
acceptance_ttbar     lnN     1.005          -              -              -              -
norm_jetFakes        lnN     -              -              -              1.2            -
xsec_diboson         lnN     -              1.05           -              -              -
```

observation

backgrounds

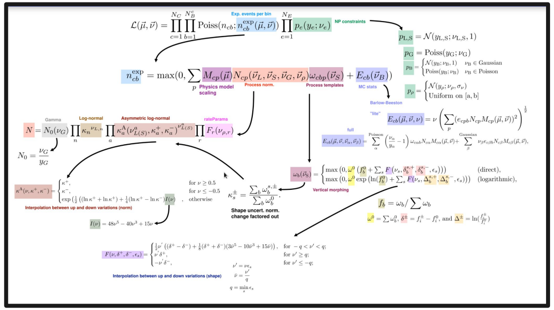
signal

Systematic uncertainties

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(\boldsymbol{x}|\mathrm{s})$, $f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$



Use a modified version of the likelihood ratio test statistic

Test statistic based on likelihood ratio

How can we choose a test's critical region in an 'optimal way', in particular if the data space is multidimensional?

Neyman-Pearson lemma states:

For a test of H_0 of size α, to get the highest power with respect to the alternative H_1 we need for all $\boldsymbol{x}$ in the critical region W

"likelihood ratio (LR)" $\rightarrow$ $$\frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_0)} \geq c_\alpha$$

inside W and $\leq c_\alpha$ outside, where c_α is a co… test of the desired size.

Equivalently, optimal scalar test statistic i…

N.B. any monotonic function of this is lead…

G. Cowan / RHUL Physics Terascale Statistics 2024 / Le…

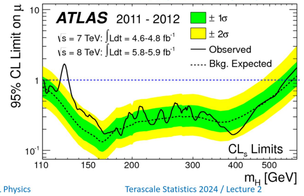
How to read the green and yellow limit plots

For every value of $m_{\rm H}$, find the upper limit on μ.

Also for each $m_{\rm H}$, determine the distribution of upper limits $\mu_{\rm up}$ one would obtain under the hypothesis of μ = 0.

The dashed curve is the median $\mu_{\rm up}$, and the green (yellow) bands give the $\pm$ 1σ (2σ) regions of this distribution.

ATLAS, Phys. Lett. B 716 (2012) 1-29

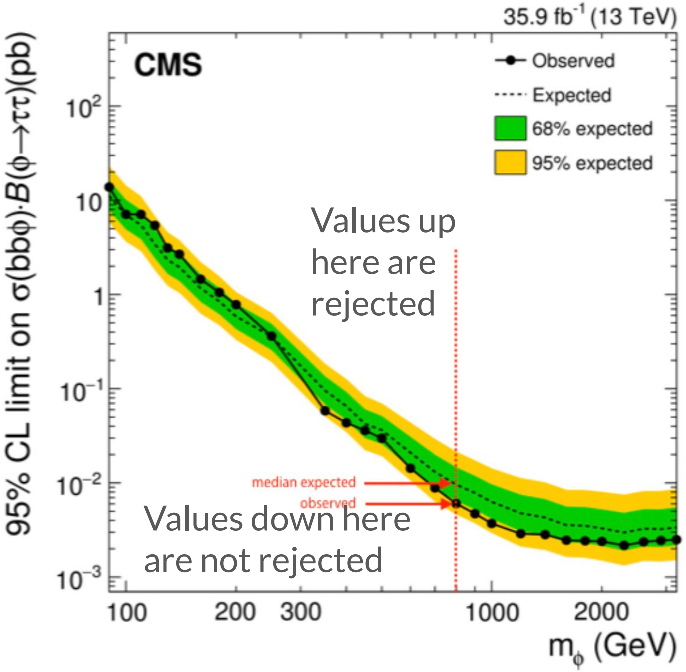


G. Cowan / RHUL Physics Terascale Statistics 2024 / Lecture 2 44

We use:

$$\tilde{q}_\mu = \begin{cases} -2\log\left(\frac{\mathcal{L}(\mu)}{\mathcal{L}(\mu=0)}\right) & \hat{\mu} < 0 \\ -2\log\left(\frac{\mathcal{L}(\mu)}{\mathcal{L}(\hat{\mu})}\right) & 0 < \hat{\mu} < \mu \\ 0 & \mu < \hat{\mu} \end{cases}$$

(And use CL_s criterion)



Asymptotic vs Using Toys

Wilks' Theorem

Wilks' Theorem: if the hypothesized $\mu_i(\boldsymbol{\theta})$, $i = 1,\ldots,N$, are true for some choice of the parameters $\boldsymbol{\theta} = (\theta_1,\ldots, \theta_M)$, then in the large sample limit (and provided regularity conditions are satisfied)

$$t_{\boldsymbol{\mu}} = -2\ln \frac{L(\boldsymbol{\mu}(\hat{\boldsymbol{\theta}}))}{L(\hat{\boldsymbol{\mu}})}$$

MLE of $(\theta_1,\ldots, \theta_M)$ (numerator); MLE of $(\mu_1,\ldots, \mu_N)$ (denominator)

follows a chi-square distribution for $N - M$ degrees of freedom.

The regularity conditions include: the model in the numerator of the likelihood ratio is "nested" within the one in the denominator, i.e., $\boldsymbol{\mu}(\boldsymbol{\theta})$ is a special case of $\boldsymbol{\mu} = (\mu_1,\ldots, \mu_N)$.

Proof boils down to having all estimators ~ Gaussian.

S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

G. Cowan / RHUL Physics   Terascale Statistics 2024 / Lecture 1   51

Under some conditions the distribution of the test statistic is known analytically:
→ Use asymptotic approximation

Otherwise:
Generate many sets of pseudodata to get an empirical distribution of the test statistic

Asymptotic Approximation

```
 <<< Combine >>>
 <<< v9.1.0 >>>
>>> Random number generator seed is 123456
>>> Method used is AsymptoticLimits

