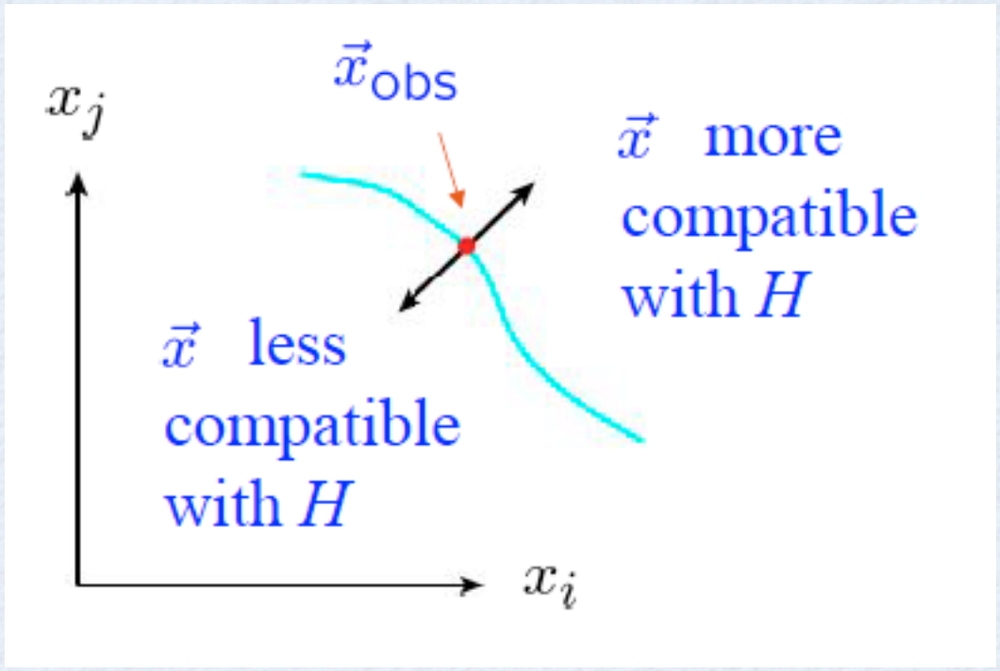
Significance, Interval Estimation, and Limits

TESTING SIGNIFICANCE

Suppose hypothesis H predicts PDF $f(\vec{x}|H)$
for a set of observations $\vec{x} = (x_1, x_2, ...x_n)$

In our experiment we make $\vec{x}_{obs}$

What can we say about the validity of H given this data?



P-VALUES

- We express the 'goodness-of-fit' of the data with a hypothesis by specifying it's p-value

- P-value = probability to observe data that is compatible with hypothesis H with lesser or equal compatibility with data we observed

P-VALUES

- We express the 'goodness-of-fit' of the data with a hypothesis by specifying it's p-value

- P-value = probability to observe data that is compatible with hypothesis H with lesser or equal compatibility with data we observed

- NOTE THIS IS NOT THE PROBABILITY THAT H IS TRUE! (i.e. P(H) or even P(H | x))

  - Common misconception but remember:

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

EXAMPLE: FLIPPING A COIN

- If you had a `fair' coin what how many heads would you expect if you had 20 coin tosses?

EXAMPLE: FLIPPING A COIN

- If you had a `fair' coin what how many heads would you expect if you had 20 coin tosses?
  - Expectation value for a fair coin is 10 heads in 20 tosses

Example: Flipping a coin

- If you had a `fair' coin what how many heads would you expect if you had 20 coin tosses?
  - Expectation value for a fair coin is 10 heads in 20 tosses
  - If you saw 11 or 12 heads you probably wouldn't think much of it. But what if you got 17 heads out of 20?

COIN TOSSING

- Probability to observe n heads in N coin tosses is given by a binomial distribution where p = probability coin will land on heads

- Hypothesis is H is that the coin is fair (p=0.5)

- Suppose we toss the coin N = 20 times and get n = 17 heads. Region of data space with equal or lesser compatibility with

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

COIN TOSSING

$$P(n; p, N) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

Suppose we toss the coin N = 20 times and get n = 17 heads. Region of data space with equal or lesser compatibility with H relative to n = 17 is: n = 17, 18, 19, 20, 0, 1, 2, 3. Adding up the probabilities for these values gives:

$$P(n = 0, 1, 2, 3, 17, 18, 19, \text{ or } 20) = 0.0026 \, .$$

i.e. p = 0.0026 is the probability of obtaining such a bizarre result (or more so) 'by chance', under the assumption of H.

OBSERVATION OF A NEW PARTICLE

- Typically when we first observe a new particle we only observe a few events
- There are exceptions - but for the most part if we were seeing a huge number of them right away chances are someone else saw them first

SIGNIFICANCE

Suppose we observe n events

n_b events from known processes (background)
n_s events from a new process (signal)

For a small number of events the binomial reduces to a Poisson distribution

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

SIGNIFICANCE

Suppose we observe n=5 events with an expectation of 0.5 events from background

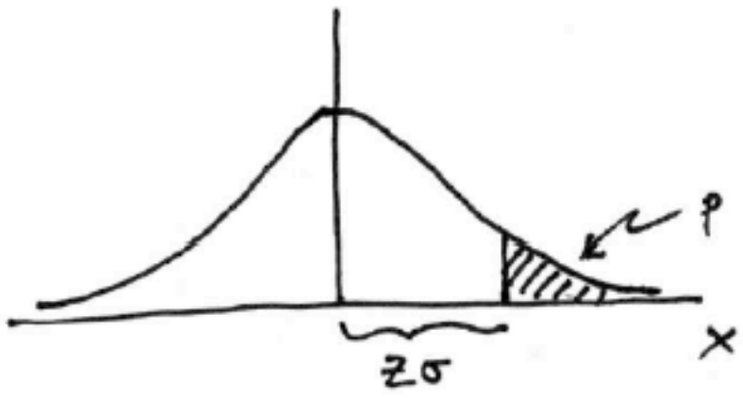
$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Give p-value for hypothesis $s = 0$:

$$p\text{-value} = P(n \geq 5; b = 0.5, s = 0)$$
$$= 1.7 \times 10^{-4} \neq P(s = 0)!$$

SIGNIFICANCE

- Often define significance Z as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value.



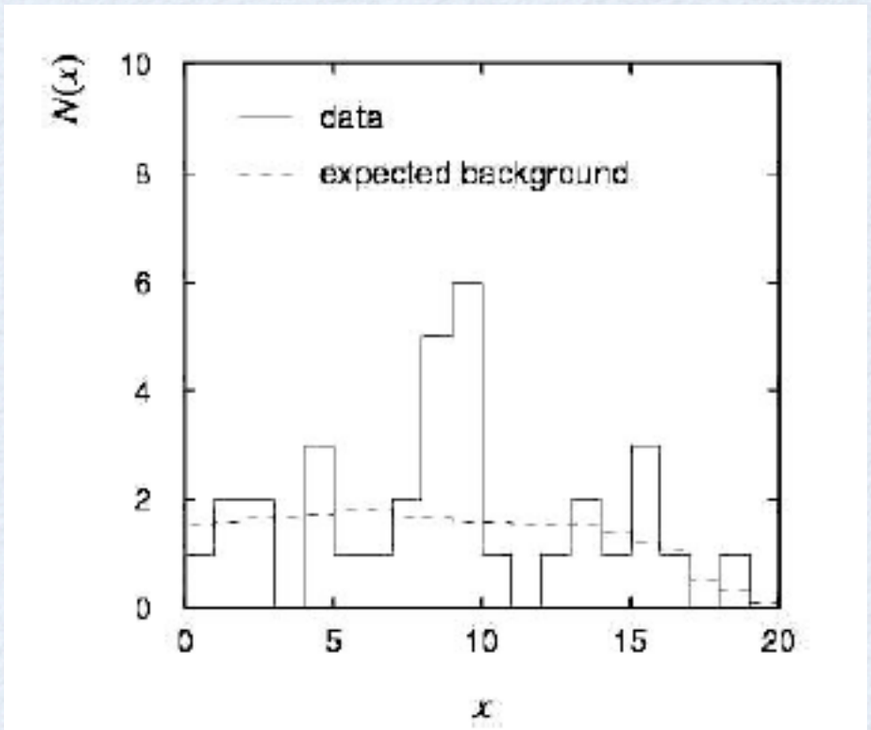
$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1 - \Phi(Z)$$ `1 - TMath::Freq`

$$Z = \Phi^{-1}(1-p)$$ `TMath::NormQuantile`

WHAT IF?

Suppose we measure a value x for each event and find:

Each bin (observed) is a Poisson r.v., means are given by dashed lines.



In the two bins with the peak, 11 entries found with b = 3.2. The p-value for the s = 0 hypothesis is:

$$P(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

WHAT TO DO NOW?

- But... did we know where to look for the peak?
  - → give P(n ≥ 11) in any 2 adjacent bins
- Is the observed width consistent with the expected x resolution?
  - → take x window several times the expected resolution
- How many bins °ø distributions have we looked at?
  - → look at a thousand of them, you'll find a 10-3 effect
- Did we adjust the cuts to 'enhance' the peak?
  - → freeze cuts, repeat analysis with new data
- How about the bins to the sides of the peak... (too low!)
- Should we publish????

WHEN TO PUBLISH

- HEP folklore is to claim discovery when p = 2.9 x10^-7, corresponding to a significance Z = 5.
- This is very subjective and really should depend on the prior probability of the phenomenon in question, e.g.,

phenomenon	reasonable *p*-value for discovery
D^0D^0 mixing	~0.05
Higgs	~ 10^{-7} (?)
Life on Mars	~10^{-10}
Astrology	~10^{-20}

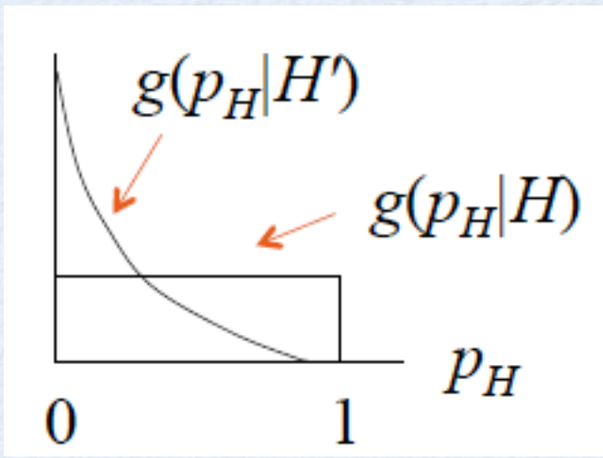
The p-value is a function of the data, and is thus itself a random variable with a given distribution. Suppose the p-value of H is found from a test statistic t(x) as

$$p_H = \int_t^\infty f(t'|H)dt'$$

The pdf of pH under assumption of H is

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H/\partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \le p_H \le 1)$$

In general for continuous data, under assumption of **H**, $\mathbf{p_H}$ ~ Uniform[0,1] and is concentrated toward zero for Some (broad) class of alternatives.



