Statistical Analysis at DØ

Reinhard Schwienhorst

MICHIGAN STATE UNIVERSITY

UDØ lecture, October 7 2010

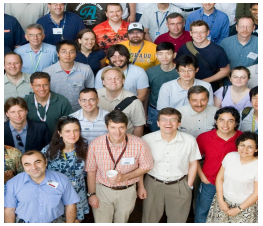
Outline

- Top discovery
- Frequentist statistics
- Modified Frequentist
- Bayesian analysis
- Systematic uncertainties
- Quick hits
- Conclusions

Statistics Example

+

= N



?

Example: Top quark discovery

LUME 74, NUMBER 14 PHYSICAL REVIEW LETTERS 3 APRIL 19

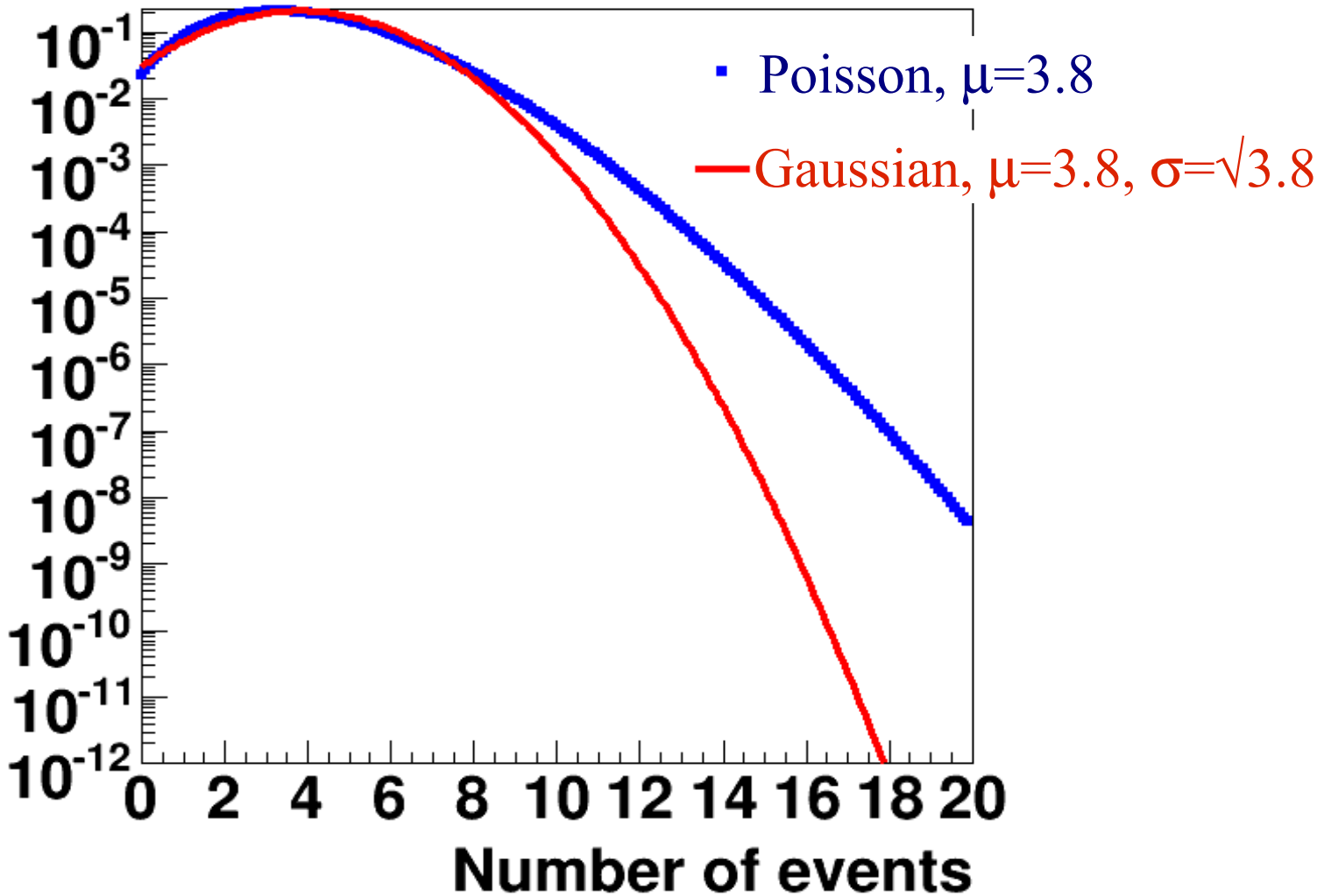
Observation of the Top Quark

(Received 24 February 1995)

The D0 Collaboration reports on a search for the standard model top quark in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV at the Fermilab Tevatron with an integrated luminosity of approximately 50 pb^{-1}. We have searched for $t\bar{t}$ production in the dilepton and single-lepton decay channels with and without tagging of b-quark jets. We observed 17 events with an expected background of 3.8 ± 0.6 events. The probability for an upward fluctuation of the background to produce the observed signal is 2×10^{-6} (equivalent to 4.6 standard deviations). The kinematic properties of the excess events are consistent with top quark decay. We conclude that we have observed the top quark and measured its mass to be 199^{+19}_{-21} (stat) ± 22 (syst) GeV/c^2 and its production cross section to be 6.4 ± 2.2 pb.

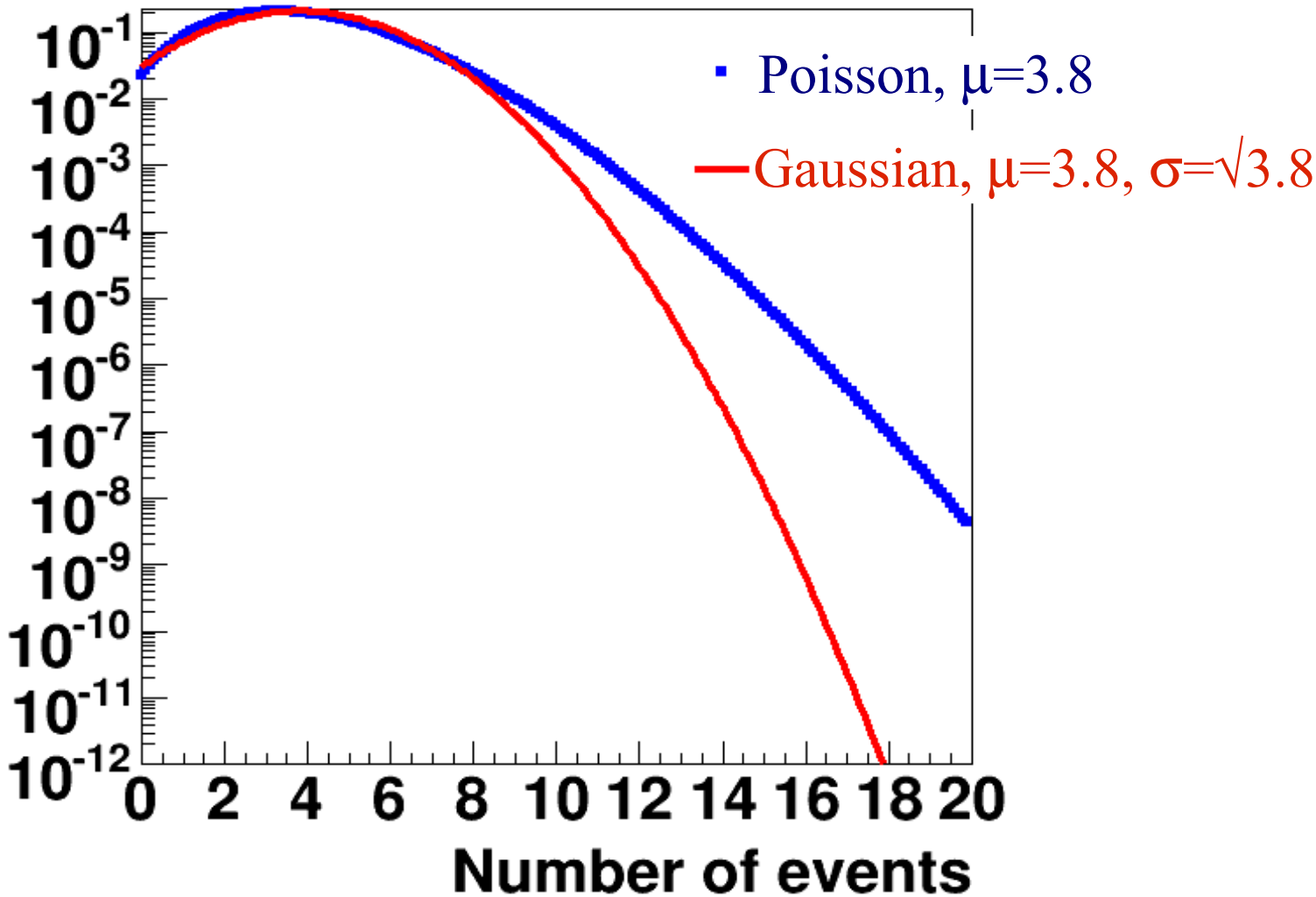
- 17 observed events
  - Expected background of 3.8 ± 0.6 events
  - Statistical fluctuation of the background: ± 1.9 events
    $\rightarrow$ statistics dominated problem

Probability density



- Quote significance based on s/√b? - No!

Probability density



- Basic assumption of particle physics:

  Event counts follow a Poisson distribution

How likely is it to observe 17 events?

- Value at 17 (10^{-7})? or fraction at 17 over peak ($10^{-7}/10^{-1}$)?
- Not an appropriate question to ask!