 -- AsymptoticLimits ( CLs ) --
Observed Limit: r < 10.8183
Expected  2.5%: r < 7.0537
Expected 16.0%: r < 9.8108
Expected 50.0%: r < 14.5625
Expected 84.0%: r < 22.3988
Expected 97.5%: r < 33.5971
```

Empirical Distribution

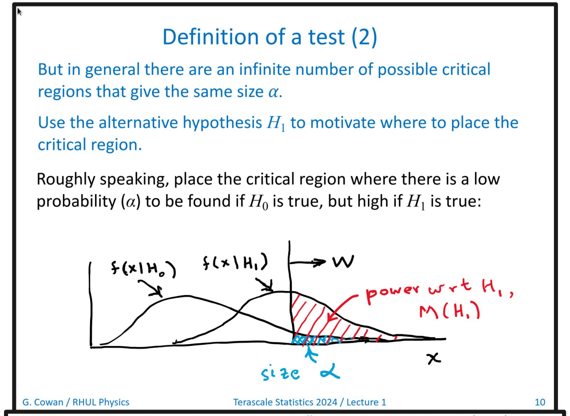
```
 -- Hybrid New --
Limit: r < 11.1291 +/- 0.163054 @ 95% CL
```

Expected

2.5%	`Limit: r < 5.46875 +/- 0.15625 @ 95% CL`
16.0%	`Limit: r < 10.4676 +/- 0.123997 @ 95% CL`
50.0%	`Limit: r < 14.5396 +/- 0.136762 @ 95% CL`
84.0%	`Limit: r < 21.7222 +/- 0.271188 @ 95% CL`
97.5%	`Limit: r < 33.2392 +/- 1.62741 @ 95% CL`

Empirical test statistic distributions

Can directly look at the distributions of the test statistics under the background-only and signal_background hypothesis



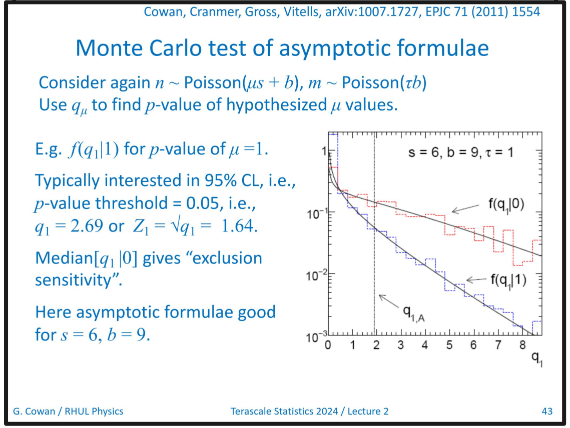
Definition of a test (2)

But in general there are an infinite number of possible critical regions that give the same size α.

Use the alternative hypothesis H_1 to motivate where to place the critical region.

Roughly speaking, place the critical region where there is a low probability (α) to be found if H_0 is true, but high if H_1 is true:

G. Cowan / RHUL Physics Terascale Statistics 2024 / Lecture 1 10



Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Monte Carlo test of asymptotic formulae

Consider again $n \sim$ Poisson($\mu s + b$), $m \sim$ Poisson(τb)
Use q_μ to find p-value of hypothesized μ values.

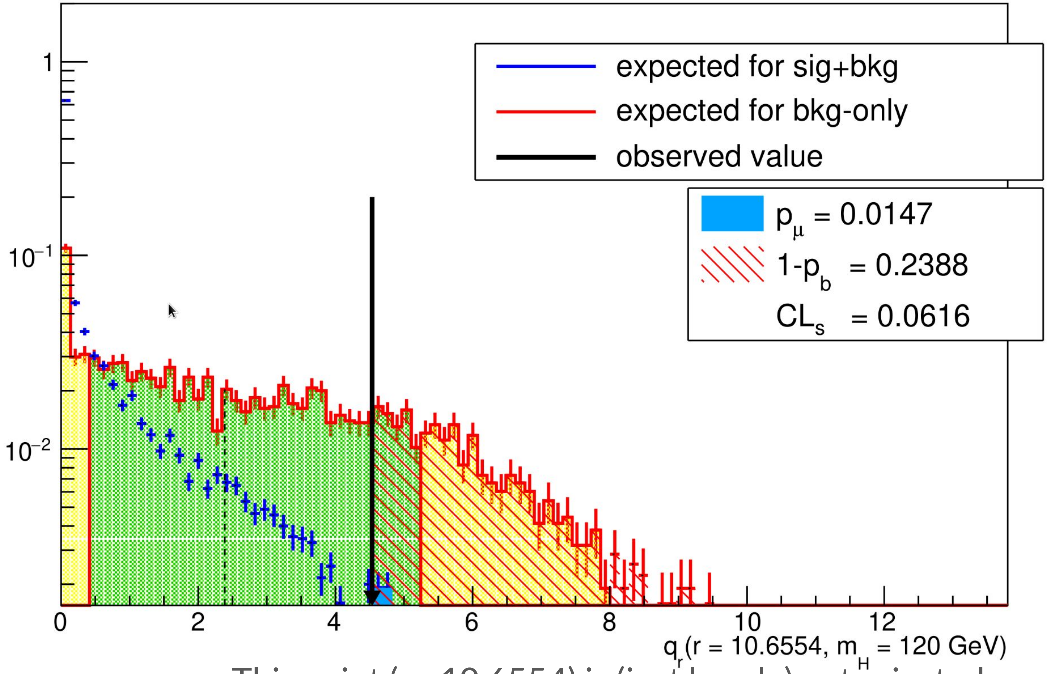
E.g. $f(q_1|1)$ for p-value of $\mu = 1$.

Typically interested in 95% CL, i.e., p-value threshold = 0.05, i.e., $q_1 = 2.69$ or $Z_1 = \sqrt{q_1} = 1.64$.

Median[q_1 |0] gives "exclusion sensitivity".

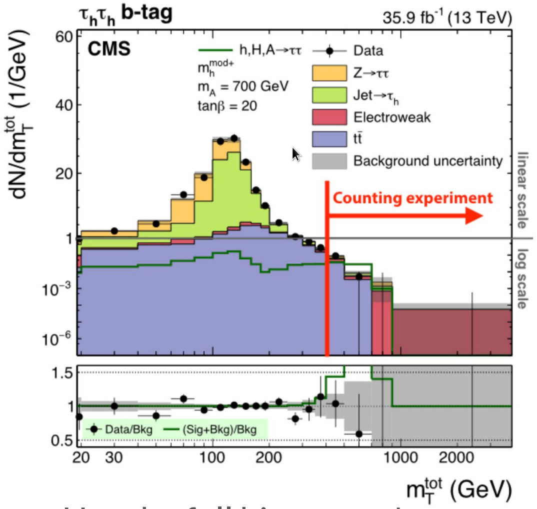
Here asymptotic formulae good for $s = 6$, $b = 9$.

G. Cowan / RHUL Physics Terascale Statistics 2024 / Lecture 2 43



This point (r=10.6554) is (just barely) not rejected

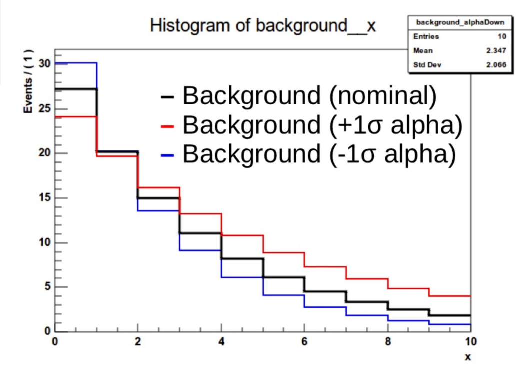
Part 2 - Shape Experiment



Use the full histogram!

Input histograms