USING P-VALUE

So the probability to find the **p**-value of $\mathbf{H}_0$, $\mathbf{p}_0$, less than α is

$$P(p_0 \leq \alpha | H_0) = \alpha$$

We started by defining critical region in the original data space (**x**), then reformulated this in terms of a scalar test statistic **t**(**x**).
We can take this one step further and define the critical region of a test of $\mathbf{H}_0$ with size α as the set of data space where $\mathbf{p}_0 \leq \alpha$.
Formally the **p**-value relates only to $\mathbf{H}_0$, but the resulting test will have a given power with respect to a given alternative $\mathbf{H}_1$.

RECALL χ^2

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\sigma_i^2} \ , \ \text{where } \sigma_i^2 = V[n_i] \ .$$

- If the n are assumed to be Gaussian one can show that the χ^2 pdf will be (with $z = \chi^2$)

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2}\Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

Then the p-value becomes

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N)\, dz$$

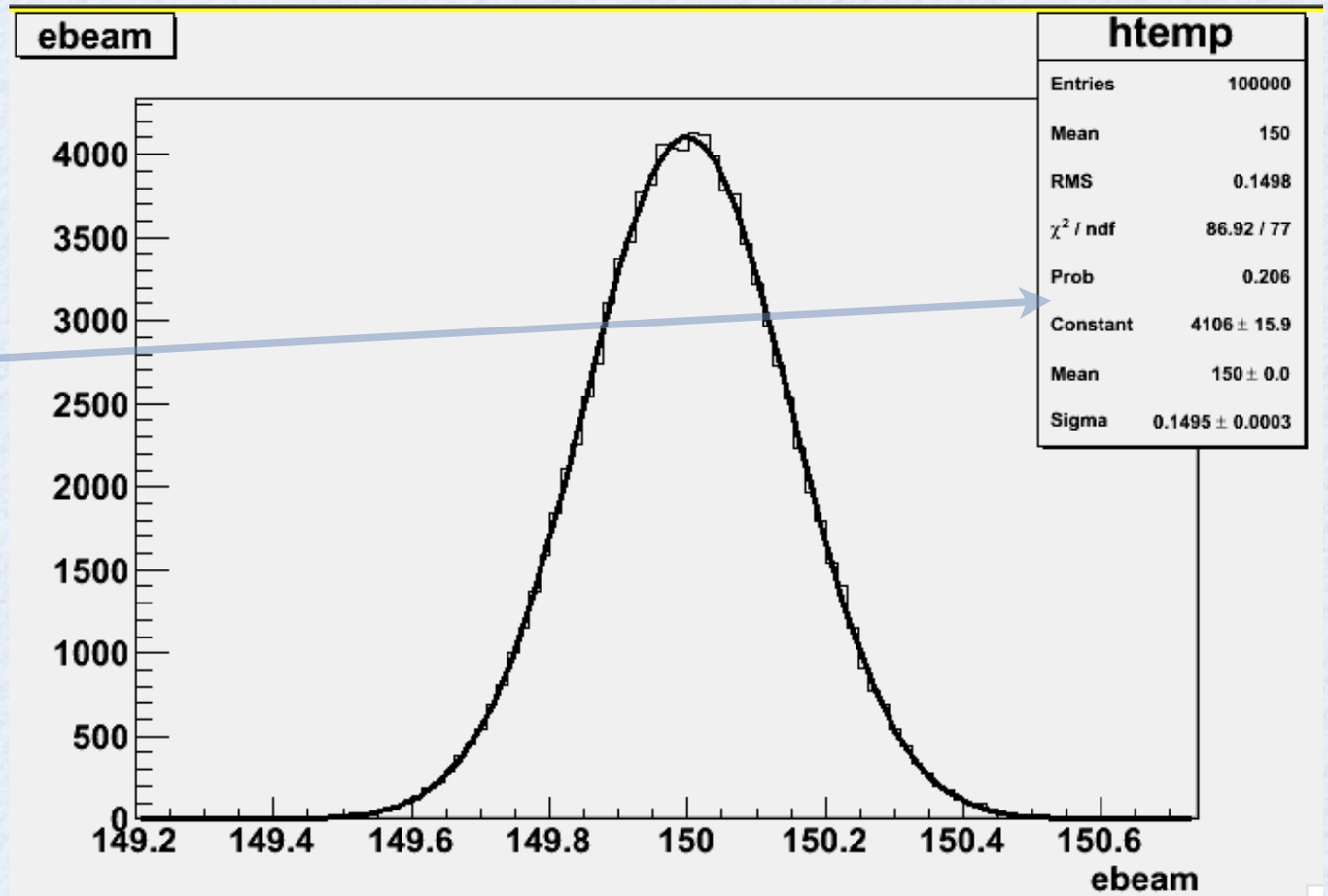
P-VALUE FOR FIT

- One often sees $\chi 2/N$ reported as a measure of goodness-of-fit.
- But... better to give $\chi 2$and N separately. Consider, e.g.,

$$\chi^2 = 15,\ N = 10 \rightarrow p - \text{value} = 0.13\ ,$$
$$\chi^2 = 150,\ N = 100 \rightarrow p - \text{value} = 9.0 \times 10^{-4}$$

ROOT FITS

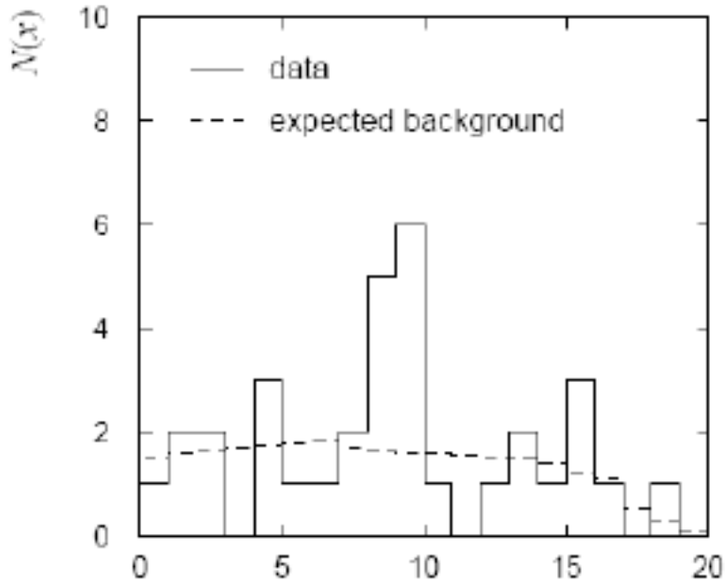
ROOT reports this - this is an important thing to check!



GET IT FROM ROOT

```
3836 Int_t TH1::GetQuantiles(Int_t nprobSum, Double_t *q, const Double_t *probSum)
3837 {
3838    //  Compute Quantiles for this histogram
3839    //     Quantile x_q of a probability distribution Function F is defined as
3840    //
3841    //         F(x_q) = q with 0 <= q <= 1.
3842    //
3843    //     For instance the median x_0.5 of a distribution is defined as that value
3844    //     of the random variable for which the distribution function equals 0.5:
3845    //
3846    //         F(x_0.5) = Probability(x < x_0.5) = 0.5
3847    //
3848    //  code from Eddy Offermann, Renaissance
3849    //
3850    // input parameters
3851    //    - this 1-d histogram (TH1F,D,etc). Could also be a TProfile
3852    //    - nprobSum maximum size of array q and size of array probSum (if given)
3853    //    - probSum array of positions where quantiles will be computed.
3854    //      if probSum is null, probSum will be computed internally and will
3855    //      have a size = number of bins + 1 in h. it will correspond to the
3856    //       quantiles calculated at the lowest edge of the histogram (quantile=0) and
3857    //      all the upper edges of the bins.
3858    //      if probSum is not null, it is assumed to contain at least nprobSum values.
3859    //  output
3860    //    - return value nq (<=nprobSum) with the number of quantiles computed
3861    //    - array q filled with nq quantiles
3862    //
3863    //  Note that the Integral of the histogram is automatically recomputed
3864    //  if the number of entries is different of the number of entries when
3865    //  the integral was computed last time. In case you do not use the Fill
3866    //  functions to fill your histogram, but SetBinContent, you must call
3867    //  TH1::ComputeIntegral before calling this function.
3868    //
3869    //  Getting quantiles q from two histograms and storing results in a TGraph,
3870    //    a so-called QQ-plot
3871    //
3872    //      TGraph *gr = new TGraph(nprob);
3873    //      h1->GetQuantiles(nprob,gr->GetX());
3874    //      h2->GetQuantiles(nprob,gr->GetY());
3875    //      gr->Draw("alp");
3876    //
```

χ^2 test



← This gives

$$\chi^2 = \sum_{i=1}^{N} \frac{(n_i - \nu_i)^2}{\nu_i} = 29.8$$

for $N = 20$ dof.