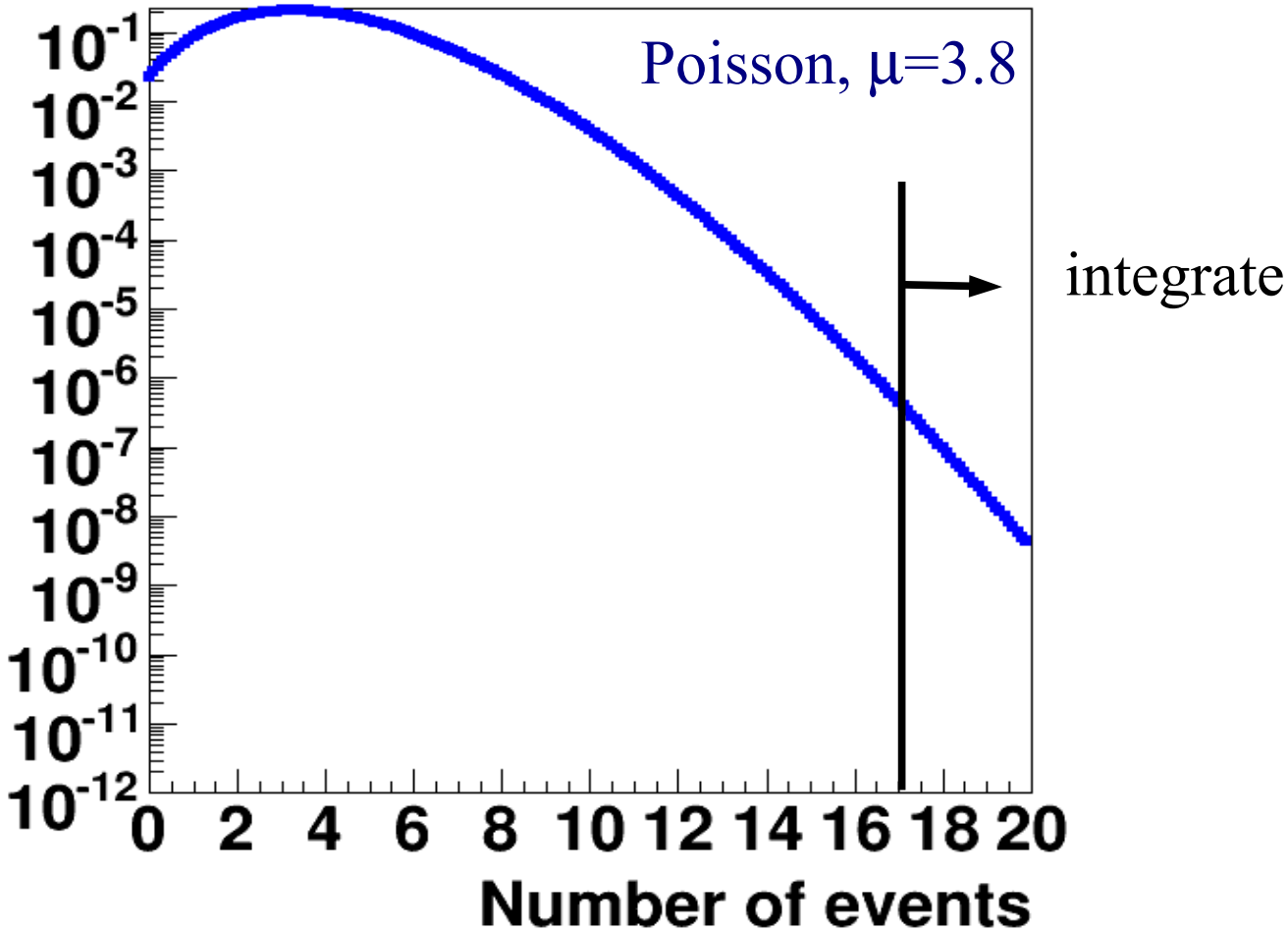
- Deep Thought, what is the right question to ask for the ultimate significance and everything?

Statistics textbook: p-value

- Proper question to ask:
  How likely is it that an expected background (true value) of 3.8 events fluctuates up to 17 events or more?"

- Wikipedia:
  "... probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true."

P-value calculation



- p-value in Root: gSystem->Load("libMathCore");
  1-ROOT::Math::poisson_cdf(17,3.8)
  1.19E-7
- Corresponds to 5σ significance

Truth vs Data

- Frequentist statistics:
  - True value is fixed but unknown
  - Measured value is one of an infinite series of statistically independent measurements
  - Probability is the limiting relative frequency of a certain outcome

$$P(A) = \lim_{n \to \infty} \frac{\#\text{ of outcome A in } n \text{ measurements}}{n}$$

  - 95% Confidence level interval:
    Contains the true value in 95% of many repeated experiments

Limit setting

Limit setting example

VOLUME 93, NUMBER 14 PHYSICAL REVIEW LETTERS week ending 1 OCTOBER 2004

Search for Doubly Charged Higgs Boson Pair Production in the Decay to $\mu^+\mu^+\mu^-\mu^-$ in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV

A search for pair production of doubly-charged Higgs bosons in the process $p\overline{p} \to H^{++}H^{--} \to \mu^+\mu^+\mu^-\mu^-$ is performed with the DØ Run II detector at the Fermilab Tevatron. The analysis is based on a sample of inclusive di-muon data collected at an energy of $\sqrt{s} = 1.96$ TeV, corresponding to an integrated luminosity of 113 pb^{-1}. In the absence of a signal, 95% confidence level mass limits of $M(H_L^{\pm\pm}) > 118.4$ GeV/c^2 and $M(H_R^{\pm\pm}) > 98.2$ GeV/c^2 are set for left-handed and right-handed doubly-charged Higgs bosons, respectively, assuming 100% branching into muon pairs.

- Example: background 1.5, observe 3 events
  - What is the limit on the doubly charged Higgs cross section?

Limit setting example

VOLUME 93, NUMBER 14 PHYSICAL REVIEW LETTERS week ending 1 OCTOBER 2004

Search for Doubly Charged Higgs Boson Pair Production in the Decay to $\mu^+\mu^+\mu^-\mu^-$ in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV

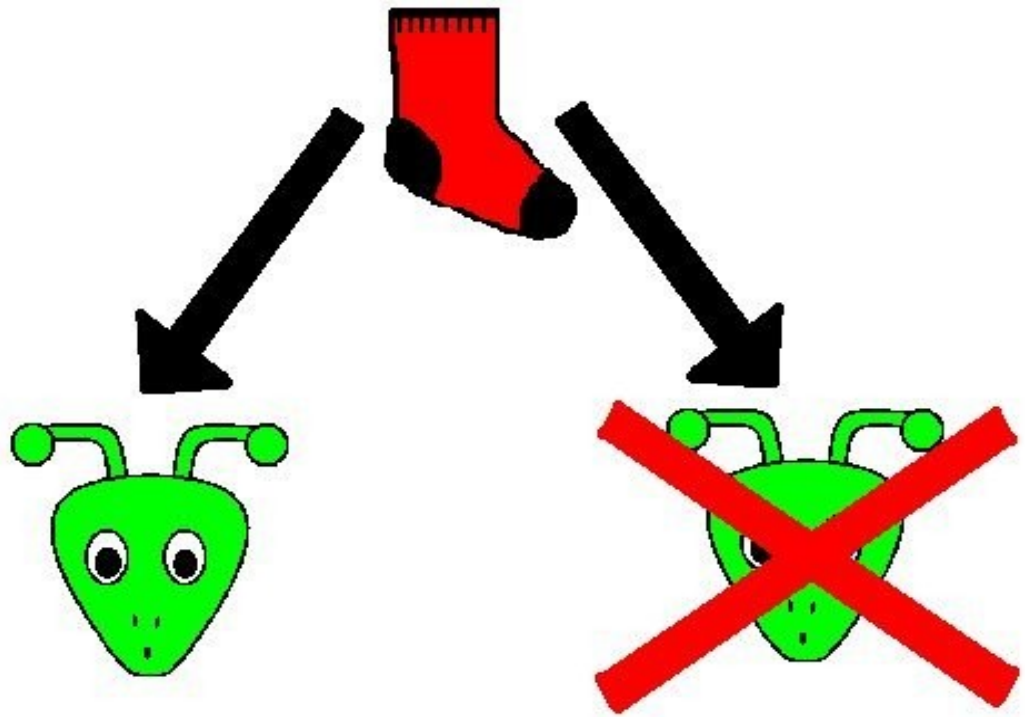
A search for pair production of doubly-charged Higgs bosons in the process $p\overline{p} \to H^{++}H^{--} \to \mu^+\mu^+\mu^-\mu^-$ is performed with the DØ Run II detector at the Fermilab Tevatron. The analysis is based on a sample of inclusive di-muon data collected at an energy of $\sqrt{s} = 1.96$ TeV, corresponding to an integrated luminosity of 113 pb^{-1}. In the absence of a signal, 95% confidence level mass limits of $M(H_L^{\pm\pm}) > 118.4$ GeV/c^2 and $M(H_R^{\pm\pm}) > 98.2$ GeV/c^2 are set for left-handed and right-handed doubly-charged Higgs bosons, respectively, assuming 100% branching into muon pairs.

- Example: background 1.5, observe 3 events
  - What is the limit on the doubly charged Higgs cross section?

Hypothesis testing

Hypothesis testing in daily life

Q. Where have all my socks gone?



Alternate Hypothesis = Extra-terrestrial beings have transported themselves into my house in order to steal my socks.

Null Hypothesis = Aliens are not to blame. There is some other explanation for the disappearing socks.

Hypothesis testing

- Null hypothesis
  - The old, default, already known
  - Data are described by background-only

- Alternative hypothesis
  - Something new, fresh, exciting
  - Data can only be described by signal + background

- Type I error: Null hypothesis is true but rejected (α)

- Type II error: Alternative hypothesis is true but null hypothesis is not rejected (β)

Can I reject the null hypothesis?

1. Form a test statistic
2. Define the critical region
   - The range of possible values of the test statistic for which we reject the null hypothesis
   - Normally determined before conducting the experiment:
   - Want to minimize the Type I error rate
   - Typically a 95% confidence level (CL) interval or limit
     - Type I error rate 5% or less
3. Test if observed data are in the critical region
   - Reject the null hypothesis or fail to reject it
   - Typically vary cross section or similar parameter and determine for which XS we reach 95%

Limit setting example

1. Form test statistic

- Neyman-Pearson lemma:
  The most powerful test statistic is the likelihood ratio

$$LR = \frac{L(signal + background)}{L(background)}$$

Limit setting example

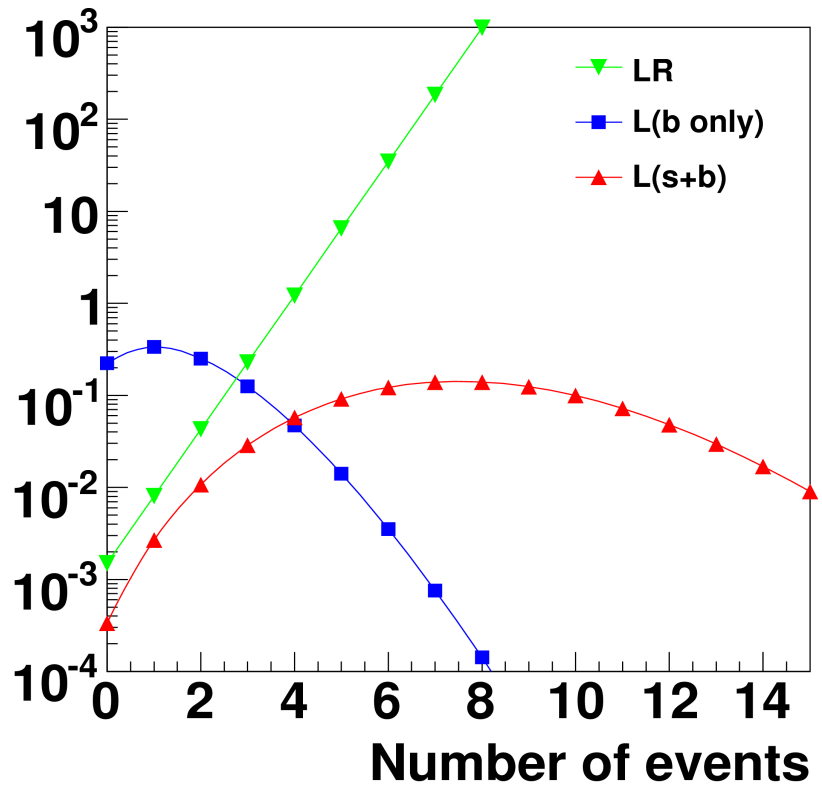
1. Form test statistic

- Neyman-Pearson lemma:
  The most powerful test statistic is the likelihood ratio