```
imax 1
jmax 1
kmax *
---------------
shapes * * simple-shapes-TH1_input.root $PROCESS $PROCESS_$SYSTEMATIC
shapes signal * simple-shapes-TH1_input.root $PROCESS$MASS $PROCESS$MASS_$SYSTEMATIC
---------------
bin bin1
observation 85
------------------------------
bin             bin1        bin1
process         signal      background
process         0           1
rate            10          100
--------------------------------
lumi     lnN    1.10        1.0
bgnorm   lnN    1.00        1.3
alpha  shape    -           1
```



```
imax    1 number of bins
jmax    4 number of processes minus 1
kmax    * number of nuisance parameters
--------------------------------------------------------------------------------
shapes bbHtautau * datacard_part2.shapes.root signal_region/$PROCESS$MASS signal_region/$PROCESS$MASS_$SYSTEMATIC
shapes * * datacard_part2.shapes.root signal_region/$PROCESS signal_region/$PROCESS_$SYSTEMATIC
--------------------------------------------------------------------------------
bin          signal_region
observation  3416.0
--------------------------------------------------------------------------------
bin                                  signal_region  signal_region  signal_region  signal_region  signal_region
process                              ttbar          diboson        Ztautau        jetFakes       bbHtautau
process                              1              2              3              4              0
rate                                 683.017        96.5185        742.649        2048.94        0.913183
--------------------------------------------------------------------------------
CMS_eff_b                    lnN     1.02           1.02           1.02           -              1.02
CMS_eff_t                    lnN     1.12           1.12           1.12           -              1.12
acceptance_Ztautau           lnN     -              -              1.08           -              -
acceptance_bbH               lnN     -              -              -              -              1.05
acceptance_ttbar             lnN     1.005          -              -              -              -
lumi_13TeV                   lnN     1.025          1.025          1.025          -              1.025
norm_jetFakes                lnN     -              -              -              1.2            -
xsec_Ztautau                 lnN     -              -              1.04           -              -
xsec_diboson                 lnN     -              1.05           -              -              -
xsec_ttbar                   lnN     1.06           -              -              -              -
# These ones are new
top_pt_ttbar_shape           shape   1              -              -              -              -
CMS_scale_t_1prong0pi0_13TeV shape   1              1              1              -              1
CMS_scale_t_1prong1pi0_13TeV shape   1              1              1              -              1
CMS_scale_t_3prong0pi0_13TeV shape   1              1              1              -              1
CMS_eff_t_highpt             shape   1              1              1              -              1
```

Shape-based analysis improved limits over simple counting experiment

```
 -- AsymptoticLimits ( CLs ) --
Observed Limit: r < 7.9771
Expected  2.5%: r < 4.7720
Expected 16.0%: r < 6.8417
Expected 50.0%: r < 10.5312
Expected 84.0%: r < 16.9959
Expected 97.5%: r < 26.5059
```

Shape-based

```
 <<< Combine >>>
 <<< v9.1.0 >>>
>>> Random number generator seed is 123456
>>> Method used is AsymptoticLimits

 -- AsymptoticLimits ( CLs ) --
Observed Limit: r < 10.8183
Expected  2.5%: r < 7.0537
Expected 16.0%: r < 9.8108
Expected 50.0%: r < 14.5625
Expected 84.0%: r < 22.3988
Expected 97.5%: r < 33.5971
```

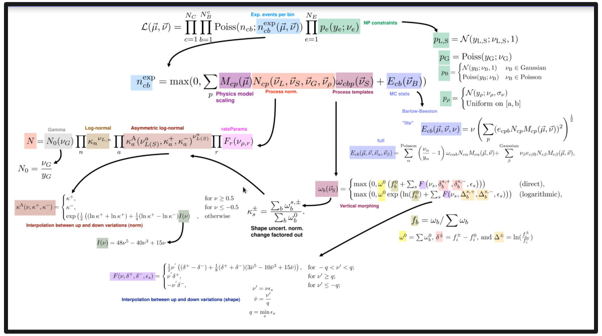
Counting

Fit diagnostics

Neyman-Pearson doesn't usually help

We usually don't have explicit formulae for the pdfs $f(\boldsymbol{x}|\mathrm{s})$, $f(\boldsymbol{x}|\mathrm{b})$, so for a given $\boldsymbol{x}$ we can't evaluate the likelihood ratio

$$t(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}$$



In this case we do have explicit formulas for the pdf, we constructed it with combine!

BUT we want to make sure the model we've constructed is sensible and the fits are running well! → **Inspect various aspects of your model!**

Finally

Two lectures only enough for a brief introduction to:

- Parameter estimation
- Hypothesis tests (→ path to Machine Learning)
- Limits (confidence intervals/regions)
- Systematics (nuisance parameters)

No time for many other interesting topics:

- Experimental sensitivity
- Bayesian parameter estimation

Final thought: once the basic formalism is fixed, most of the work focuses on writing down the likelihood, e.g., $P(\boldsymbol{x}|\boldsymbol{\theta})$, and including in it enough parameters to adequately describe the data (true for both Bayesian and frequentist approaches) so often best to invest most of your time with it.

G. Cowan / RHUL Physics Terascale Statistics 2024 / Lecture 2 45

Fit Parameter Values