- can use this as a test of goodness of fit!
- Now need to find p-value, but... many bins have few (or no) entries, so here we do not expect $\chi 2$ to follow the chi-square pdf

WARNING

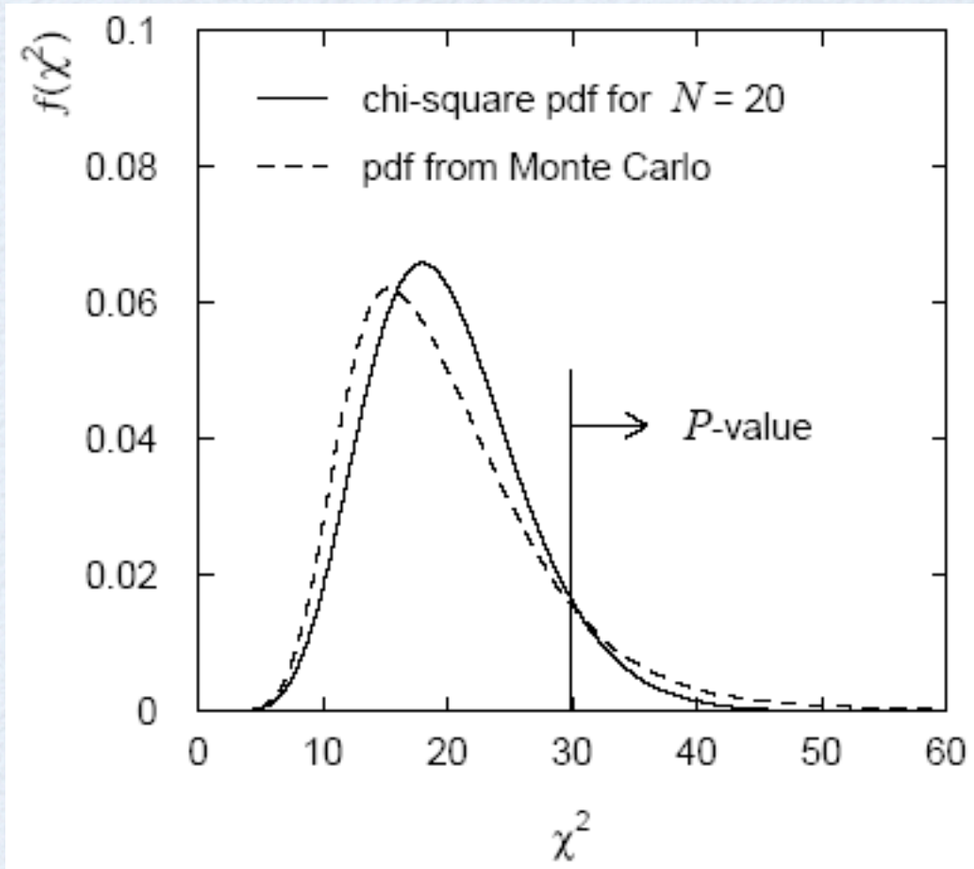
- $\chi 2$ Test assumes that entries in each bin are drawn from a gaussian distribution
- Excellent results for high statistic histograms
- With very small number of events (especially 0!) in bins the results can be difficult to interpret

- The χ2 statistic still reflects the level of agreement between data and prediction, i.e., it is still a 'valid' test statistic.

- To find its sampling distribution, simulate the data with a Monte Carlo program:

    - Here data sample simulated 106 times. The fraction of times wefind $\chi 2 > 29.8$ gives the p-value: $p = 0.11$

- If we had used the chi-square pdf we would find $p = 0.073$.



INTERVAL ESTIMATION

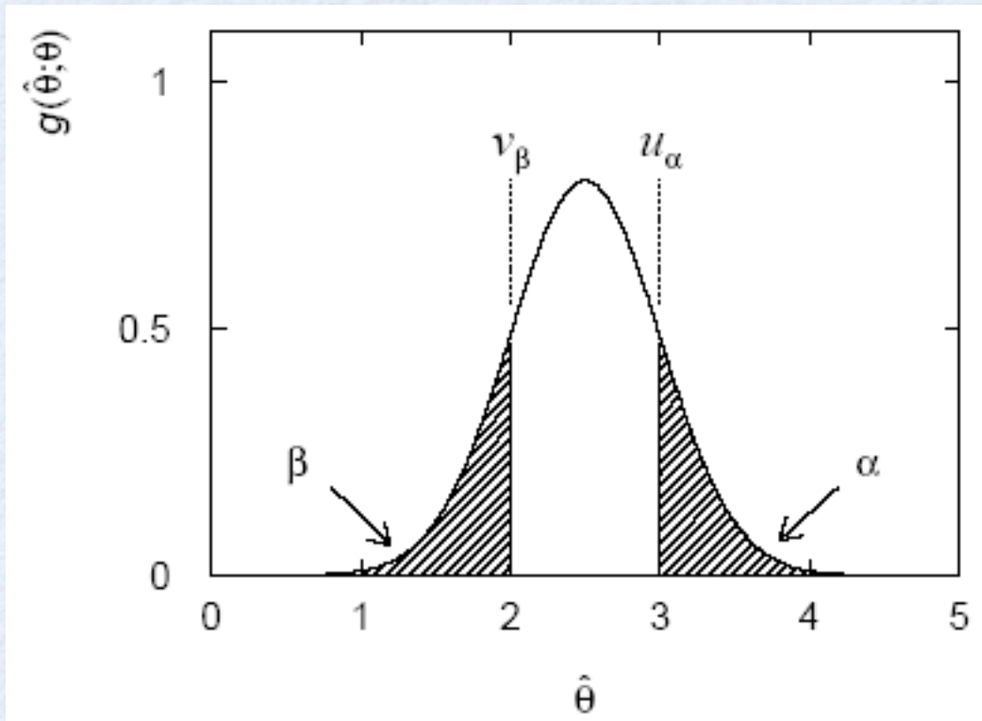
- In addition to a 'point estimate' of a parameter we should report an interval reflecting its statistical uncertainty.
- Desirable properties of such an interval may include:
  - communicate objectively the result of the experiment;
  - have a given probability of containing the true parameter;
  - provide information needed to draw conclusions about
  - the parameter possibly incorporating stated prior beliefs

INTERVAL ESTIMATION

- Consider an estimator for a parameter θ and an estimate. We also need for all possible θ its sampling distribution

- Specify upper and lower tail probabilities, e.g., $\alpha = 0.05$, $\beta = 0.05$, then find functions $u_{\alpha(\theta)}$ and $v_{\beta(\theta)}$ such that:

$$\begin{aligned}
\alpha &= P(\hat{\theta} \geq u_\alpha(\theta)) \\
&= \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta} \\
\beta &= P(\hat{\theta} \leq v_\beta(\theta)) \\
&= \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta}
\end{aligned}$$

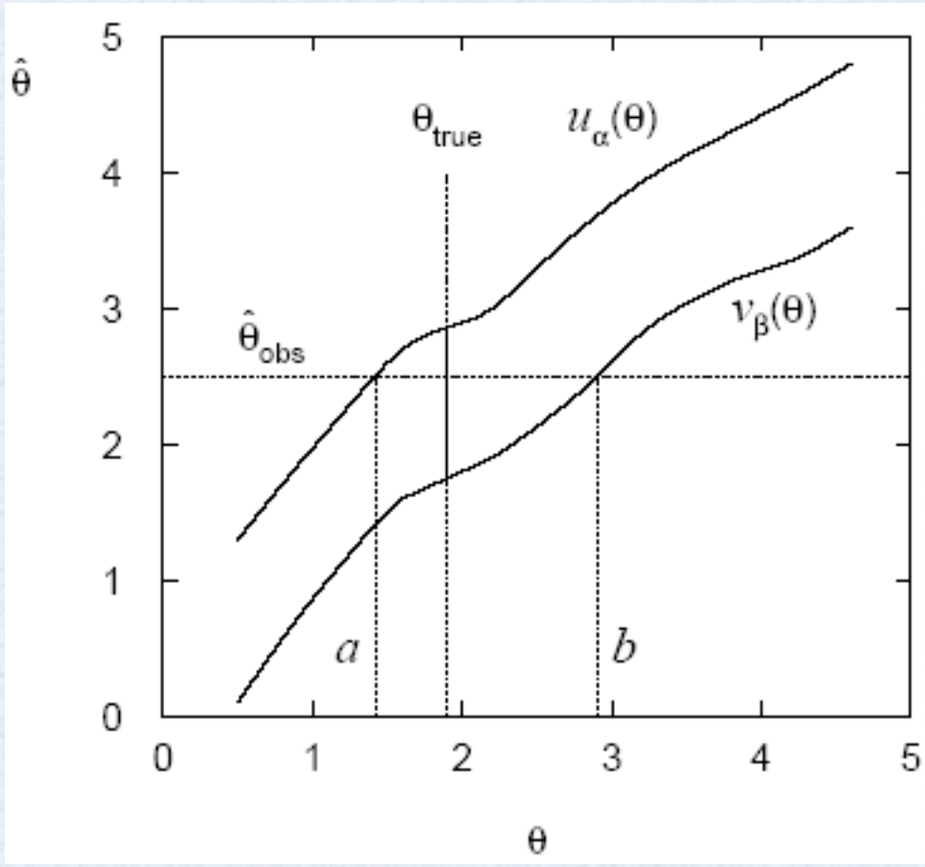


CONFIDENCE LEVEL

- the region between these two functions is called the confidence belt

- [a,b] - confidence level

- Confidence level = 1 - α - β = probability for the interval to cover true value of the parameter (holds for any possible true θ).



- Confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (do this for all θ):

- Specify values of the data that are 'disfavored' by θ (critical region) such that P (data in critical region) $\leq \gamma$ or a pre specified γ, e.g., 0.05 or 0.1.

- If data observed in the critical region, reject the value θ .

- Now invert the test to define a confidence interval as:

  - set of θ values that would not be rejected in a test of size γ (confidence level is $1 - \gamma$).

- The interval will cover the true value of θ with probability $\geq 1 - \gamma$.

- Equivalent to confidence belt construction; confidence belt is acceptance region of a test.

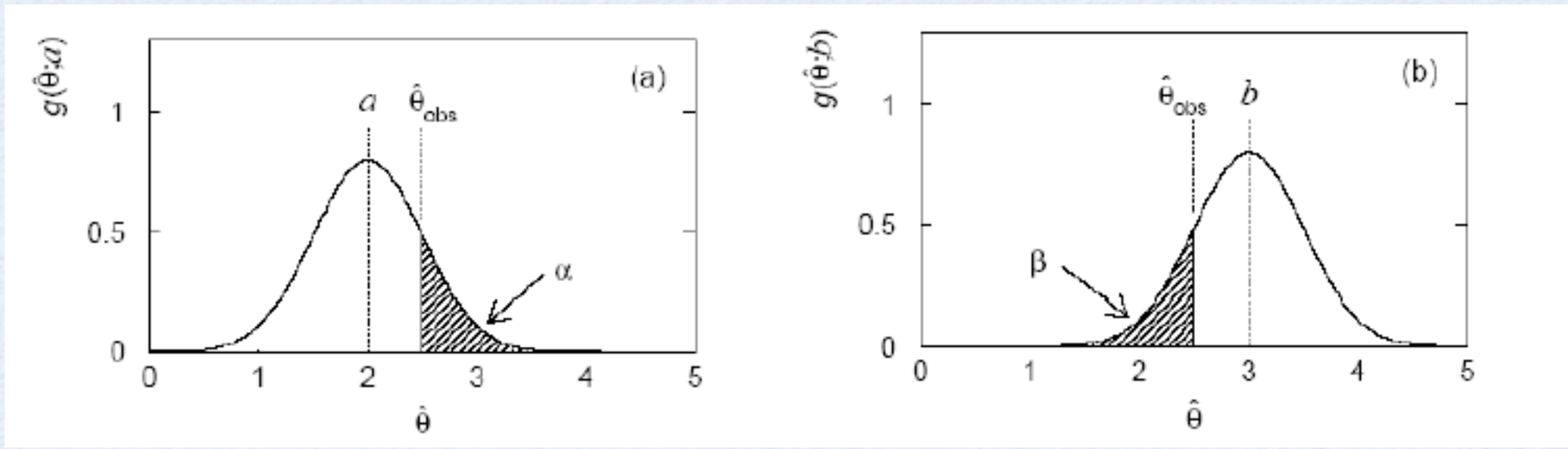
Relation with p-value

- Equivalently we can consider a significance test for each hypothesized value of θ, resulting in a p-value, $p\theta$..
- If $p\theta < \gamma$, then we reject θ.
- The confidence interval at CL = $1 - \gamma$ consists of those values of θ that are not rejected.
- E.g. an upper limit on θ is the greatest value for which $p\theta \geq \gamma$.
- In practice find by setting $p\theta = \gamma$ and solve for θ.