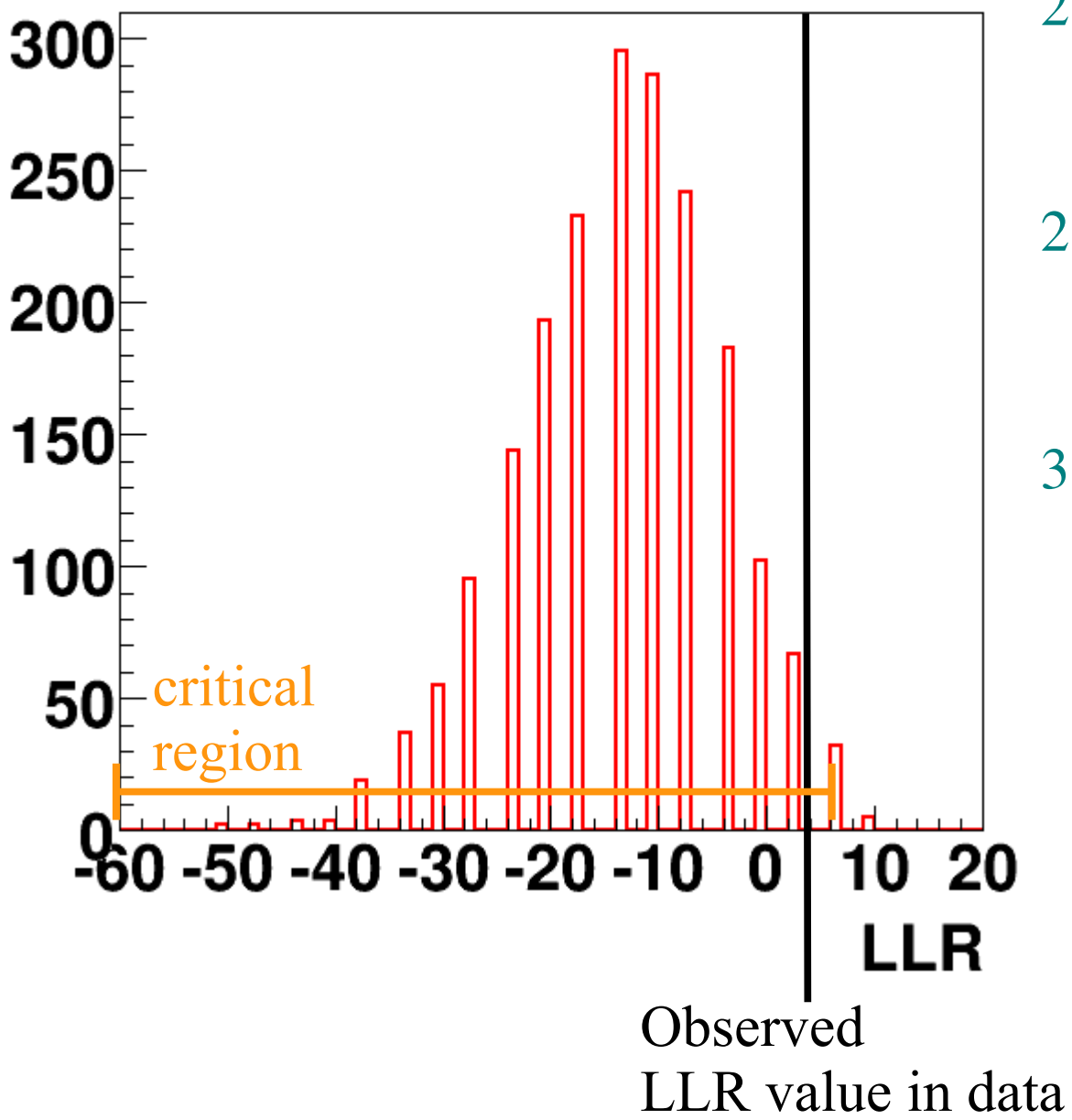
$$LR = \frac{L(signal + background)}{L(background)}$$

- Example:
  L(b): Poisson(μ=1.5)
  L(s+b): Poisson(μ=8.0)
  - M(H++) = 120GeV

- Plot
  LLR = -2log(LR)



Limit setting example



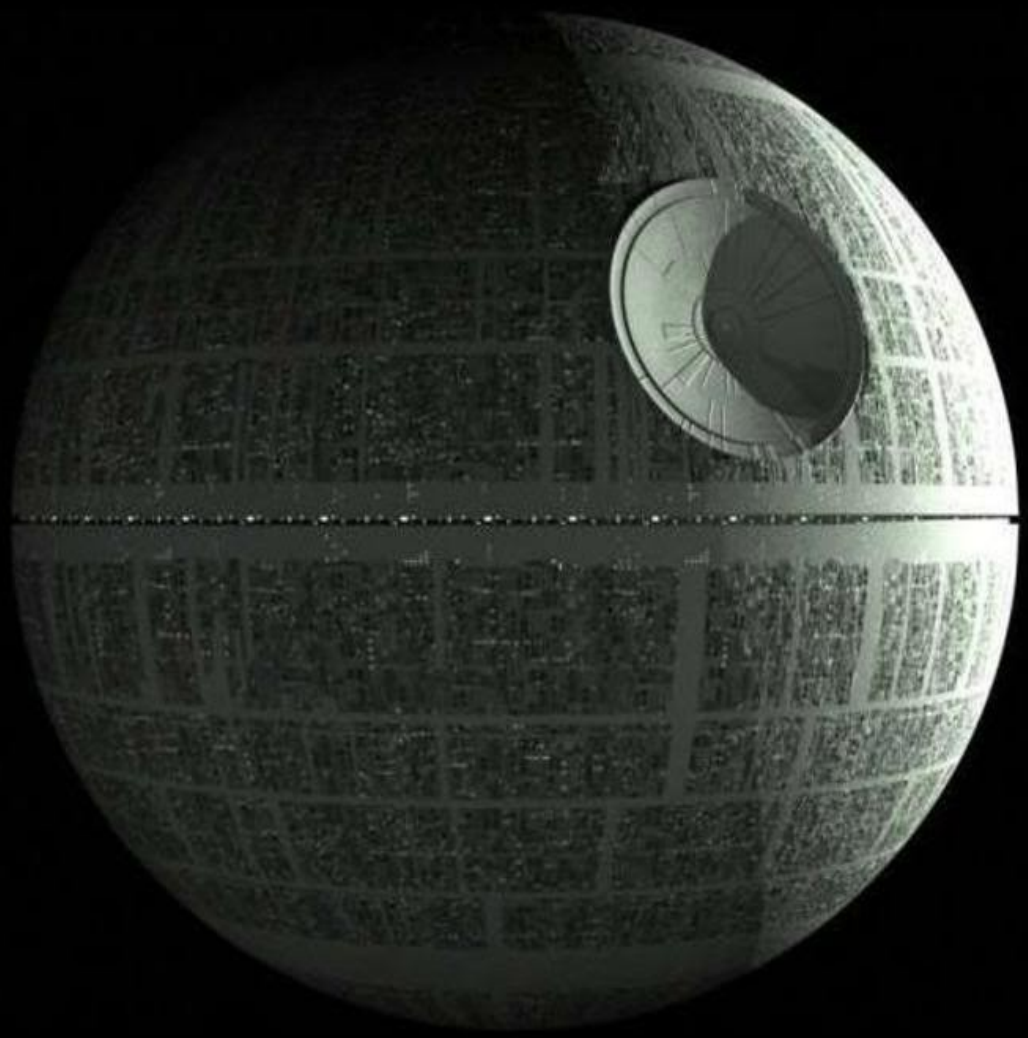
2. Critical region is the integral starting at left covering 95% of area