```
RooFitResult: minimized FCN value: -2.55338e-05, estimated distance to minimum: 7.54243e-06
                covariance matrix quality: Full, accurate covariance matrix
                Status : MINIMIZE=0 HESSE=0

    Floating Parameter    FinalValue +/-  Error
  --------------------  --------------------------
             CMS_eff_b   -4.5380e-02 +/-  9.93e-01
             CMS_eff_t   -2.6311e-01 +/-  7.33e-01
      CMS_eff_t_highpt   -4.7146e-01 +/-  9.62e-01
  CMS_scale_t_1prong0pi0_13TeV   -1.5989e-01 +/-  5.93e-01
  CMS_scale_t_1prong1pi0_13TeV   -1.6426e-01 +/-  4.94e-01
  CMS_scale_t_3prong0pi0_13TeV   -3.0698e-01 +/-  6.06e-01
    acceptance_Ztautau   -3.1262e-01 +/-  8.62e-01
        acceptance_bbH   -2.8676e-05 +/-  1.00e+00
      acceptance_ttbar    4.9981e-03 +/-  1.00e+00
            lumi_13TeV   -5.6366e-02 +/-  9.89e-01
         norm_jetFakes   -9.3327e-02 +/-  2.56e-01
                     r   -2.7220e+00 +/-  2.59e+00
    top_pt_ttbar_shape    1.7586e-01 +/-  7.00e-01
          xsec_Ztautau   -1.6007e-01 +/-  9.66e-01
          xsec_diboson    3.9758e-02 +/-  1.00e+00
            xsec_ttbar    5.7794e-02 +/-  9.46e-01
```

```
name                                        b-only fit       s+b fit       rho
CMS_eff_b                                   -0.04, 0.99      -0.05, 0.99    +0.01
CMS_eff_t                                 * -0.24, 0.73*   * -0.26, 0.73*   +0.06
CMS_eff_t_highpt                          * -0.56, 0.94*   * -0.47, 0.96*   +0.02
CMS_scale_t_1prong0pi0_13TeV              * -0.17, 0.58*   * -0.16, 0.59*   -0.04
CMS_scale_t_1prong1pi0_13TeV              ! -0.12, 0.45!   ! -0.16, 0.49!   +0.20
CMS_scale_t_3prong0pi0_13TeV              * -0.31, 0.61*   * -0.31, 0.61*   +0.02
acceptance_Ztautau                        * -0.31, 0.86*   * -0.31, 0.86*   -0.05
acceptance_bbH                              +0.00, 1.00      -0.00, 1.00    +0.05
acceptance_ttbar                            +0.01, 1.00      +0.00, 1.00    +0.00
lumi_13TeV                                  -0.05, 0.99      -0.06, 0.99    +0.01
norm_jetFakes                             ! -0.09, 0.26!   ! -0.09, 0.26!   -0.05
top_pt_ttbar_shape                        * +0.24, 0.69*   * +0.18, 0.70*   +0.22
xsec_Ztautau                                -0.16, 0.97      -0.16, 0.97    -0.02
xsec_diboson                                +0.03, 1.00      +0.04, 1.00    -0.02
xsec_ttbar                                  +0.08, 0.95      +0.06, 0.95    +0.02
```

Part 3

Rate Parameters

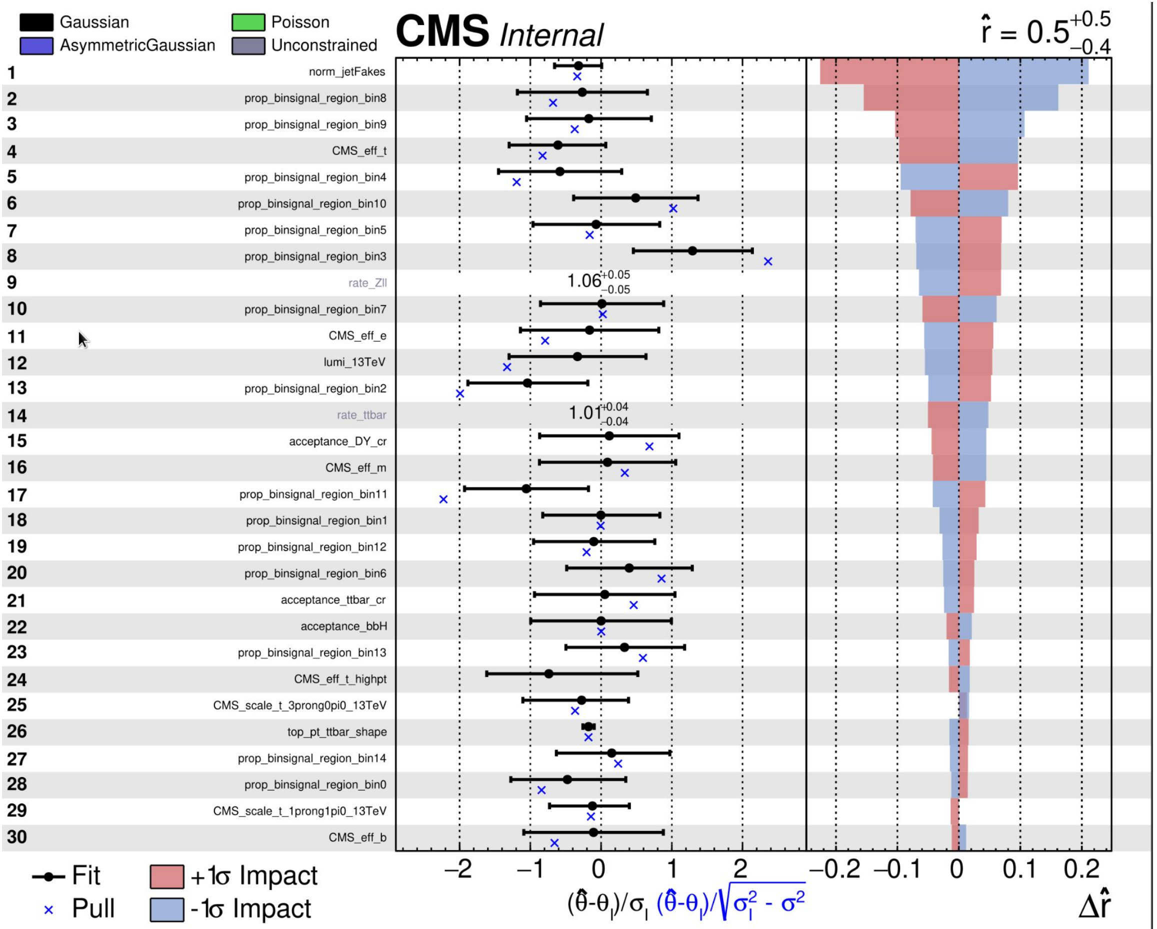
Can add rate parameters which scale certain processes