RECIPE

$$\alpha = \int_{u_\alpha(\theta)}^{\infty} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{\hat{\theta}_{\text{obs}}}^{\infty} g(\hat{\theta};a)\, d\hat{\theta}\,,$$

$$\beta = \int_{-\infty}^{v_\beta(\theta)} g(\hat{\theta};\theta)\, d\hat{\theta} = \int_{-\infty}^{\hat{\theta}_{\text{obs}}} g(\hat{\theta};b)\, d\hat{\theta}\,.$$



→ a is hypothetical value of θ such that $P(\hat{\theta} > \hat{\theta}_{\text{obs}}) = \alpha$.

→ b is hypothetical value of θ such that $P(\hat{\theta} < \hat{\theta}_{\text{obs}}) = \beta$.

N.B. the interval is random, the true θ is an unknown constant.

Often report interval $[a, b]$ as $\hat{\theta}^{+d}_{-c}$, i.e. $c = \hat{\theta} - a$, $d = b - \hat{\theta}$.

So what does $\hat{\theta} = 80.25^{+0.31}_{-0.25}$ mean? It does not mean:

$P(80.00 < \theta < 80.56) = 1 - \alpha - \beta$, but rather:

repeat the experiment many times with same sample size,

construct interval according to same prescription each time,

in $1 - \alpha - \beta$ of experiments, interval will cover θ.

POISSON PROCESS

Find the hypothetical value of s such that there is a given small probability, say, $\gamma = 0.05$, to find as few events as we did or less:

$$\gamma = P(n \leq n_{\text{obs}}; s, b) = \sum_{n=0}^{n_{\text{obs}}} \frac{(s+b)^n}{n!} e^{-(s+b)}$$

Solve numerically for $s = s_{\text{up}}$, this gives an upper limit on s at a confidence level of $1-\gamma$.

Example: suppose $b = 0$ and we find $n_{\text{obs}} = 0$. For $1-\gamma = 0.95$,

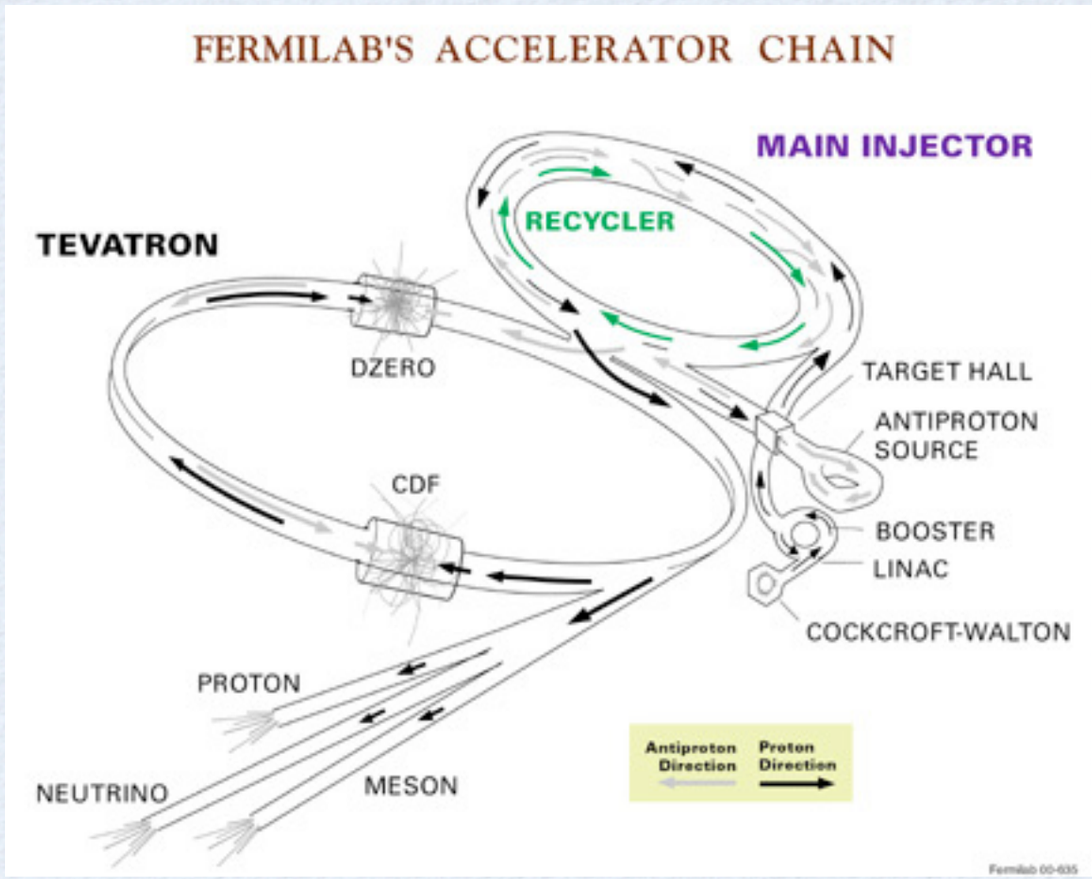
$$\gamma = P(n = 0; s, b = 0) = e^{-s} \quad \rightarrow \quad s_{\text{up}} = -\ln \gamma \approx 3.00$$

EXAMPLES

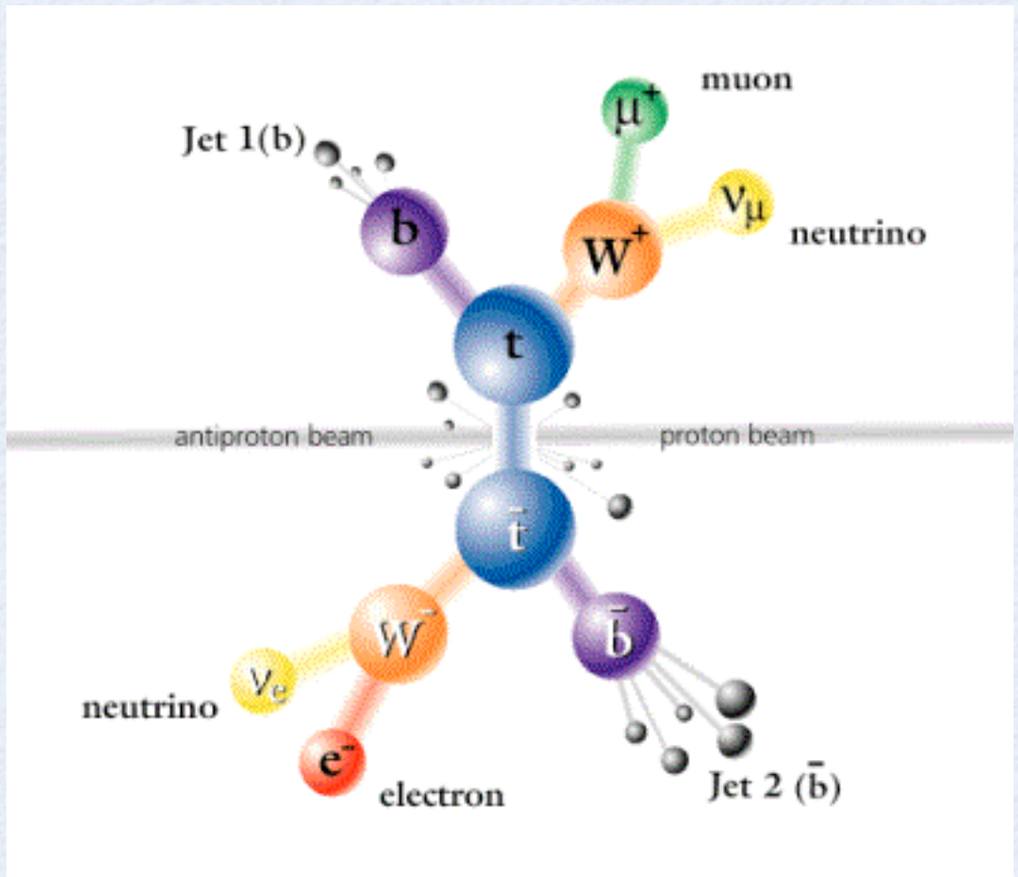
- Observation of top quark
- Limits on Higgs