2. Observed value is inside critical region

- Cannot exclude it

3. Observed confidence level CL_{s+b} is integral of LLR distribution above observed value

LEP strikes back

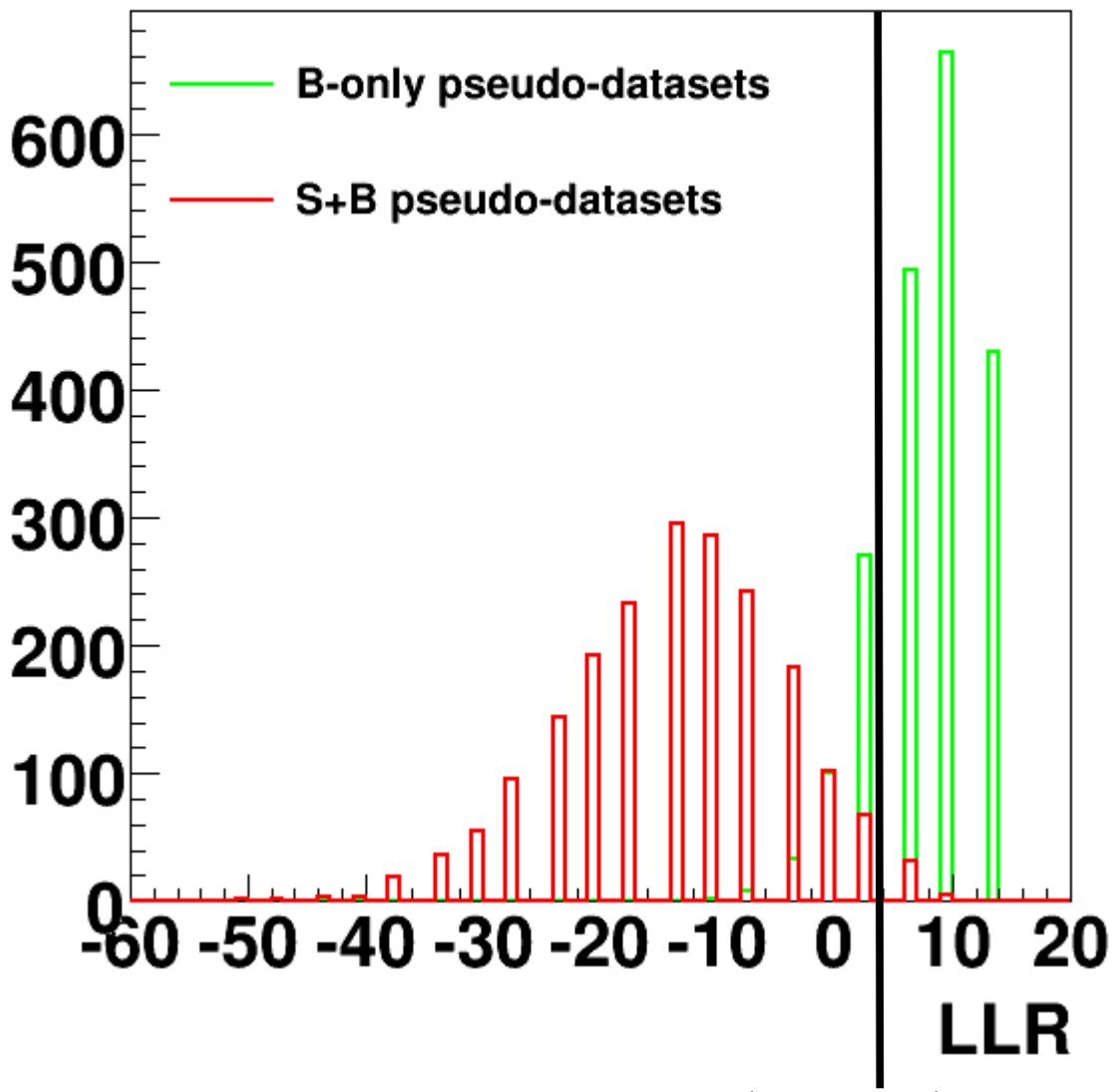


Learning from LEP

Modified Frequentist approach

- If we observe fewer events than predicted by the background model, the limit gets better
  - Something wrong with background model?
- Modify Frequentist approach to avoid this scenario
  - Also called CLs
- Procedure:
  - Generate ensembles of pseudo-datasets
    - Signal+background
    - Background only
  - Determine LLR = -2log(LR) value for each of them

Example LLR distribution

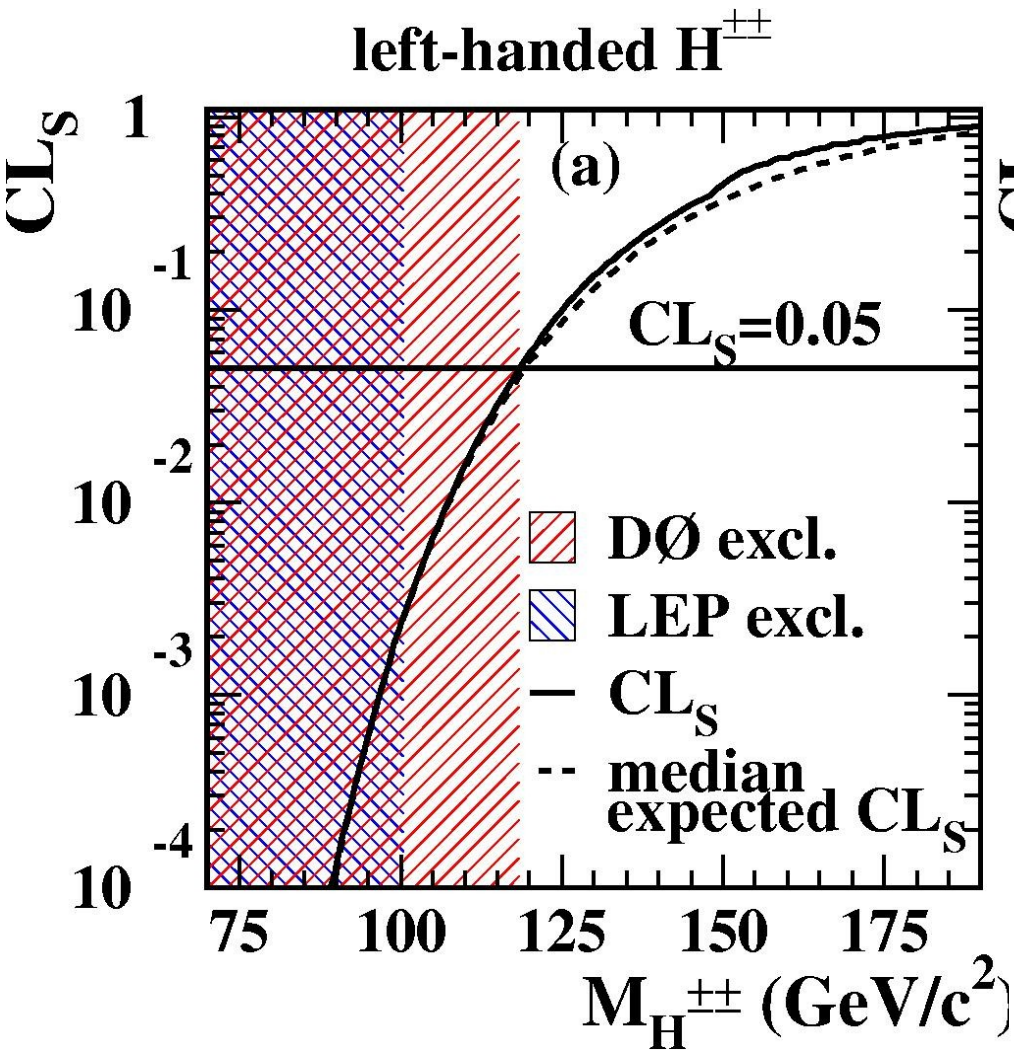


Observed
LLR value in data

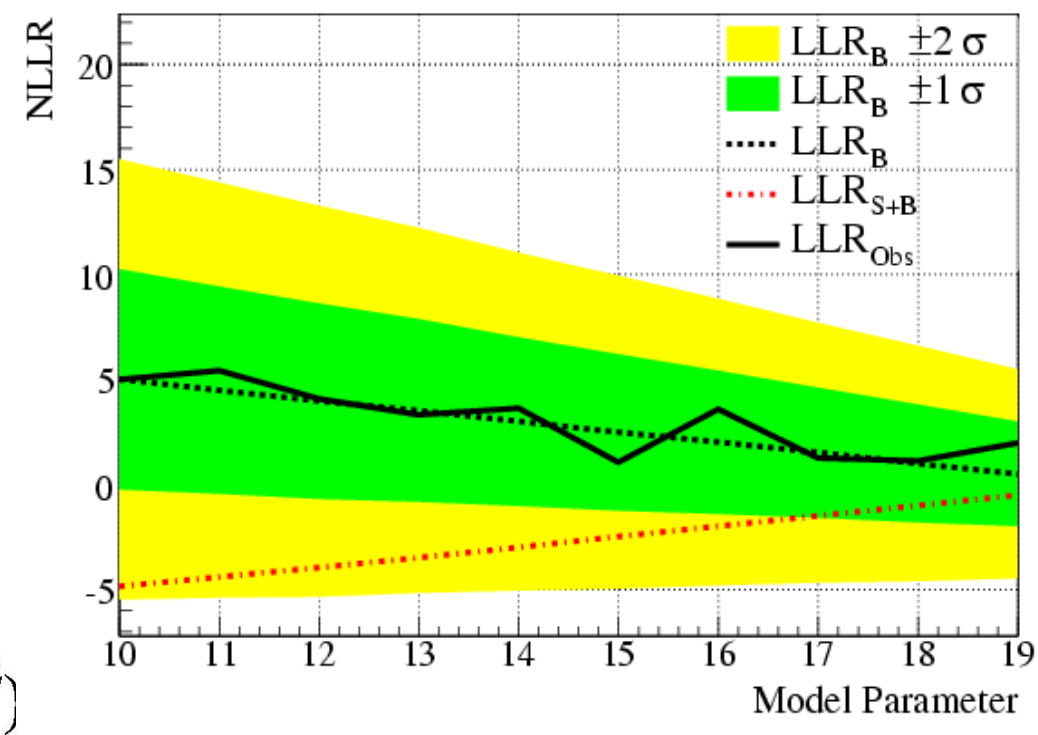
- CL_{s+b} is standard statistical procedure from statistics textbooks
- Integrate red above data to obtain $CL_{s+b} = 0.052$
- Integrate green above data to obtain $CL_b = 0.93$
- $CL_s = 0.055$