```
[name] rateParam [channel] [process] [init] [min,max]
```

To allow the rates of the ttbar and Z->ll process in the control regions to influence those in the signal region,
Connect them to each other via a rate parameter

```
rate_ttbar rateParam * ttbar 1
rate_Zll rateParam * Ztautau 1
rate_Zll rateParam * Zmumu 1
```

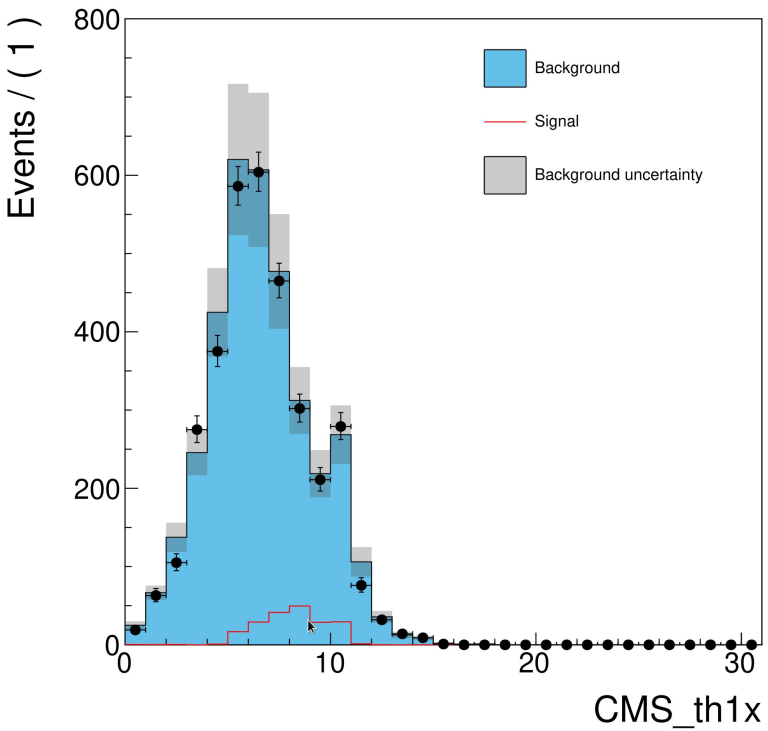
Impacts



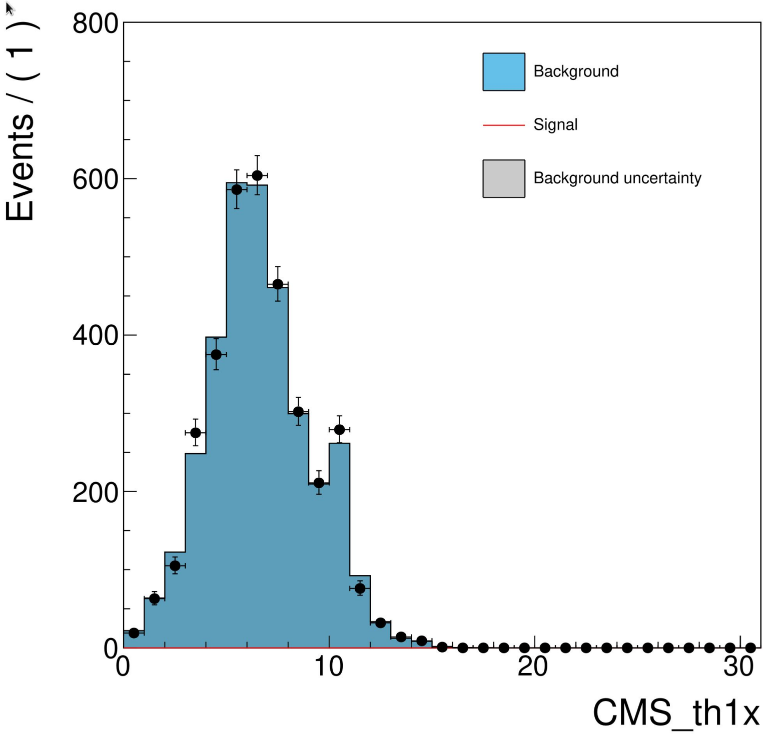
Gaussian
Poisson
AsymmetricGaussian
Unconstrained
CMS Internal
r̂ = 0.5 +0.5 −0.4
1 norm_jetFakes
2 prop_binsignal_region_bin8
3 prop_binsignal_region_bin9
4 CMS_eff_t
5 prop_binsignal_region_bin4
6 prop_binsignal_region_bin10
7 prop_binsignal_region_bin5
8 prop_binsignal_region_bin3
9 rate_Zll
1.06 +0.05 −0.05
10 prop_binsignal_region_bin7
11 CMS_eff_e
12 lumi_13TeV
13 prop_binsignal_region_bin2
14 rate_ttbar
1.01 +0.04 −0.04
15 acceptance_DY_cr
16 CMS_eff_m
17 prop_binsignal_region_bin11
18 prop_binsignal_region_bin1
19 prop_binsignal_region_bin12
20 prop_binsignal_region_bin6
21 acceptance_ttbar_cr
22 acceptance_bbH
23 prop_binsignal_region_bin13
24 CMS_eff_t_highpt
25 CMS_scale_t_3prong0pi0_13TeV
26 top_pt_ttbar_shape
27 prop_binsignal_region_bin14
28 prop_binsignal_region_bin0
29 CMS_scale_t_1prong1pi0_13TeV
30 CMS_eff_b
Fit
Pull
+1σ Impact
-1σ Impact
−2 −1 0 1 2
$(\hat{\theta}-\theta_I)/\sigma_I$ $(\hat{\theta}-\theta_I)/\sqrt{\sigma_I^2 - \sigma^2}$
−0.2 −0.1 0 0.1 0.2
$\Delta\hat{r}$

Visualizing fits

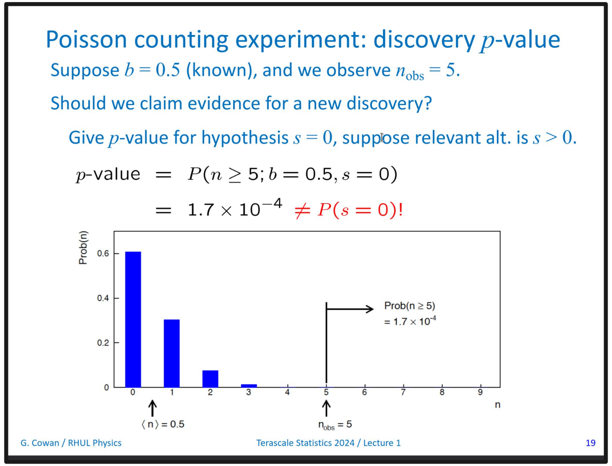
