Statistics Overview

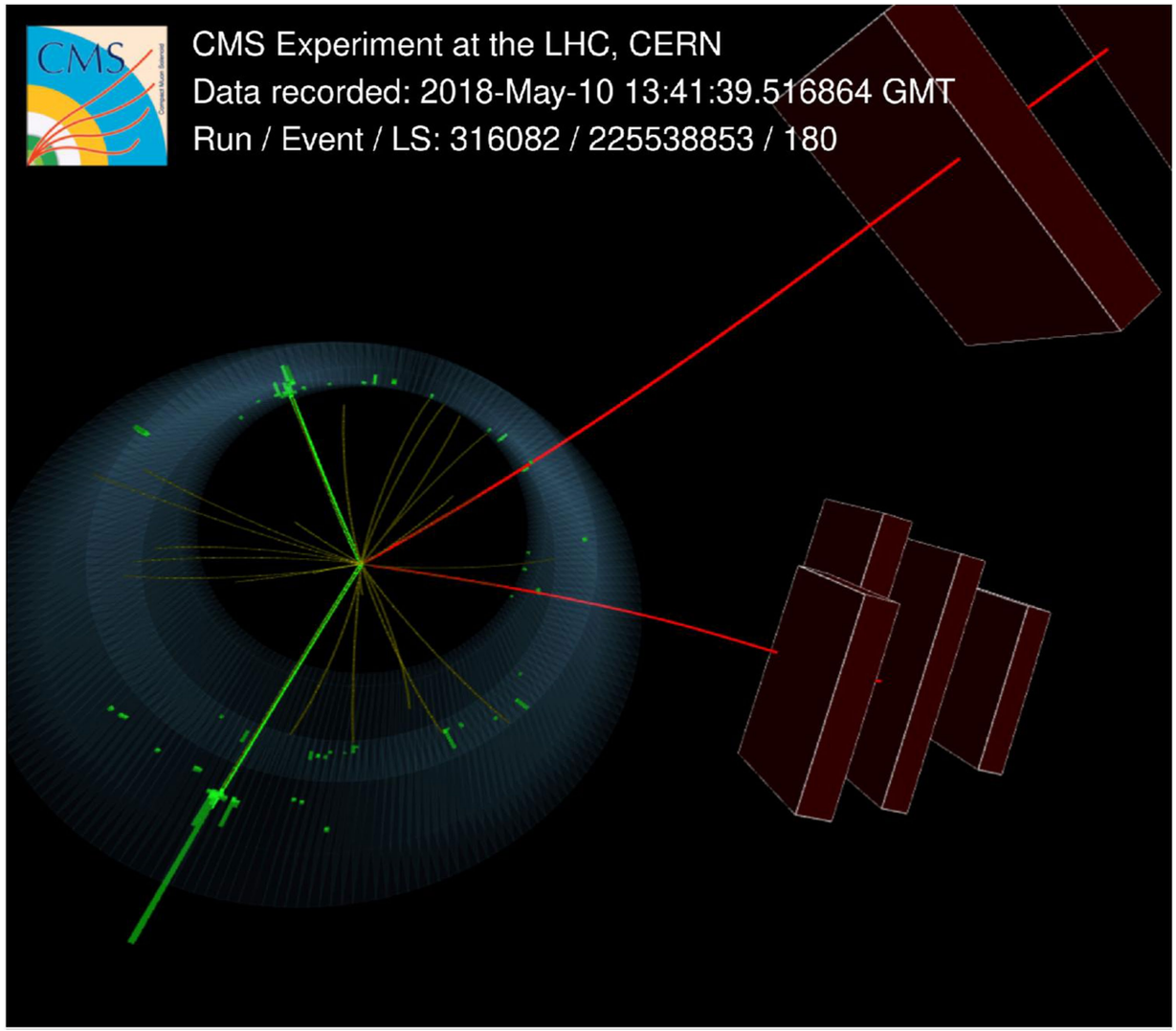
Susan Dittmer

CMS Topical Workshop on Off-shell Higgs Boson Production at LPC

March 25, 2024

Statistical Analysis in HEP

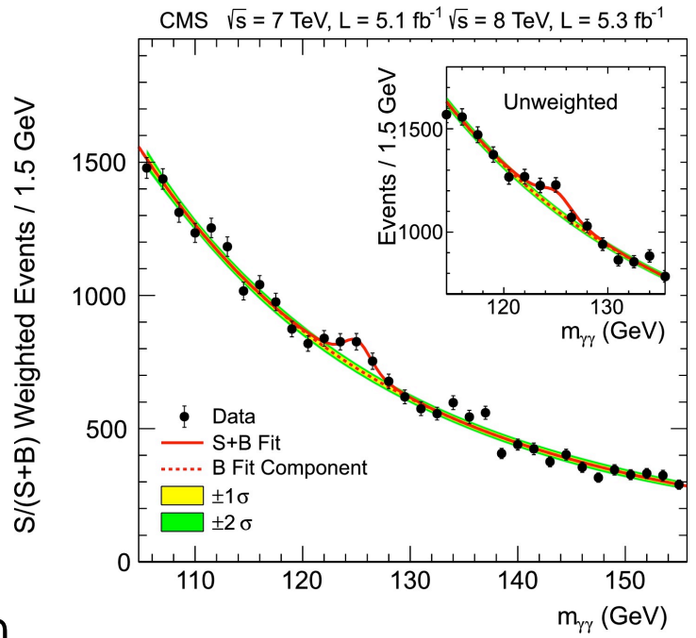
- Particle physics is probabilistic
  - Parton distribution functions, branching ratios, momentum and energy resolution, etc…