Limit setting example

5. Compute CLs for several Higgs mass



6. Define expected CLs from ensembles

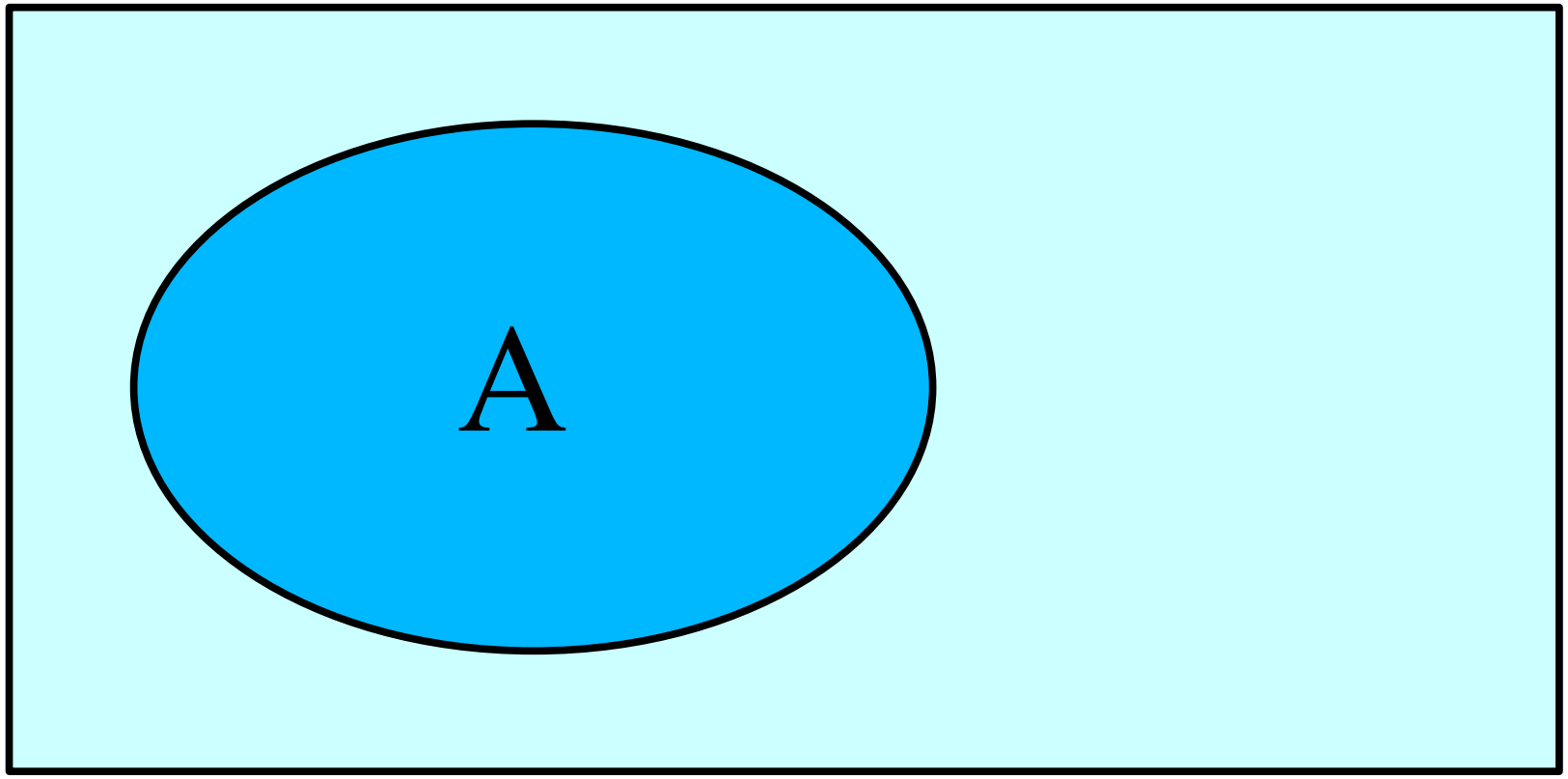


So far, we have only dealt with statistical uncertainties.

How do we include systematic uncertainties?

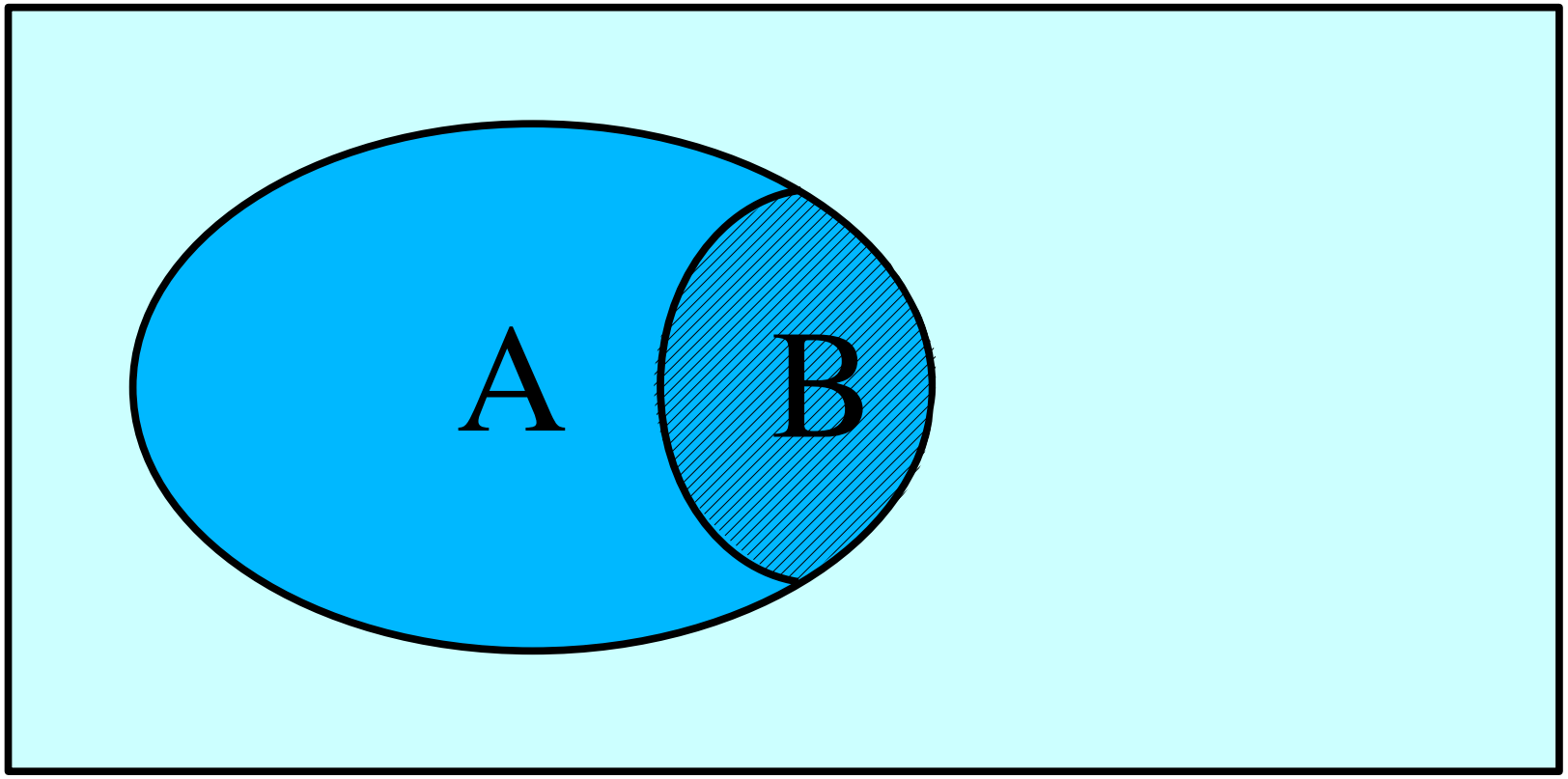
Bayesian Analysis

Simple probability



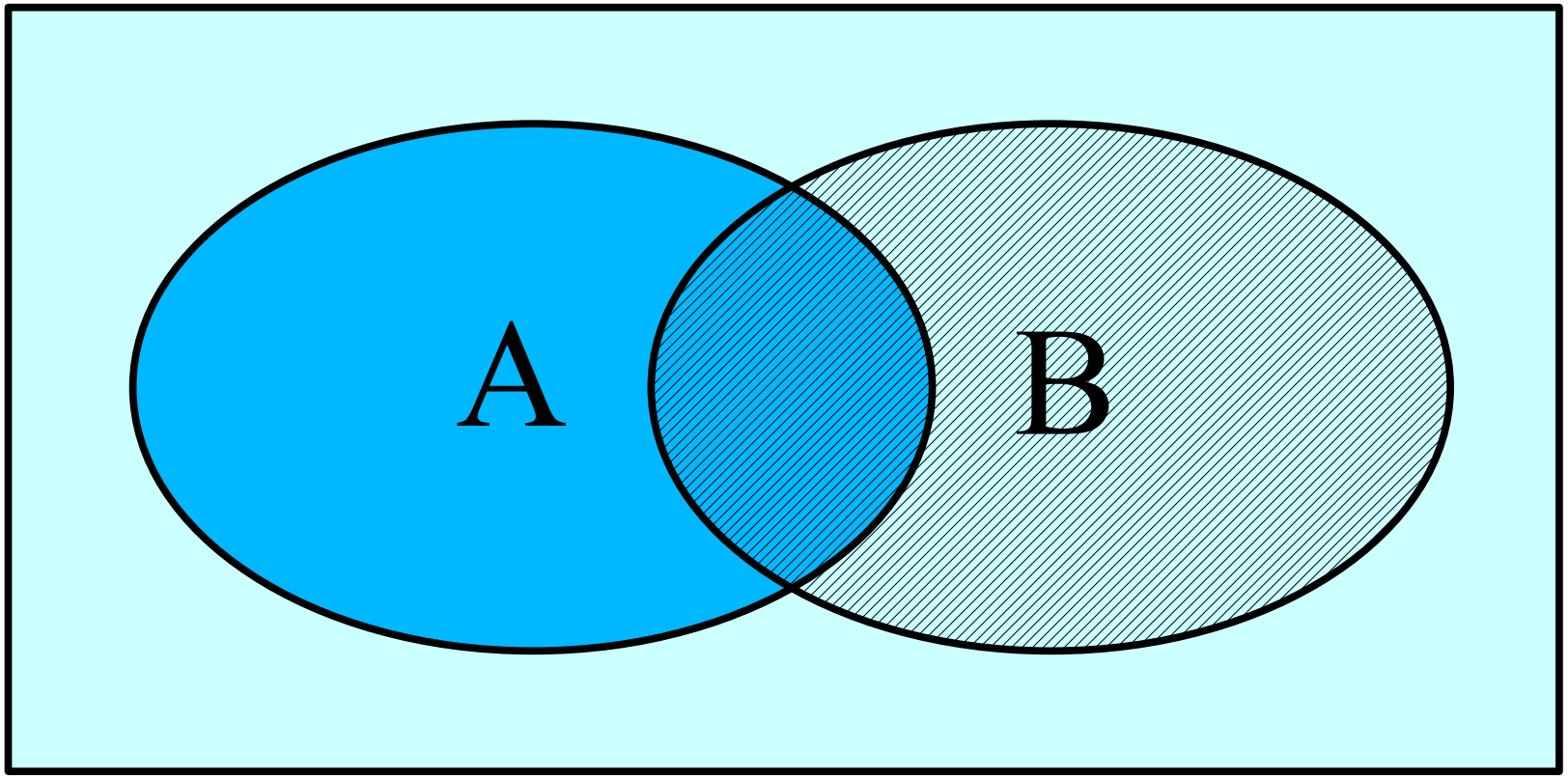
P(A): Probability that A is true

Conditional probability



P(B|A): conditional probability for B, given that A is true.

Bayes Theorem



$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

Bayesian statistical analysis

$$P(\text{signal}|\text{data}) = \frac{P(\text{data}|\text{signal},\text{bkg}) \times P(\text{signal})}{P(\text{data})}$$

"Posterior probability"

- Probability as the subjective degree of belief
  - P(A) – degree of believe in hypothesis A
  - As opposed to frequentist approach that true values can never be determined and P(A) is the result of many repeated experiments

Bayesian Statistical Analysis

$$P(\text{signal}|\text{data}) = \frac{P(\text{data}|\text{signal},\text{bkg}) \times P(\text{signal})}{P(\text{data})}$$

"Posterior probability"

Likelihood

Poisson distribution

$$L(x,\mu) = \frac{\mu^x e^{-x}}{x!}$$

$$\mu = \text{signal} + \text{bkg}$$

Bayesian Statistical Analysis

$$P(\text{signal}|\text{data}) = \frac{P(\text{data}|\text{signal},\text{bkg}) \times P(\text{signal})}{P(\text{data})}$$

"Posterior probability" | Likelihood | Normalization factor | "prior probability"

- Much discussion about the prior by statisticians
  - Often choice is not clear
    - For example different result if prior is flat in XS or flat in coupling=√XS
  - One goal for statisticians is "uninformed prior"
  - For us the choice is always prior flat in cross section
    - Only violated once so far!