Prefit

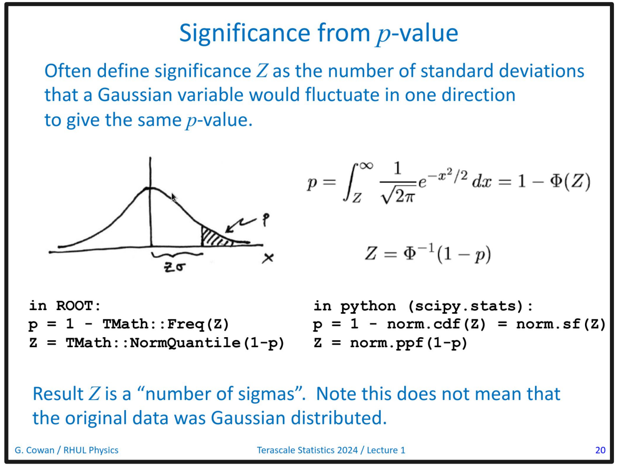


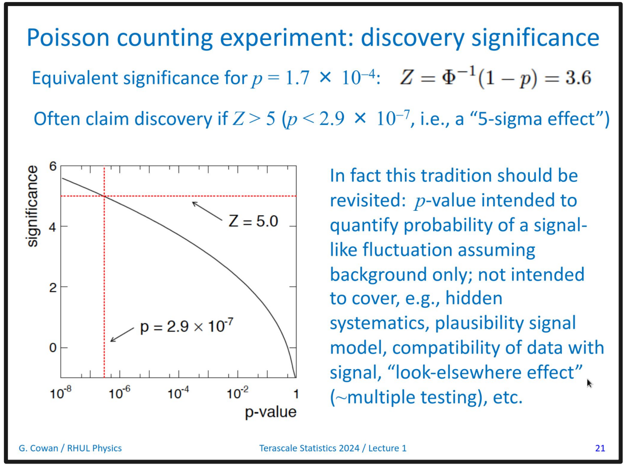
background -only fit



Significance







We calculate the p-value of the modified Likelihood Ratio test-statistic q_0 and quote a significance

$$q_0 = \begin{cases} 0 & \hat{\mu} < 0 \\ -2\log\left(\frac{\mathcal{L}(\mu_{\mathrm{NP}}=0)}{\mathcal{L}(\hat{\mu}_{\mathrm{NP}})}\right) & \hat{\mu} \geq 0 \end{cases}$$

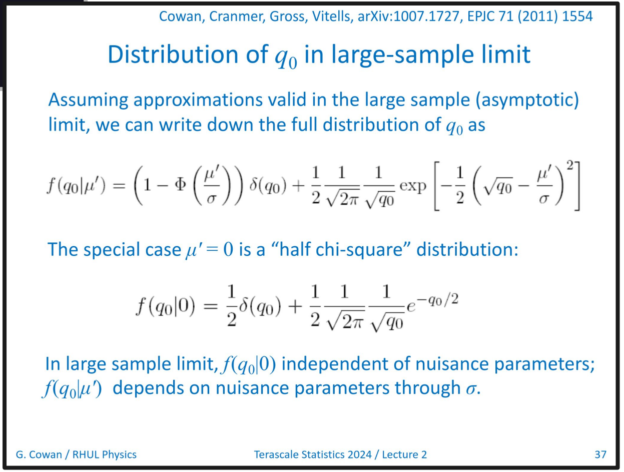
Significance

Simple Asymptotic Calculation – assume known distribution of test-statistic

```
 <<< Combine >>>
 <<< v9.1.0 >>>
>>> Random number generator seed is 123456
>>> Method used is Significance

 -- Significance --
Significance: 1.11273
```

Asymptotic method using the large sample limit (+ some other conditions, like Wilk’s theorem)


Cowan, Cranmer, Gross, Vitells, arXiv:1007.1727, EPJC 71 (2011) 1554

Distribution of q_0 in large-sample limit

Assuming approximations valid in the large sample (asymptotic) limit, we can write down the full distribution of q_0 as

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]$$

The special case $\mu' = 0$ is a “half chi-square” distribution:

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}$$

In large sample limit, $f(q_0|0)$ independent of nuisance parameters; $f(q_0|\mu')$ depends on nuisance parameters through σ.

G. Cowan / RHUL Physics Terascale Statistics 2024 / Lecture 2 37

Can also instruct combine to give the p-value directly with the `--pvalue` flag