  - No signatures impossible under SM — just (very) improbable
- Need statistical analysis + many measurements to draw conclusions on underlying model

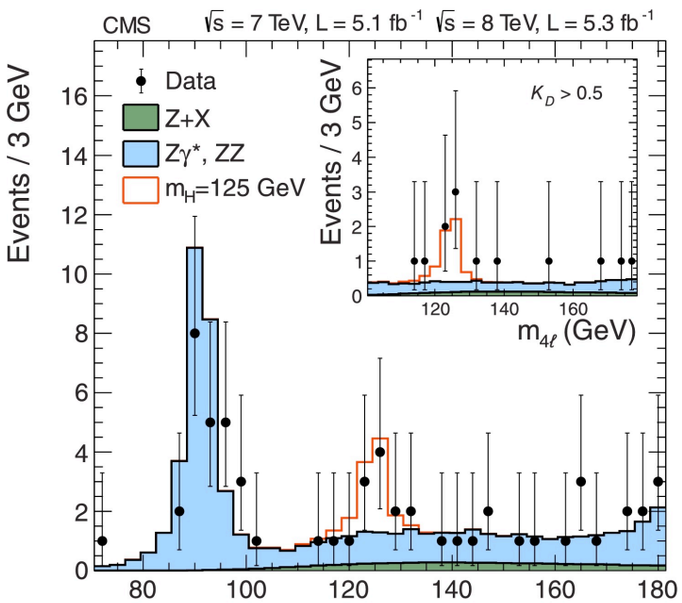


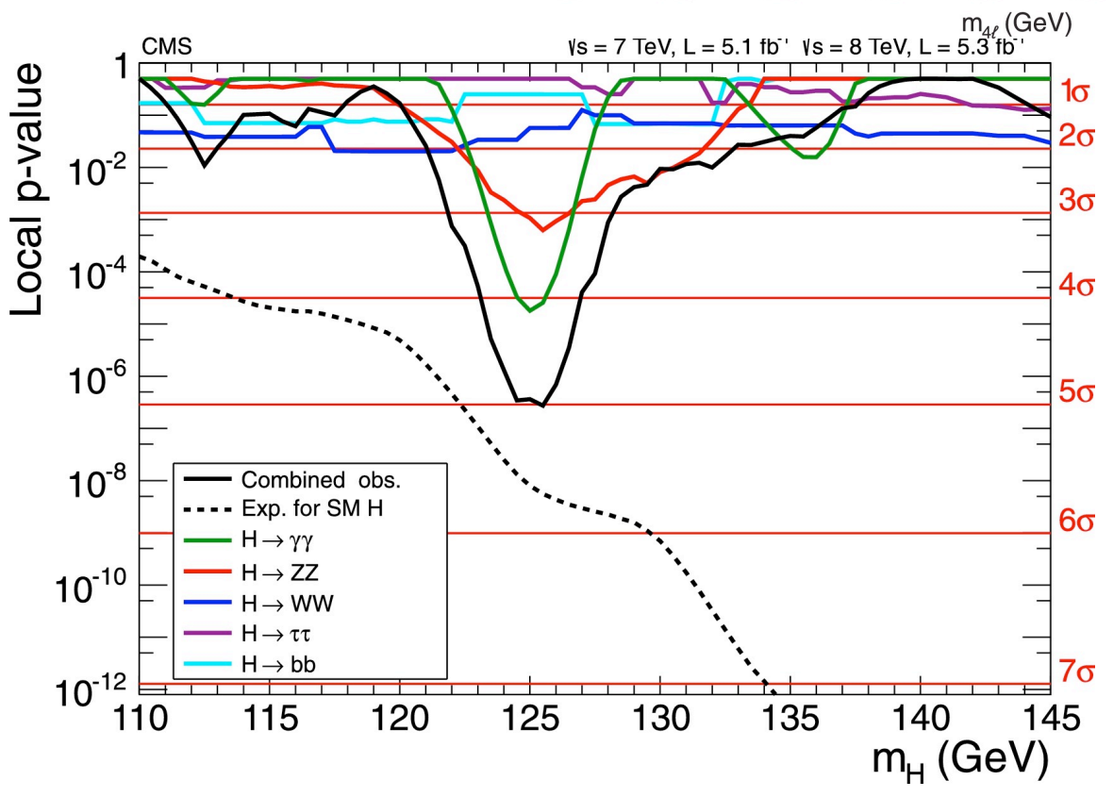
Single event could be Higgs … or could be nonresonant ZZ