VOLUME 82, NUMBER 25 PHYSICAL REVIEW LETTERS 21 JUNE 1999

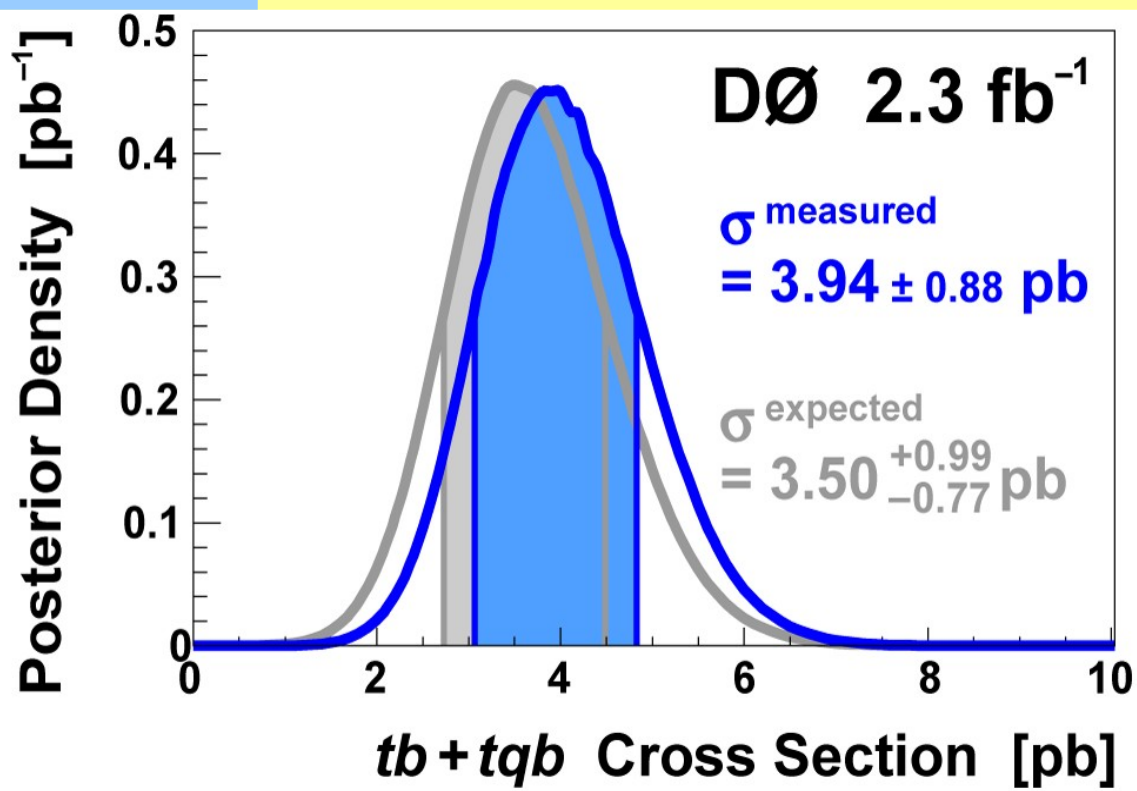
- Chose prior flat in M and tanβ

Search for Charged Higgs Bosons in Decays of Top Quark Pairs

Bayesian Statistical Analysis

$$P(\text{signal}|\text{data}) = \frac{P(\text{data}|\text{signal,bkg}) \times P(\text{signal})}{P(\text{data})}$$

"Posterior probability"



Simple Bayesian example

$$n_{predicted} = 6.5 * ratio_{XS} + n_{bkg}$$

- $N_{obs} = 3$, $n_{bkg} = 1.5$
  - i.e. rule out this Higgs mass if limit on $ratio_{XS}$ is 1
- Compute Bayesian posterior for XS using simple spreadsheet

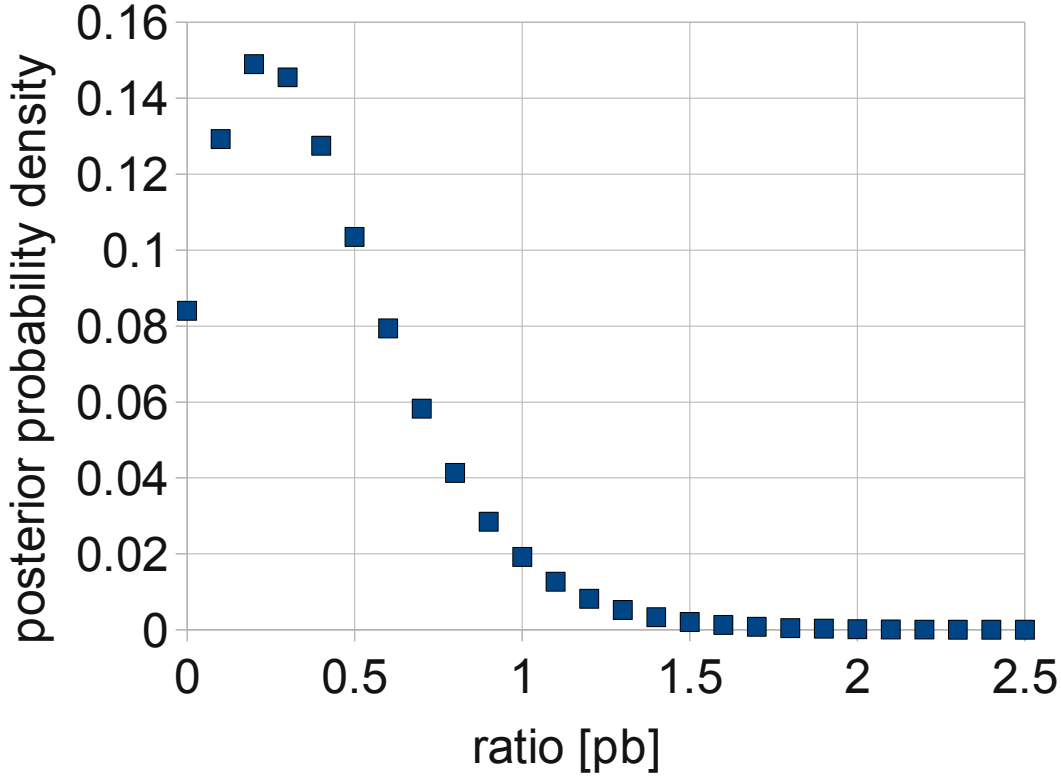
$$P(Nobs,\mu) = \frac{\mu^{Nobs}\, e^{-Nobs}}{Nobs!} \quad \text{where } \mu = 6.5*ratio_{XS} + n_{bkg}$$

- Prior is flat in XS, thus flat in $ratio_{XS}$

Bayesian example on a Spreadsheet

Nobs=3 Nbkg=1.5

XS	μ	P(ratio|μ)
0	1.5	0.13
0.1	2.15	0.19
0.2	2.8	0.22
0.3	3.45	0.22
0.4	4.1	0.19
0.5	4.75	0.15
0.6	5.4	0.12
0.7	6.05	0.09



- 95% CL limit: integrate from 0 to 95% of area
- In this simple example, limit is 0.9
  - Numerically same as LLR
  - But different interpretation!

Systematic uncertainties

Statistician says: “What do you mean you don't know exactly what your background is?
You can't even formulate a null hypothesis!?!”

Systematic uncertainties

- Neither background nor signal can be predicted precisely
  - Background prediction and other parameters are themselves probability distributions

$$L(Nobs,s,b) = \frac{(s+b)^{Nobs}\, e^{-Nobs}}{Nobs!}$$

    - s and b are not fixed anymore

$$L(s) = ?,\ \ L(b) = ?$$

  - Parameters that go into making the predictions are probability distributions – example: JES
- Nuisance parameters

Systematic uncertainties

- Bayesian approach:
  - Prior for each systematic uncertainty
  - Gaussian shape for prior, but not always symmetric

$$L(b,\mu_b,\sigma_b) = e^{-(b-\mu_b)^2/\sigma_b^2}$$

  - Evaluate systematic twice, once for $+1\sigma$ and once for -1σ
    - Only find the $\pm 1\sigma$ shifted event yield or histogram
  - Then shift yields according to the Gaussian
  - Bayesian posterior is then product of likelihoods

$$L(Nobs,\mu_s,\sigma_s,\mu_b,\sigma_b) = e\frac{(s+b)^{Nobs}\, e^{-Nobs}}{Nobs!} e^{-(s-\mu_s)^2/\sigma_s^2}\, e^{-(b-\mu_b)^2/\sigma_b^2}$$

Including systematic uncertainties

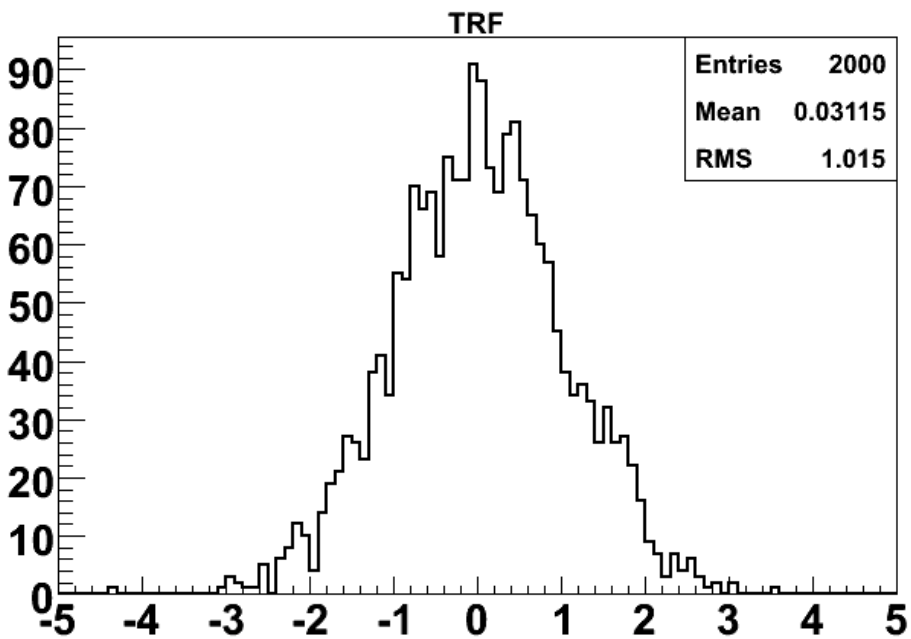
- Bayesian posterior including systematics