```
 -- Significance --
p-value of background: 0.132912
Done in 0.00 min (cpu), 0.00 min (real)
```

Significance

Can also check expected significance for various signal strengths, e.g. with `-t -1 --expectSignal 1.5`

```
 -- Significance --
Significance: 3.52007
```

Can also use fit model after a first fit to the data to get model parameters with `--toysFrequentist`

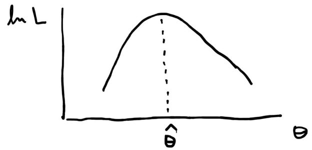
```
 -- Significance --
Significance: 3.13954
```

Signal Strength measurement

Maximum Likelihood Estimators (MLEs)

We *define* the maximum likelihood estimators or MLEs to be the parameter values for which the likelihood is maximum.

Maximizing L equivalent to maximizing $\log L$



$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

Could have multiple maxima (take highest).

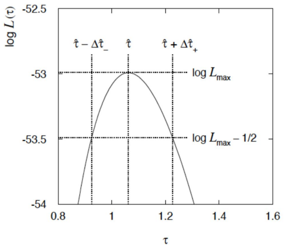
MLEs not guaranteed to have any 'optimal' properties, (but in practice they're very good).

G. Cowan / RHUL Physics — Terascale Statistics 2024 / Lecture 1 — 25

Example of variance by graphical method

ML example with exponential:

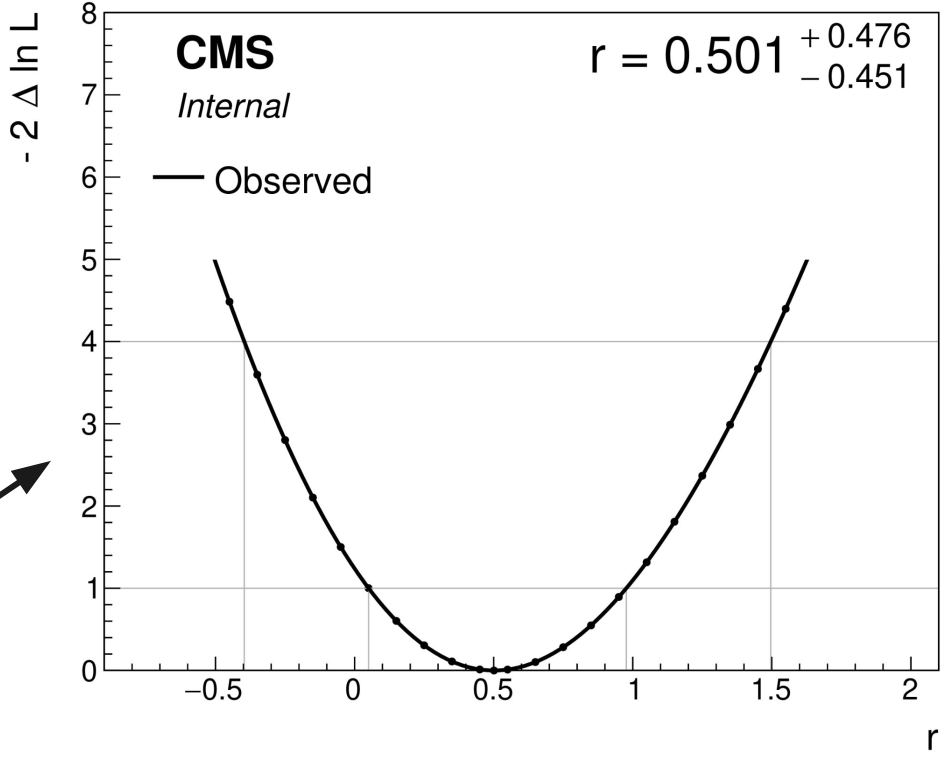
$$\hat{\tau} = 1.062$$
$$\Delta\hat{\tau}_{-} = 0.137$$
$$\Delta\hat{\tau}_{+} = 0.165$$