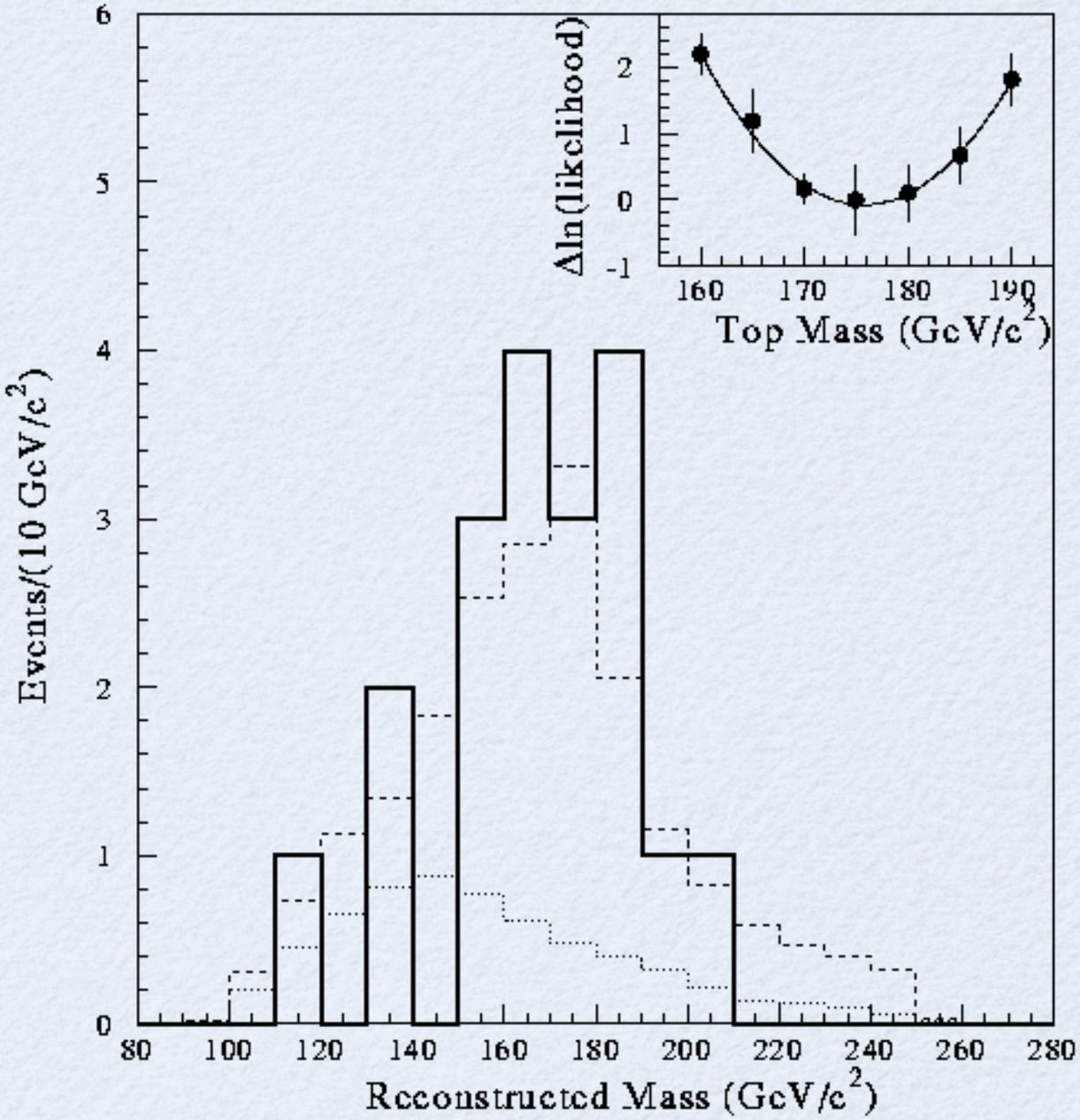
TOP QUARK DISCOVERY

- Top quark discovered in 1995
- D0 / CDF



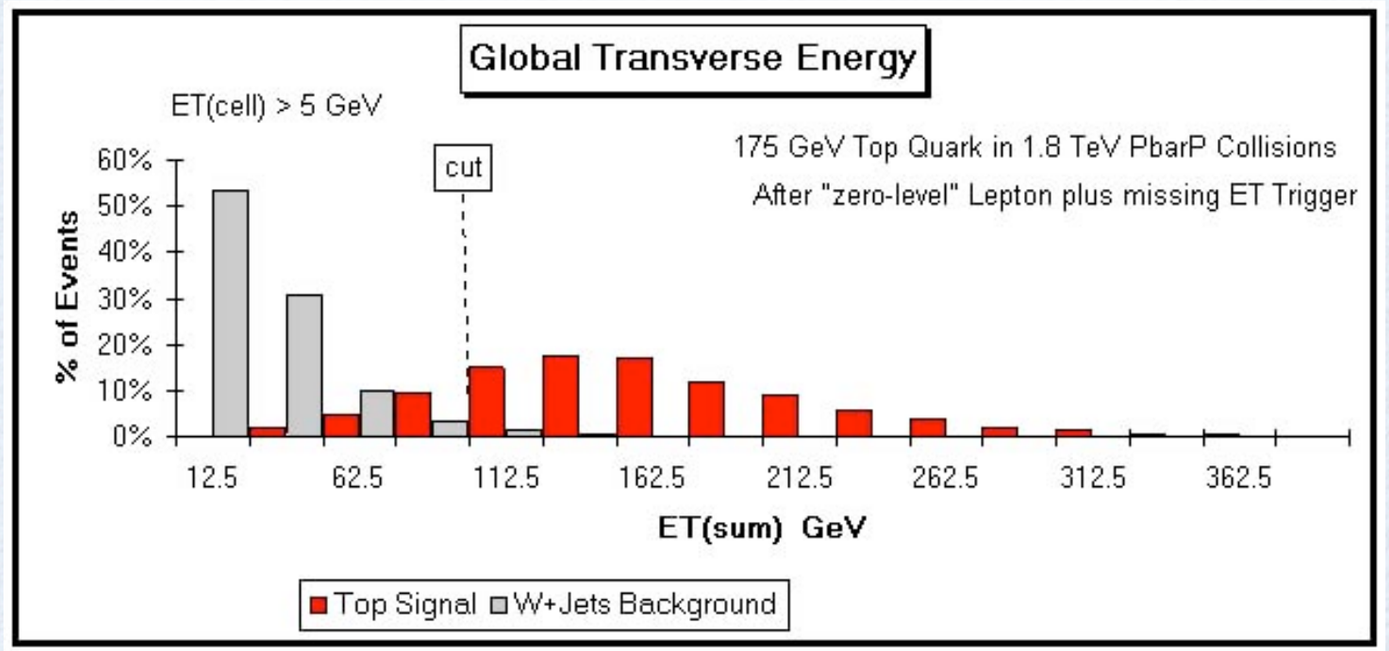
TOP QUARK


Jet 1(b)
muon
μ^+
ν_μ
neutrino
b
W^+
t
antiproton beam
proton beam
$\bar{t}$
W^-
$\bar{b}$
neutrino
ν_e
e^-
electron
Jet 2 ($\bar{b}$)


Events/(10 GeV/c^2)
Reconstructed Mass (GeV/c^2)
Δln(likelihood)
Top Mass (GeV/c^2)

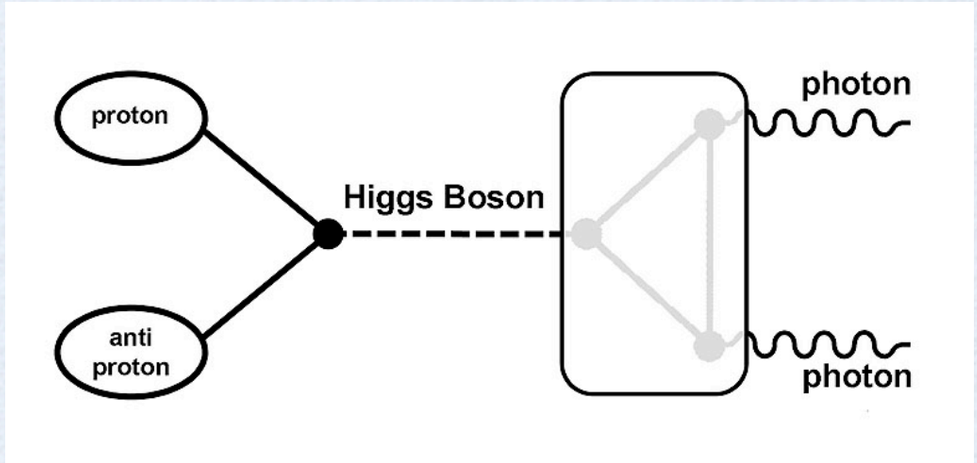
OTHER VARIABLES

- Top quark is heaviest elementary particle
- Sum of transverse 'energy' or momentum of decay products is large because the decay products are more energetic
- Compare background to signal +background with data