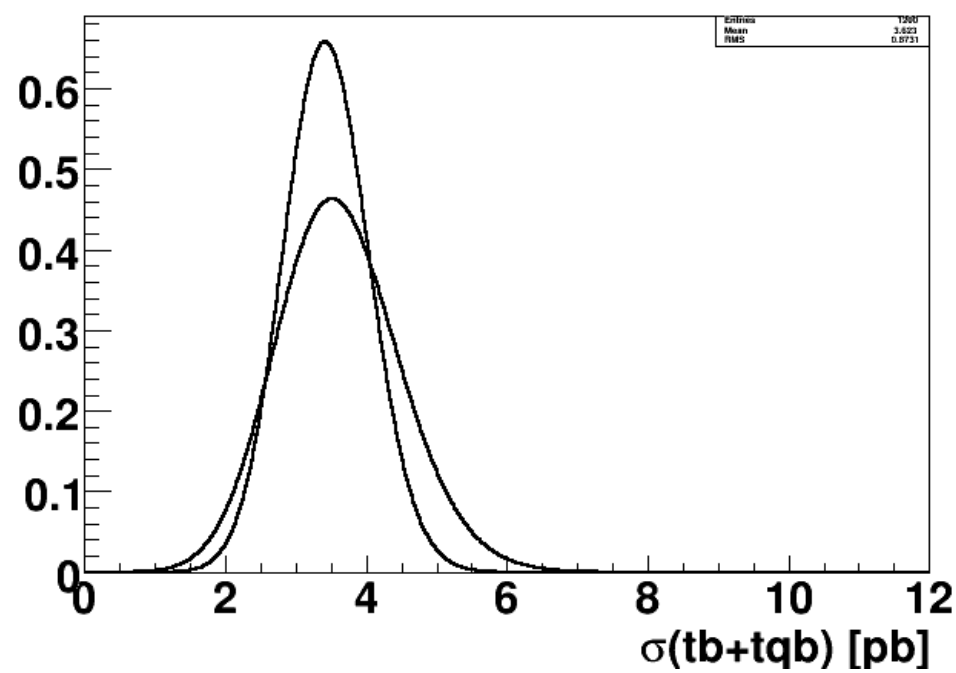
$$P(\text{signal}|\text{data}) = \iint_{\text{sys}} \frac{P(\text{data}|\text{signal},\text{bkg},\text{sys}) \times P(\text{signal}) \times P(\text{sys})}{P(\text{data})}$$

- P(sys) is a multivariate Gaussian prior
  - One Gaussian for each systematic
  - Other shapes also in use
    - Gamma function for MC statistics
    - Quadratic matching for asymmetric uncertainties
- Integrate over systematic uncertainties

Systematic uncertainty sampling

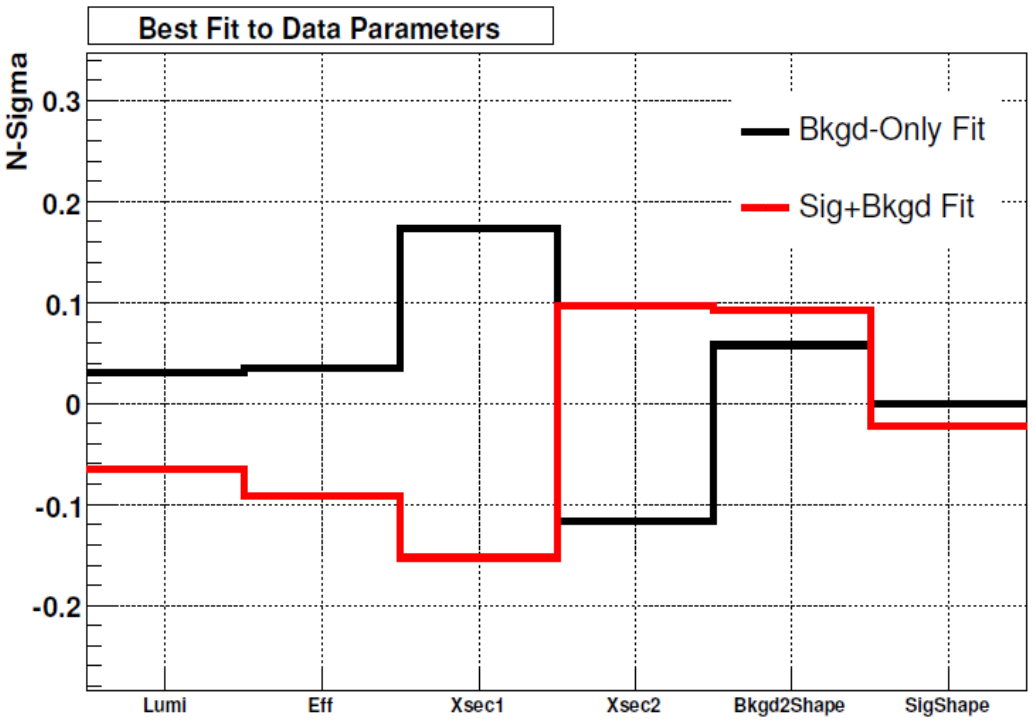
- Integration over systematics by MC sampling
  - Need sufficient # of samples from multivariate Gaussian
    - Typically ~1,000 to ~50,000 samples
  - Widens Bayesian posterior
    - And possibly shifts it
  - Widens Frequentist LLR ensemble distributions





Profile likelihood in modified frequentist analysis

- Limit setting procedure is unchanged, simply replace the test statistic
  - Generation of pseudo-experiments still samples from multivariate Gaussian
- Instead of integrating over each systematic, perform a Minuit fit to the data
  - With Gaussian prior as an extra constraint
  - Do this for the data and for pseudo-datasets
  - Separately for the null and alternative hypothesis



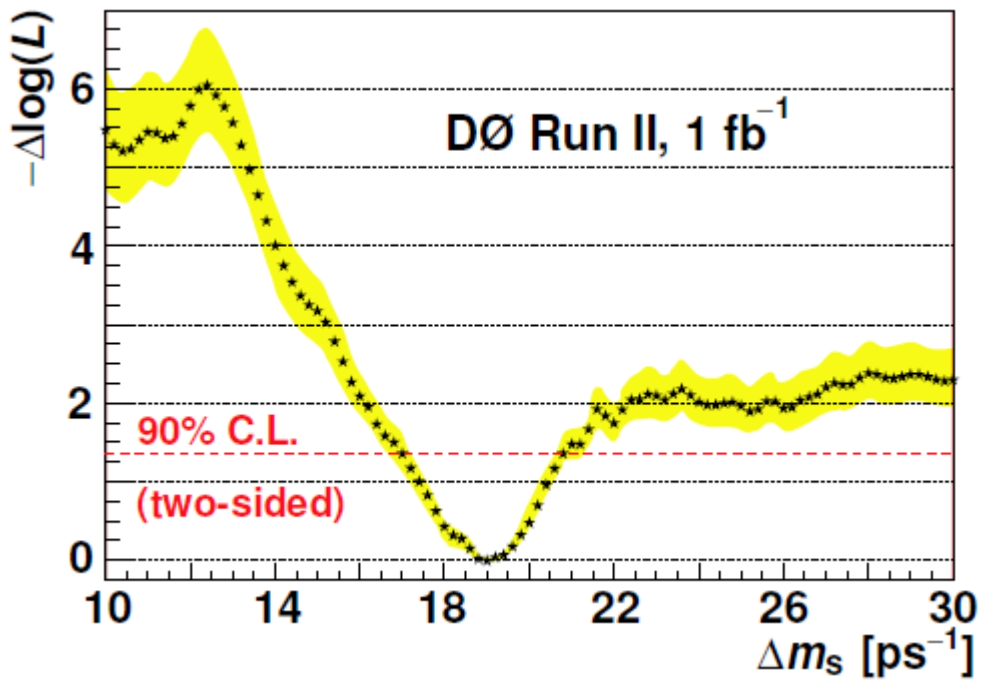
Quick hits

Multiple analysis channels

- Multiple analysis channels and bins are very straightforward:
  - As long as bins are statistically independent
- Multiply likelihoods of individual bins
  - Sum LLR values of individual bins
- Treat systematic uncertainty correlations bin-to-bin
  - Same σ shift for this systematic in all bins
  - Also include correlations between signal and background and between different background sources

Log likelihood

- Widely used in statistics and HEP
  - Originally based on Bayesian ideas
    - Keep data fixed, vary model parameter
- Commonly used in measurements
  - Minimum as measured value, min+1/2 for uncertainty
  - Also available for



Systematic uncertainties

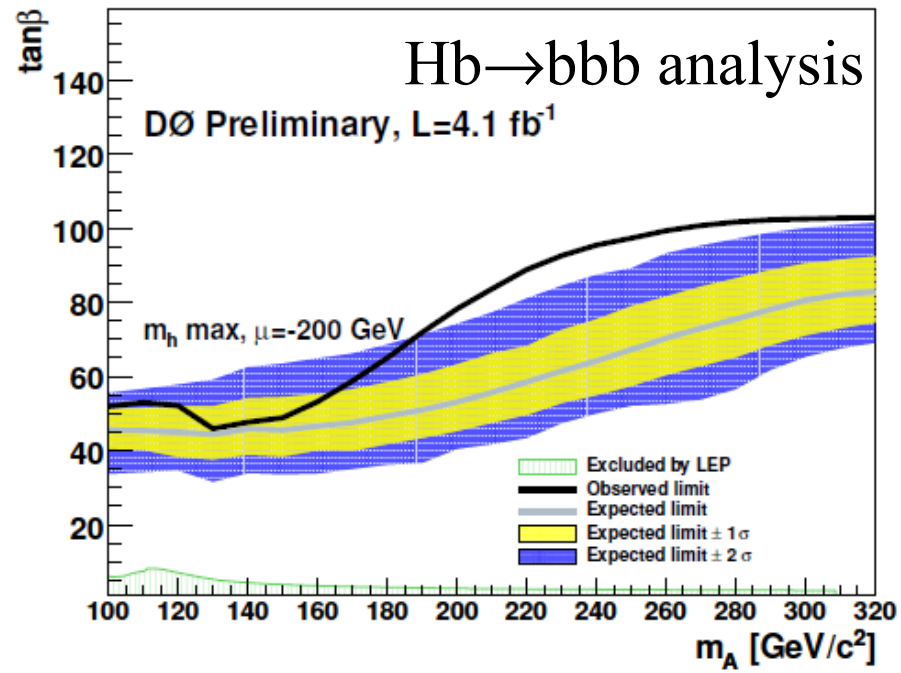
- Summing systematics in quadrature
  - Standard procedure for many measurements
  - Propagate each uncertainty to the final measurement
    - i.e. how does my measured parameter change if JES was shifted up 1σ?
  - Then add all uncertainties in quadrature
  - Assumptions
    - All systematics are independent, uncorrelated and small
    - Central measurement doesn't change when including systematics
- If any of these assumptions are not valid, need to evaluate systematics correctly
  - Integration over systematics

Pseudo-experiments

- Pseudo-experiments are very useful for a wide variety of statistical questions
- LLR calculation, significance calculation
  - Single top observation significance calculation based on 67,800,000 pseudo-datasets
- Other applications:
  - Test correlations between variables and/or channels
  - Test linearity
  - Test complete analysis chain

Why can't the Higgs mass be fixed and known?