$$\hat{\sigma}_{\hat{\tau}} \approx \Delta\hat{\tau}_{-} \approx \Delta\hat{\tau}_{+} \approx 0.15$$



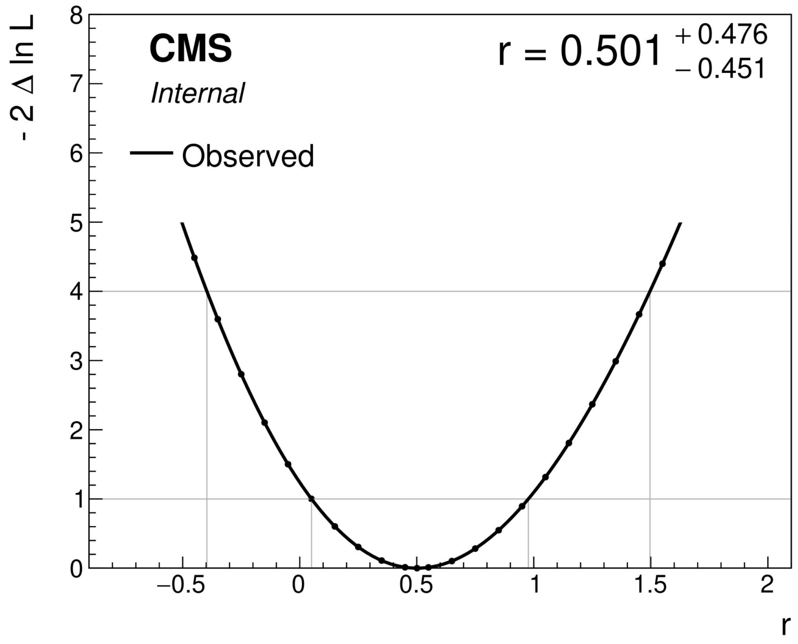
Not quite parabolic $\ln L$ since finite sample size (n = 50).

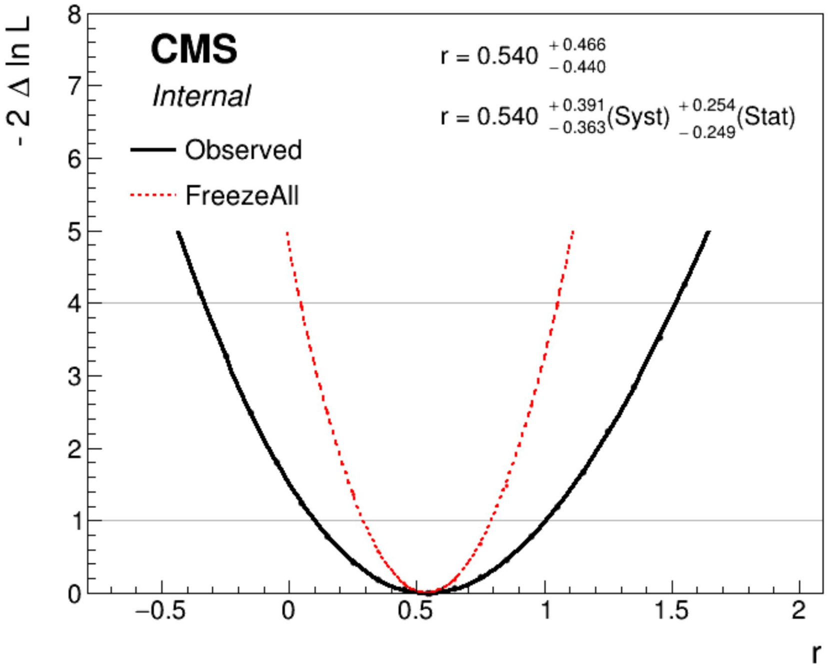
G. Cowan / RHUL Physics — Terascale Statistics 2024 / Lecture 1 — 33

With a sign flip



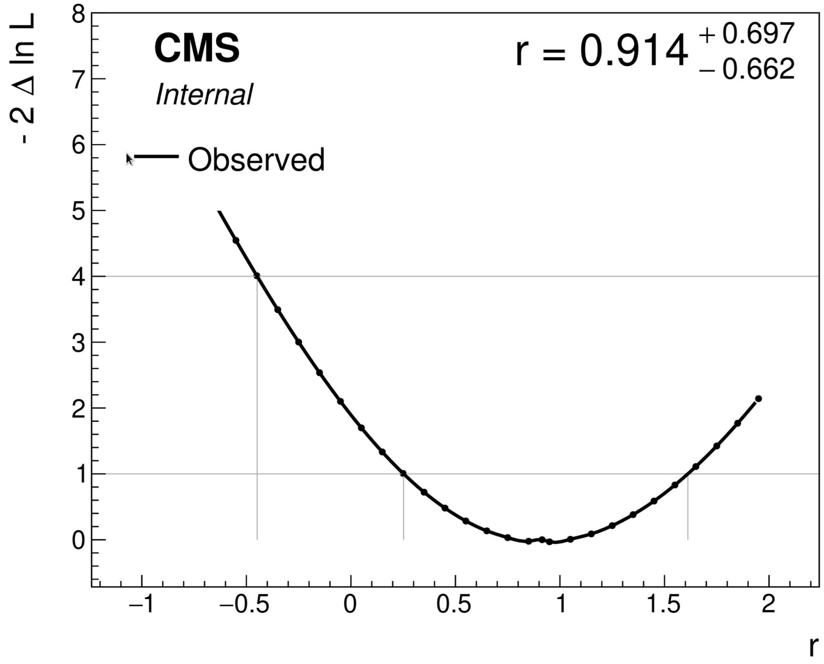
Uncertainty Breakdowns

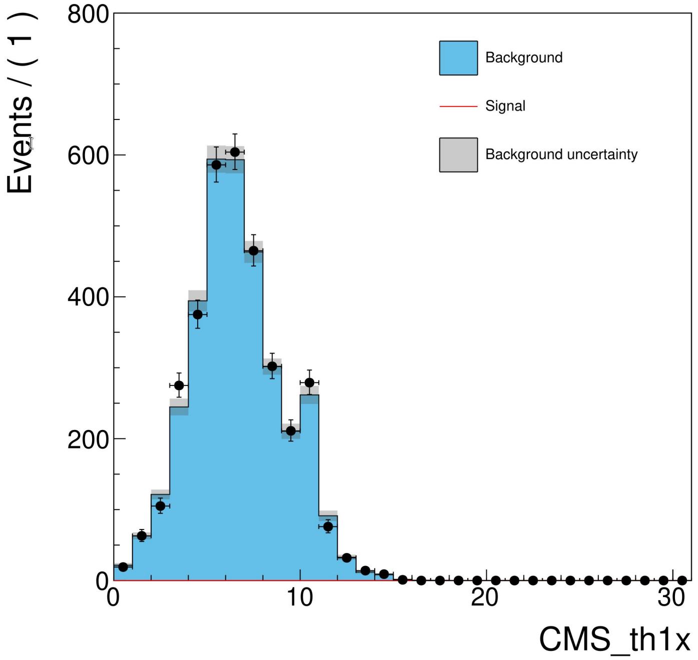




Channel Masks

Can also mask particular channels to investigate the fit, e.g. masking the control regions:





Part 4

Two Parameters of Interest

```
from HiggsAnalysis.CombinedLimit.PhysicsModel import PhysicsModel


class DASModel(PhysicsModel):
    def doParametersOfInterest(self):
        """Create POI and other parameters, and define the POI set."""
        self.modelBuilder.doVar("r[0,0,10]")
        self.modelBuilder.doSet("POI", ",".join(["r"]))

    def getYieldScale(self, bin, process):
        "Return the name of a RooAbsReal to scale this yield by or the two special values 1 and 0
        if self.DC.isSignal[process]:
            print("Scaling %s/%s by r" % (bin, process))
            return "r"
        return 1


dasModel = DASModel()
```

2-Dimensional NLL map