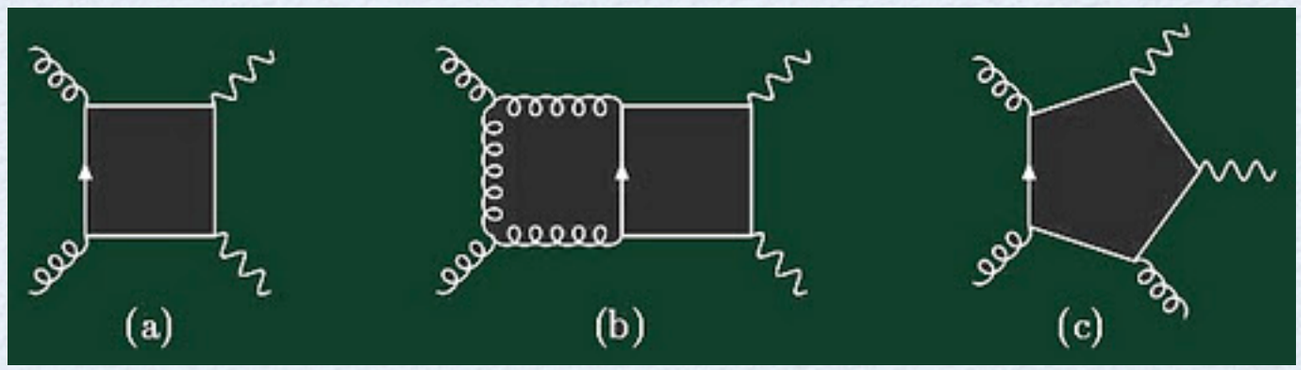


HIGGS SEARCHES

- One decay channel of the Higgs is into two photons

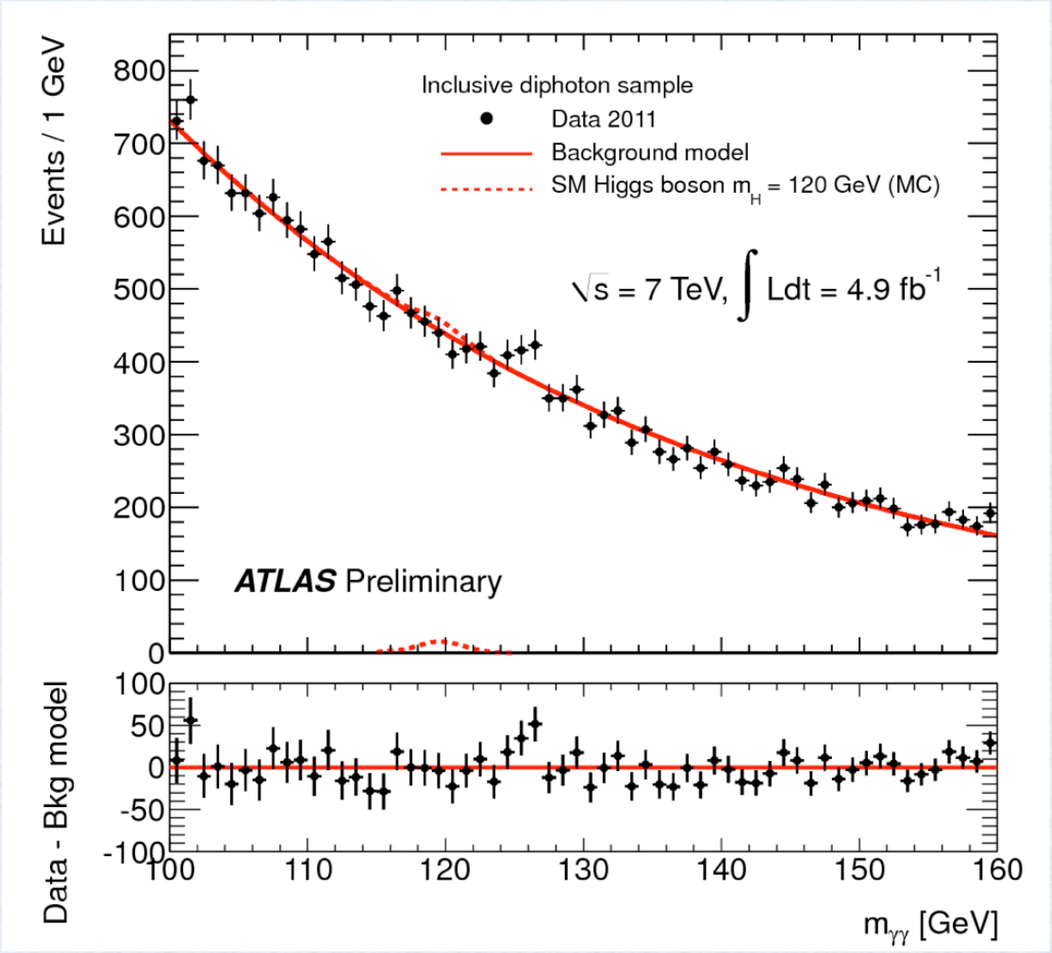


- Background is (mostly) production of diphotons

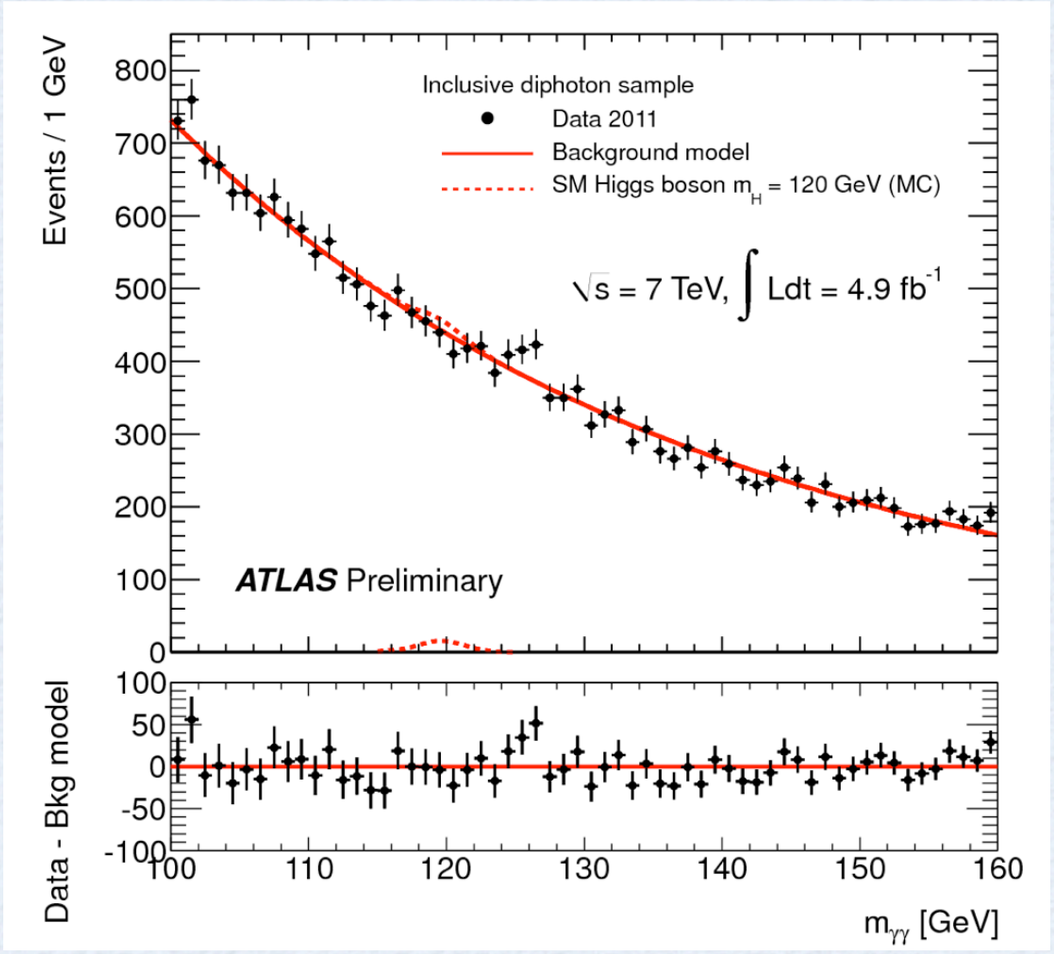


DATA

- This is the spectra from ATLAS using the 2011 data
- Small excess at 125 GeV



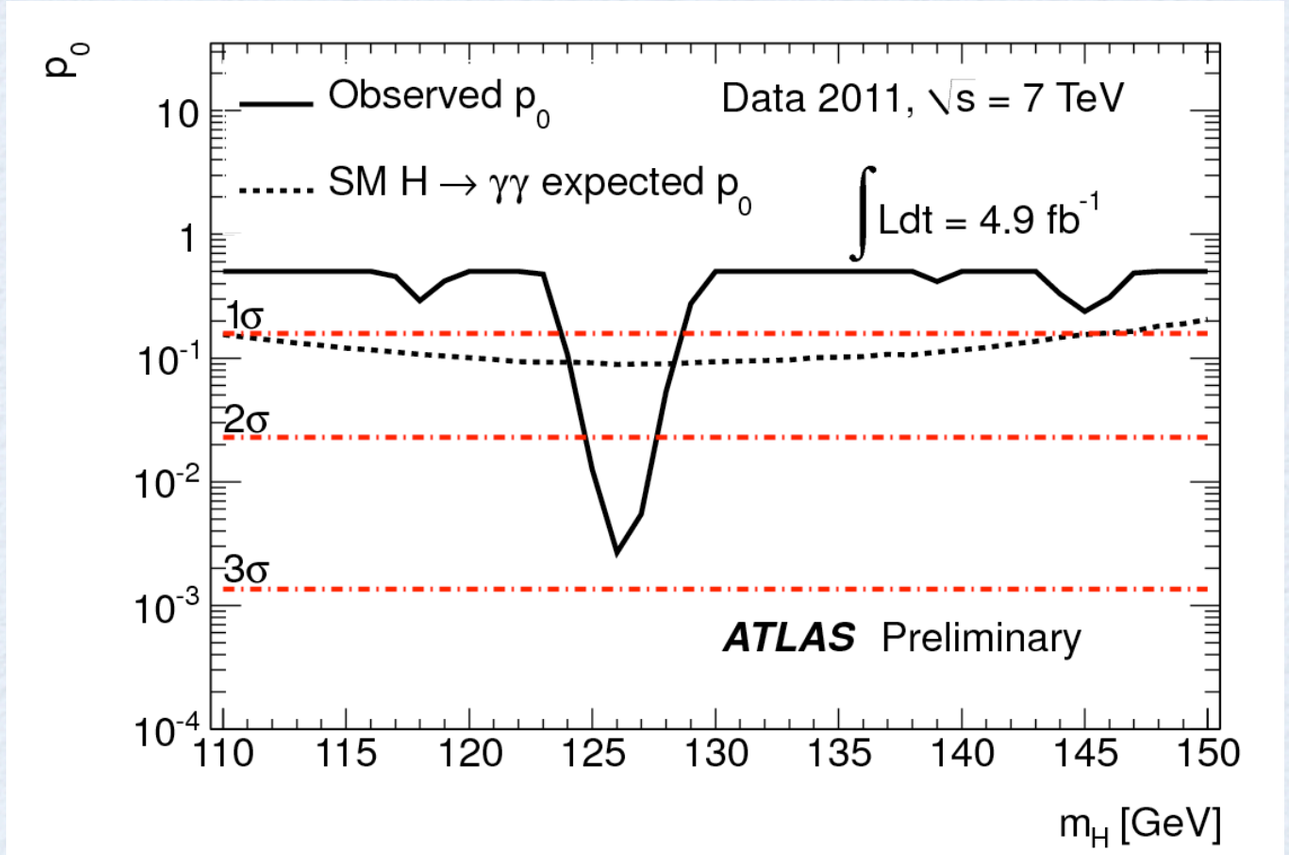
WHAT DO WE CONCLUDE



- Note that there are other deviations from the background model
- Are these deviations :
  - statistical in nature?
  - due to a mis-modeling of the background?
  - due to a new particle ?