– When exploring a distribution with an unknown parameter (such as the mass), statistical fluctuations may occur anywhere in the distribution
  - Look-elsewhere effect
  - The significance for finding a signal at one mass is reduced



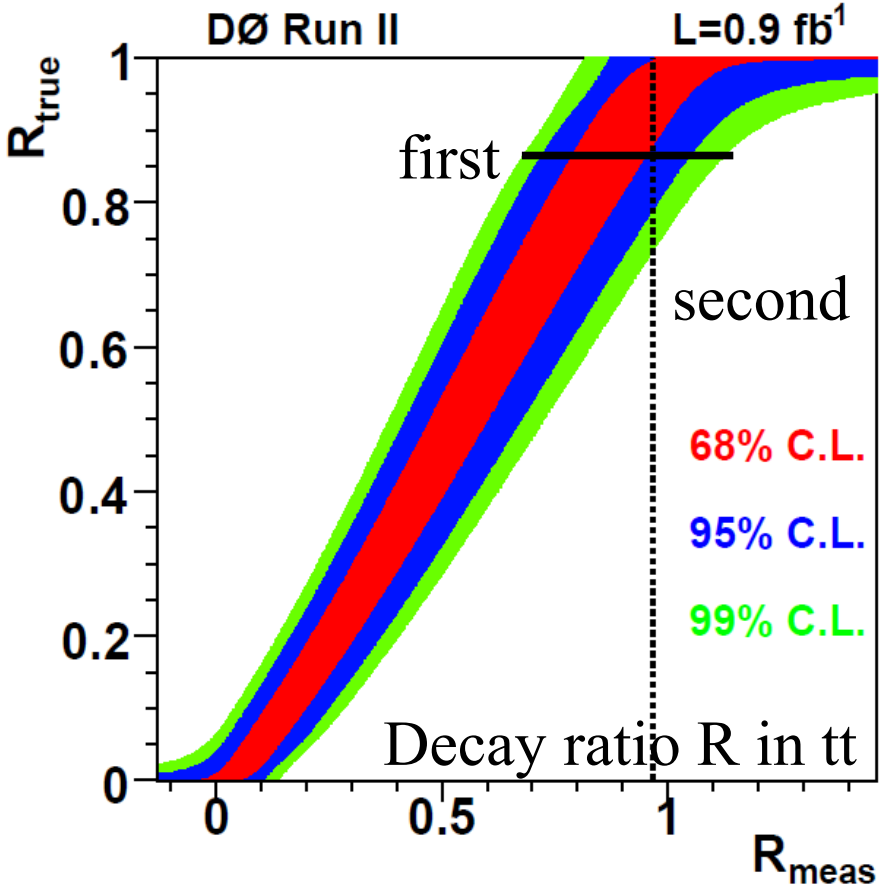
– Trials factor
  - Correct p-value for possibility of statistical fluctuations elsewhere

– Derive trials-factor in pseudo-experiments
  - What fraction of pseudo-experiments, when generated over the entire mass range, gives a signal as large as the one observed or larger?

– Or through various approximations
  - Eg: Divide mass region into bins roughly the width of the mass resolution, then correct for varying acceptance

Feldman Cousins

- Start with known true value and determine intervals
  - Then read off data interval
  - Avoids flip-flopping and empty intervals
- Procedure
  - For several values of the true, known parameter, determine the 95% CL interval i.e. along the horizontal axis
  - Plot all of them in a 2d plot
  - Take the data measurement and read the 95% CL interval vertically



Conclusions

- Statistical analysis practices in use today have been around for a long time
  - Textbooks are from the 1930s
  - Bayes lived in the 18th century
  - Important to understand the fundamentals
- HEP has expanded on standard statistical procedures and invented a few new ones
  - To deal with low statistics and systematic errors
  - Tools are easily available
- DØ is using all of the current tools
  - Employs Frequentist and Bayesian and mixed medhods

Disclaimers

- Statistics doesn't have to be hard!
- There are many ways to introduce statistical data analysis, this is just my view of it.
- This is neither complete nor unique.
- To paraphrase NPR: Sticklers for the truth should get their own UDØ talk.
- This talk is about current practices, there are other equally valid ways to do it.
- This talk is not about what you should not do. There are many ways to go wrong.
- There are many other important statistics topics that didn't get discussed here.
- A lot of basic statistics information is available on wikipedia and other online sources.
- Almost every statistics problem can now be addressed in root.
- Roostat is the tool currently in development for ATLAS and CMS that does all of this easily.
- There are regular workshops and lectures on statistics with much valuable information.
- Statisticians use the tool and language R. This has many statistical analysis tools including limit setting and multivariate analysis methods. http://www.r-project.org/

"How do you want it—the crystal mumbo-jumbo or statistical probability?"

Book recommendations

- Louis Lyons – easy to read, not very theoretical
  Statistics for Nuclear and Particle Physicists
  http://www.amazon.com/Statistics-Nuclear-Particle-Physicists-Louis/dp/0521379342
- Glen Cowan – more details, well written
  Statistical Data Analysis
  http://www.amazon.com/Statistical-Analysis-Oxford-Science-Publications/dp/0198501552
- Fred James – more detailed, useful reference
  Statistical Methods in Experimental Physics
  (2nd Edition)
  http://www.amazon.com/Statistical-Methods-Experimental-Physics-2nd/dp/981256795X

Links

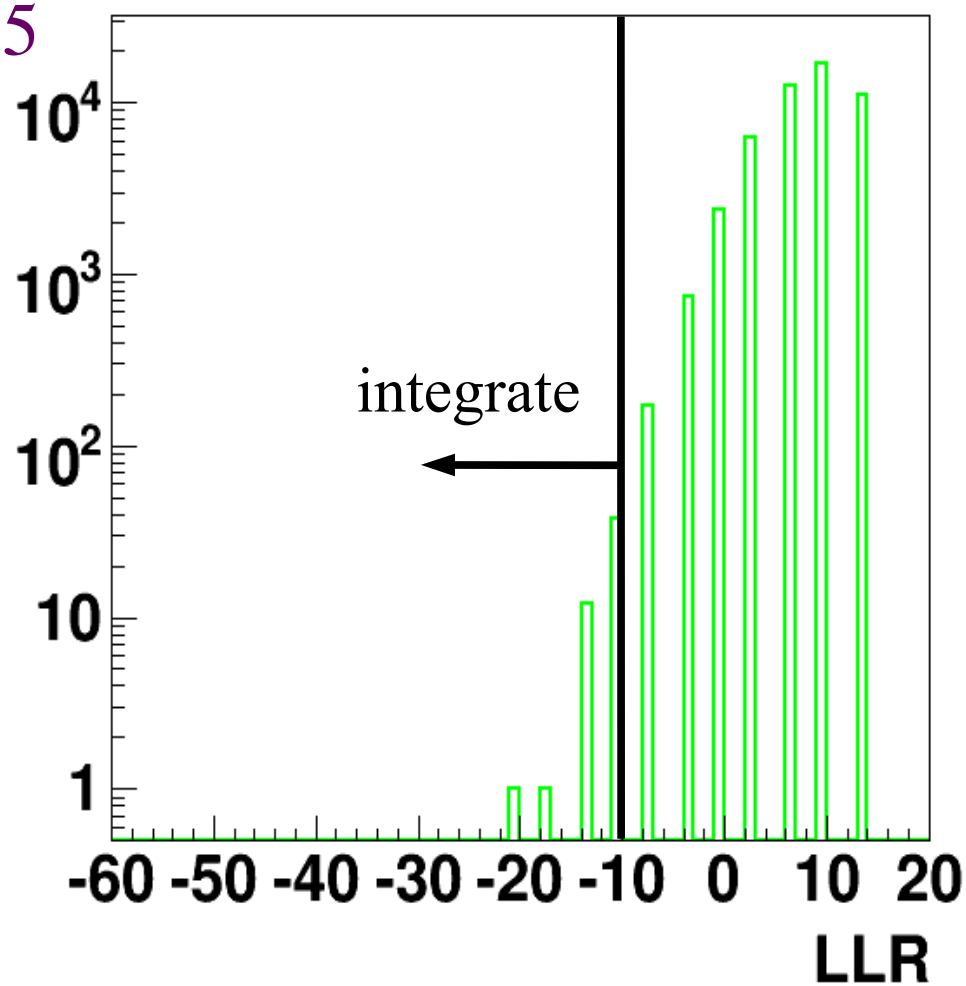
- Jim Linnemann's statistics web page
  http://www.pa.msu.edu/people/linnemann/stat_resources.html
- phystat.org
  http://www.phystat.org/
- Roostats is the ROOT statistical analysis package, it is in development
  https://twiki.cern.ch/twiki/bin/view/RooStats
- Wikipedia is excellent. Start with
  http://en.wikipedia.org/wiki/Bayesian_inference
  http://en.wikipedia.org/wiki/Frequentist_inference
  and explore the links starting from there
- Glen Cowan's lecture notes and other resources
  http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Answers to some questions

- In the Bayesian analysis, there are several different possible priors. Why not assign a systematic uncertainty to the prior?
  - Answer: Because the prior is a likelihood. Saying the prior should have a systematic uncertainty would simply result in a different prior.

Answers to some questions

- In the Modified Frequentist approach, the limits get better when profiling. But in the Tevatron Higgs combination, even after profiling, the limits are similar to those from the Bayesian analysis. Why are they not better?
  - Modified Frequentist and profiling together give roughly the same numerical limits as the Bayesian method
    - But the two numbers have different meanings of course
  - Modified Frequentist without profiling over-covers conservatively (the coverage is typically around 97.5%)

Example: discovery from ensembles

- Double-charged Higgs analysis. What if we had observed 7 events instead of 3?
  - Background expectation 1.5
  - 7 observed events corresponds to LLR=-10
  - Integrate from -∞ to -10: Observed p-value: 0.001
  - 3 sigma evidence