Statistical Analysis in HEP

- Particle physics is probabilistic
  - Parton distribution functions, branching ratios, momentum and energy resolution, etc…
    - No signatures impossible under SM — just (very) improbable
  - Need statistical analysis + many measurements to draw conclusions on underlying model







Probability and Likelihood

- P(x|H) : probability to observe x given hypothesis H
  - If x is discrete: standard probability
  - If x is continuous: probability density function p(x|H), such that integrating p(x|H) over x sums to 1
  - Usually we want to invert this: given observation x, how likely is a specific hypothesis?
    - Likelihood L(H) = P(x|H)
    - Not a probability (density function) — no guarantee that integral over hypotheses is 1
- Assuming the hypothesis is characterized by parameters θ, can write instead L(θ) = P(x|θ)

Bayesian vs Frequentist Statistics

- Bayesian: probability of model (quantifies belief)
  - ‘I am 50% confident that Schrödinger’s cat is alive’
  - How to get P(θ|x)?
    - Bayes’ theorem: P(A|B) = P(B|A)*P(A)/P(B)
    - $P(\theta|x) = \frac{P(x|\theta) * P(\theta)}{P(x)} = \frac{P(x|\theta) * P(\theta)}{\sum_{\theta_i} P(x|\theta_i)P(\theta_i)}$
    - Relies on prior P(θ) — make informed assumptions or take from other measurements
- Frequentist: probability of (repeated identical) observation, given model
  - ‘If I repeat this experiment 100 times, Schrödinger’s cat will be alive 50 times’
  - Relies solely on P(x|θ) —> no need to know P(θ)
  - However need to perform (or simulate) large set of measurements to determine result

Estimating Model Parameters

- When performing SM measurements, usually want to know value and uncertainty of interesting parameters
  - Cross section, branching ratio, particle mass and width, etc.
- Statistical treatment: measured parameter = parameter estimator
  - Estimator: function $\hat{\theta}(x)$ of data which estimates true θ
  - Good estimators are:
    - Consistent: $\hat{\theta}(x) \to \theta$ as size of $x \to \infty$
    - Unbiased: expectation $E[\hat{\theta}(x)] = \theta$ over multiple similar measurements
    - Efficient: lowest possible variance
      - Specifically, variance is at Rao-Cramér-Fréchet bound —> won't elaborate here
    - Robust: insensitive to small changes in pdf

Maximum Likelihood Estimator

- Maximum likelihood estimator (MLE): $\hat{\theta}(x)$ which maximizes likelihood L(θ) = P(x|θ)
  - 'Best fit' values of model parameters are those which yield most probable model for given data
- In practice, instead of maximizing likelihood we minimize negative log likelihood (NLL)
  - Product of likelihoods becomes sum of logs —> easier computation
  - Find θ_i for which $\frac{\partial \ln L}{\partial \theta_i} = 0$
- MLE is efficient and asymptotically unbiased (i.e. for large datasets)

Parameters of Interest

- Full model usually has many parameters — only interested in a few
- Split parameters θ into parameters of interest and nuisance parameters
- Parameters of interest (μ)
    - Physics definition: quantities to be measured (cross section, mass, etc.)
    - Statistics definition: free parameters of likelihood
- Nuisance parameters (ν)
    - Physics definition: parameters modeling systematic uncertainties
    - Statistics definition: parameters to be reduced from likelihood, usually constrained by auxiliary measurements

Building Likelihood

- Single measurement: $L(\mu, \nu) = p(x|\mu, \nu)$
- Multiple independent measurements: $L(\mu, \nu) = \prod_{i=1}^{n} p(x_i|\mu, \nu)$
- Multiple independent measurements, where number of measurements also depends on parameters
  - $L(\mu, \nu) = \frac{n_{exp}(\mu, \nu)^n}{n!} e^{-n_{exp}(\mu,\nu)} \prod_{i=1}^{n} p(x_i|\mu, \nu)$
    - Extended likelihood
  - Addition of auxiliary measurements
    - $L(\mu, \nu) = \prod_{i=1}^{n} p(x_i|\mu, \nu) \prod_{i=1}^{m} p(y_i|\nu_i)$
    - Independent measurements of certain model parameters, e.g. constraints on systematic uncertainties

Parameter Uncertainties

MLE What determines uncertainty?