FIRST CHECK

- After including full uncertainties in the model and the data
- p-value is smallest for the 125 GeV
- Note that this is a "local" p-value

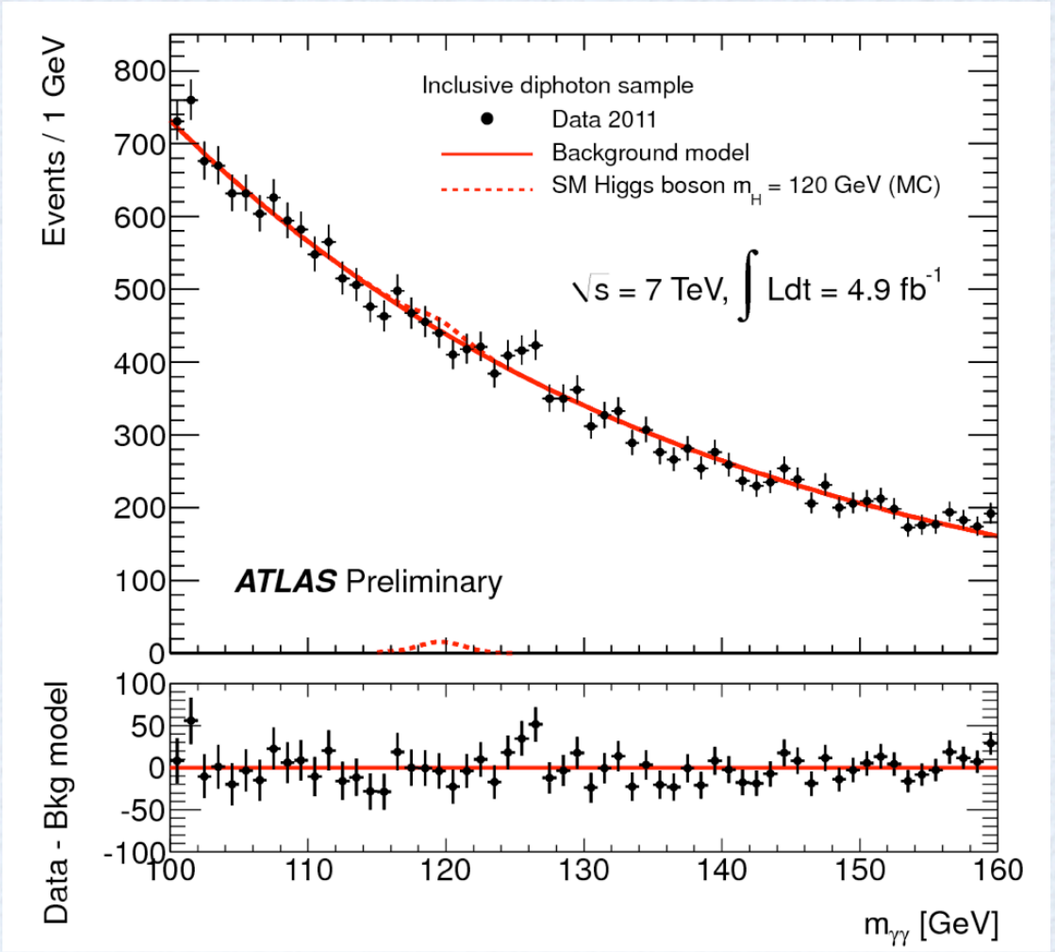


LOOK ELSEWHERE

- The previous slide shows the probability for each bin to have fluctuated given the null hypothesis (background only)

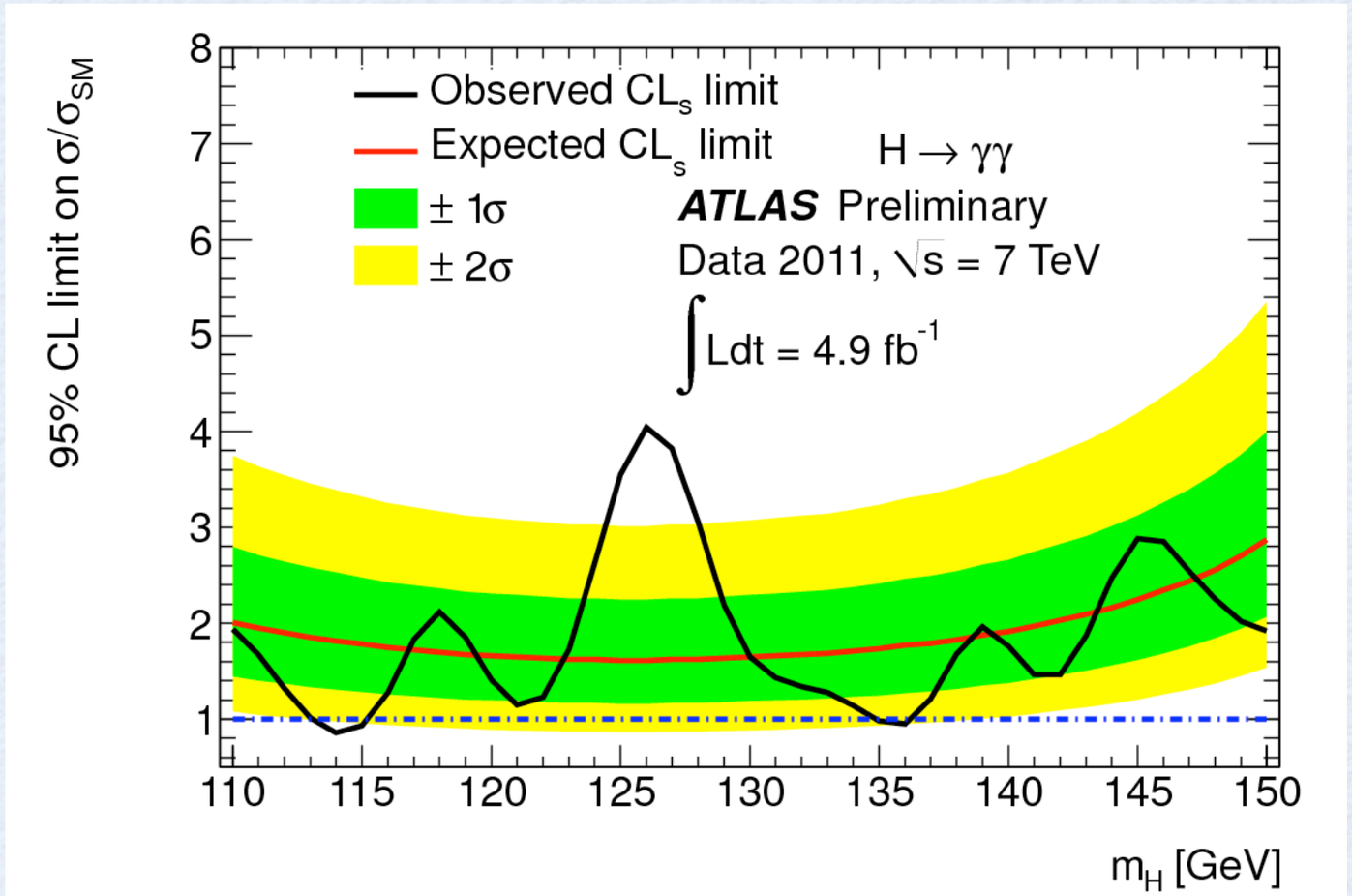
- However - it does not take into account that a fluctuation could have come from any bin

- For example - if you had reason to believe that it was at 125 GeV you might just be interested in that particular bin



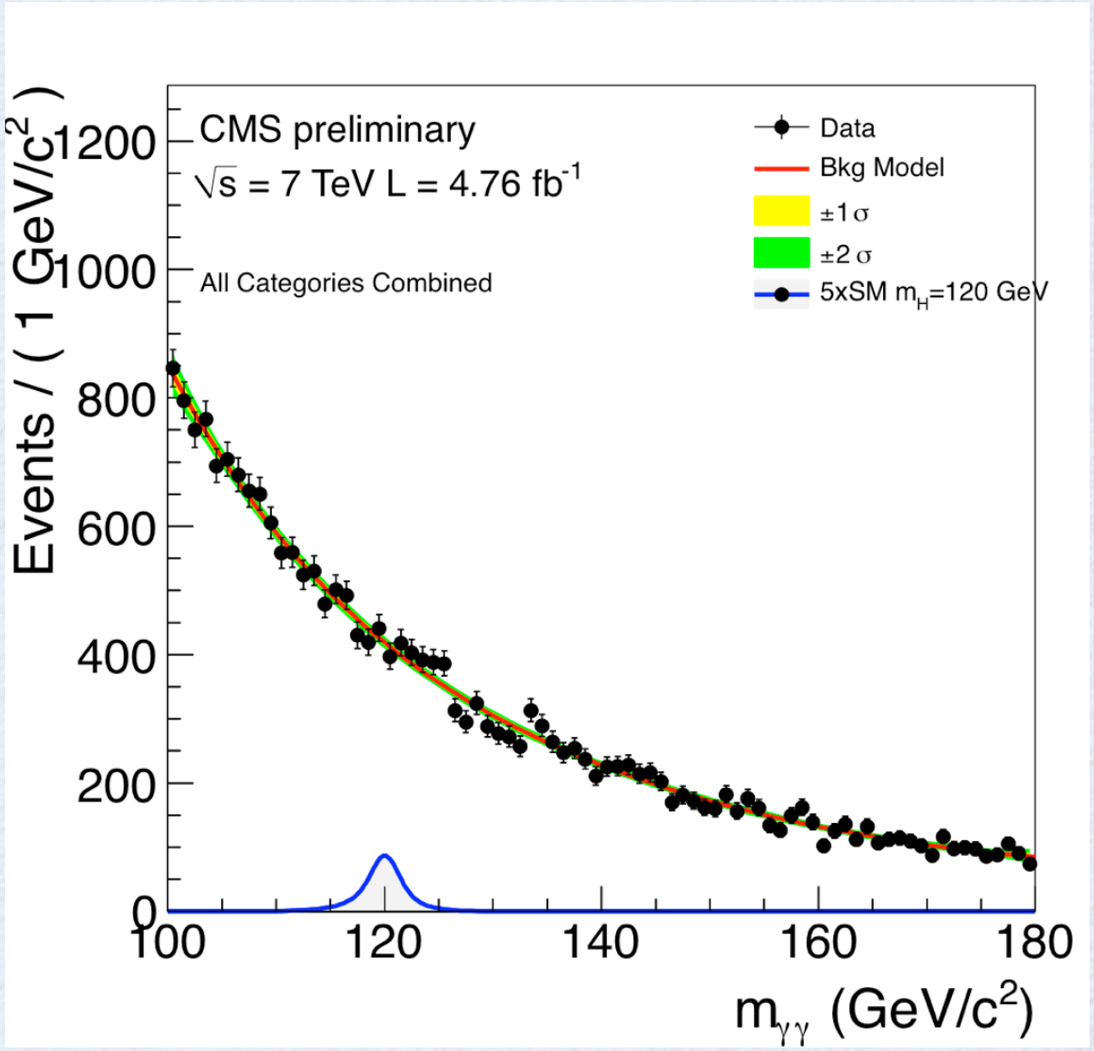
- Assuming that it is just a fluctuation you can turn that non-observation in to a limit on the Higgs Boson.

- Normalized to changing cross-section such that when the black line is below 1 that is placing a 95% CL limit



WHAT ABOUT CMS

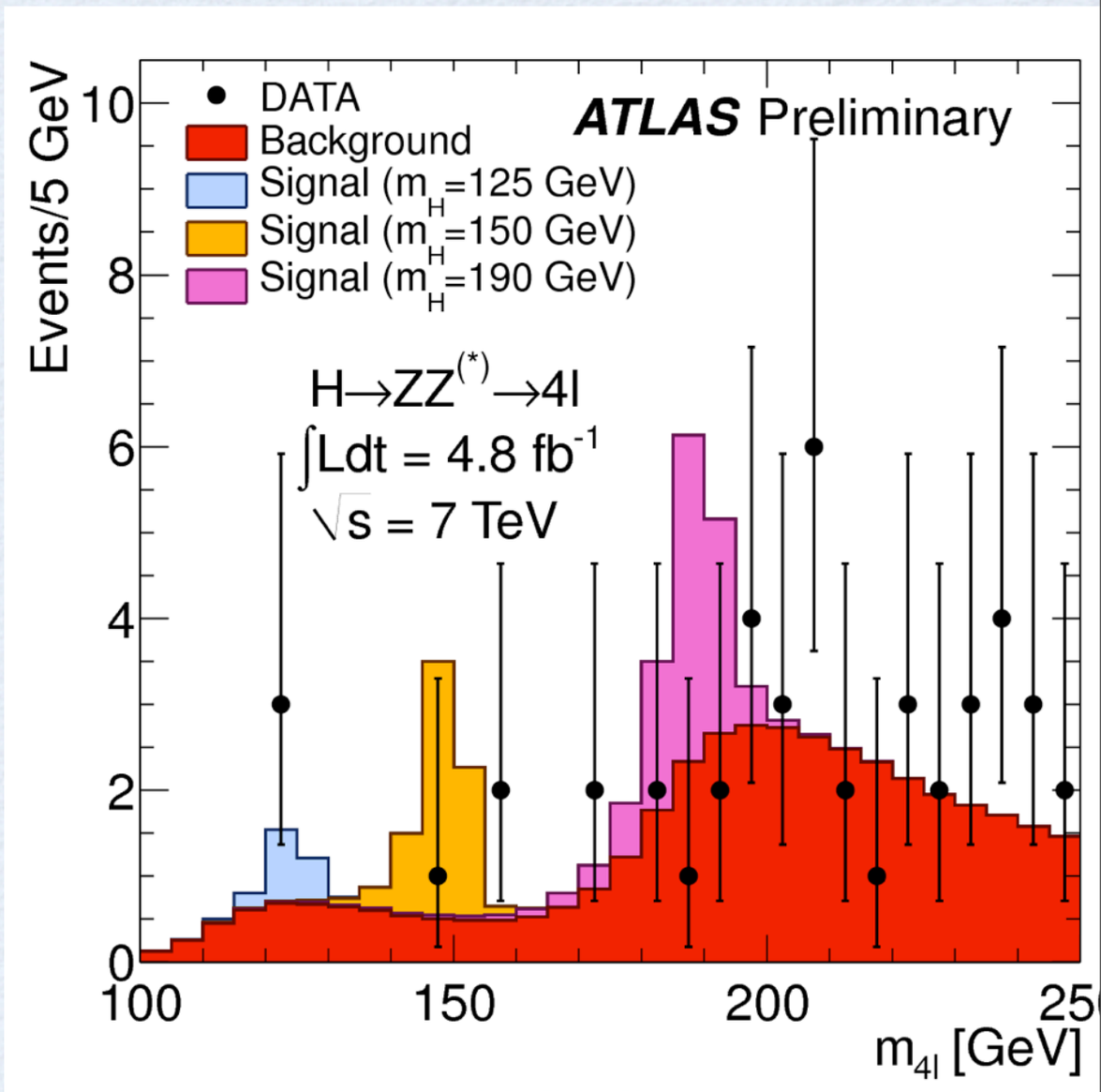
- CMS sees a (less significant) slight bump in the diphoton spectrum
- At first glance this may seem to support what ATLAS sees
  - but - background models are correlated
  - not so unlikely to get a 2 sigma'ish bump that sort of aligns with ATLAS
- What about other decay channels



ZZ

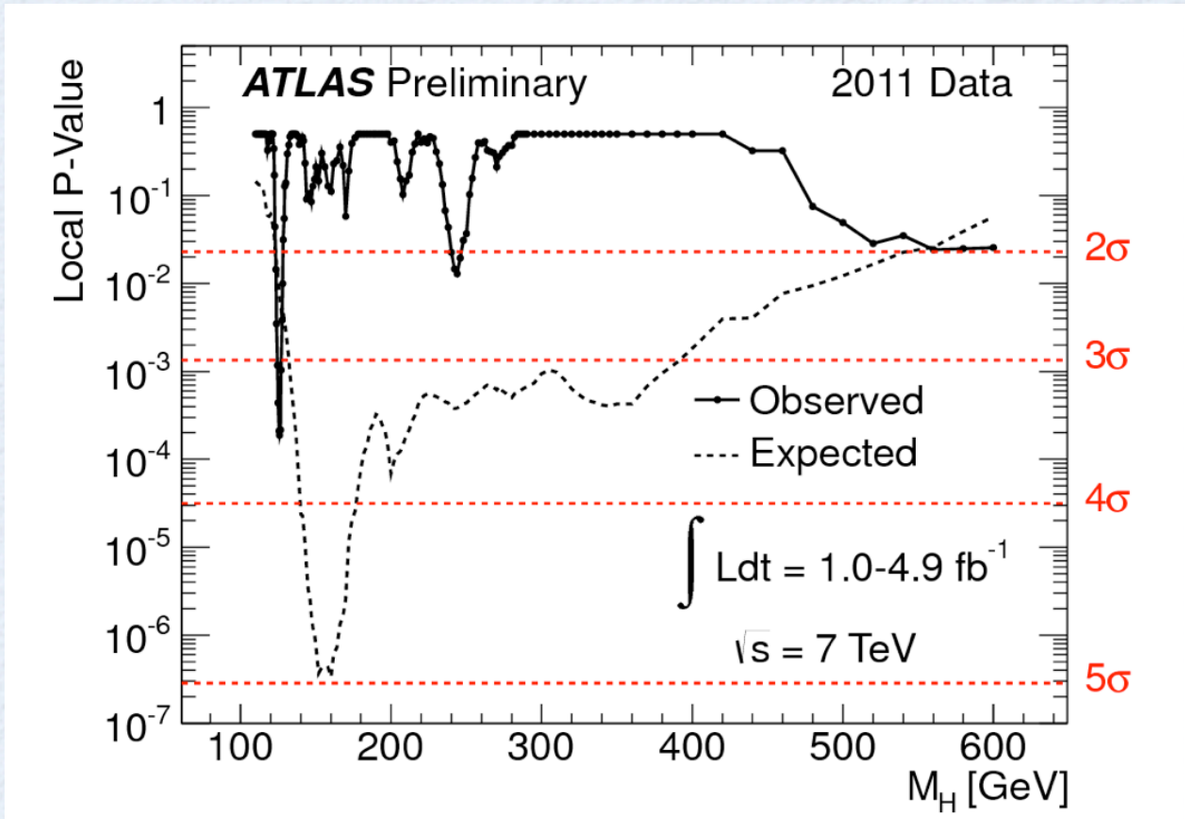
- Tantalizing that we have 3 events close invariant mass at approximately same mass as small bump in diphoton channel

- 2 sigma or so by itself



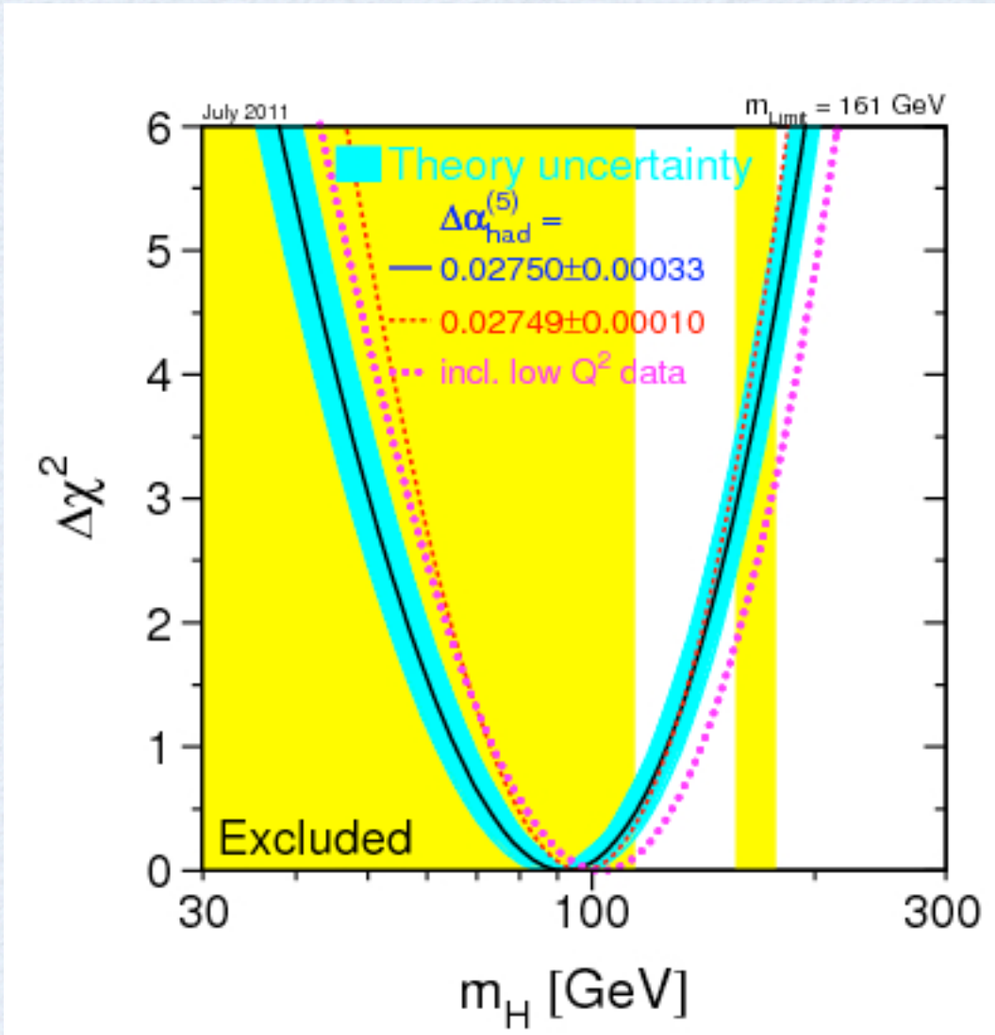
COMBINATION

- Combining the two channels enhances the discrepancy at 125 GeV
- Again - after look elsewhere effect reduced to 2.5 sigma



MORE INTRIGUING

- Data from LEP prefers low mass region
- Interestingly most preferred region is excluded!
- 125 GeV certainly allowed



SUMMARY

- 125 GeV hint at 2-3 sigma level
- Consistent with cross-sections of SM Higgs
- Consistent with theory and other data
- 2012 Data should tell us the answer!

Next few days

- More examples of analysis
- Presentations
- Post final project which is due a week after last class