- ‘Top mass is measured to be 172.5 ± 0.3 GeV’
- Interval covering true value of parameter with some probability
  - For this type of result, usually 68%, or 1σ (under Gaussian assumption)
- Bayesian: credibility region
  - ‘True value of top mass is 68% likely to fall within 172.2 < mtop < 172.8’
- Frequentist: confidence interval
  - ‘If we repeat the experiment multiple times, 68% of intervals will contain true top mass’

Parameter Uncertainties

MLE What determines uncertainty?

- ‘Top mass is measured to be 172.5 ± 0.3 GeV’
- Interval covering true value of parameter with some probability
  - For this type of result, usually 68%, or 1σ (under Gaussian assumption)
- Bayesian: credibility region
  - ‘True value of top mass is 68% likely to fall within 172.2 < mtop < 172.8’
- Frequentist: confidence interval
  - ‘If we repeat the experiment multiple times, 68% of intervals will contain true top mass’

We usually report confidence intervals

Bayesian Credibility Region

- Bayesian credibility region for probability 1-**α** is [μ_{lo},μ_{hi}] such that

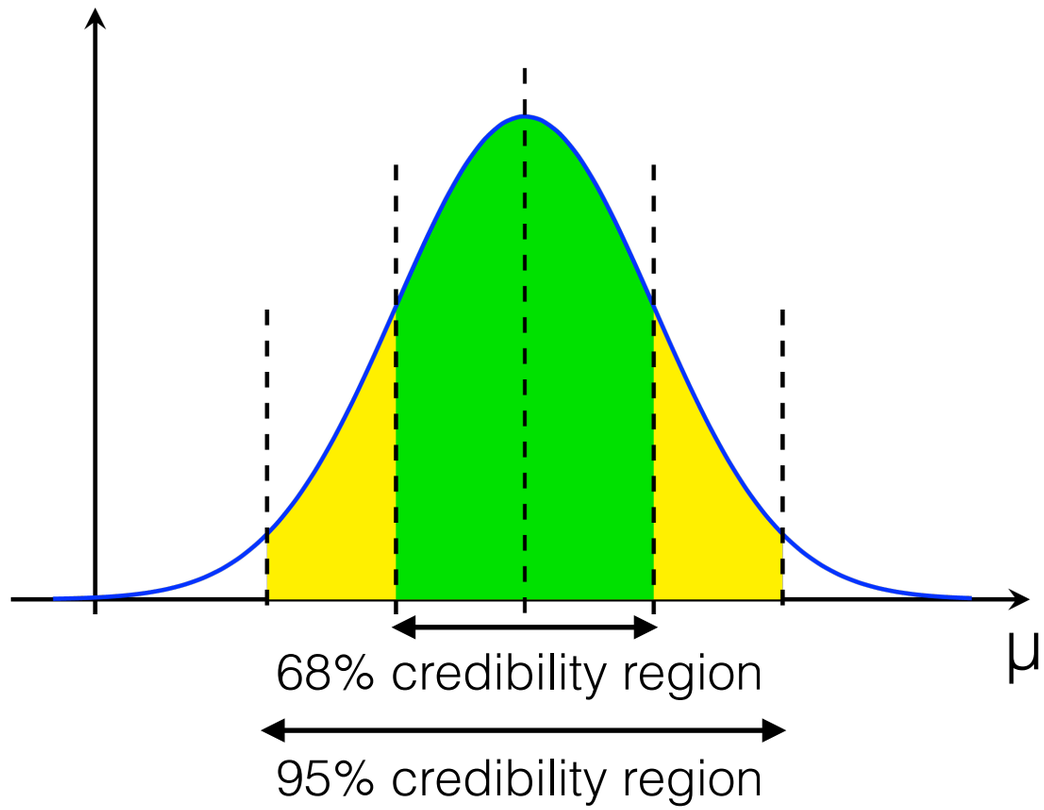
$$1-\alpha = \int_{\mu_{lo}}^{\mu_{hi}} p(\mu|x)d\mu$$

  - N.B. p(μ|x) is the posterior probability,

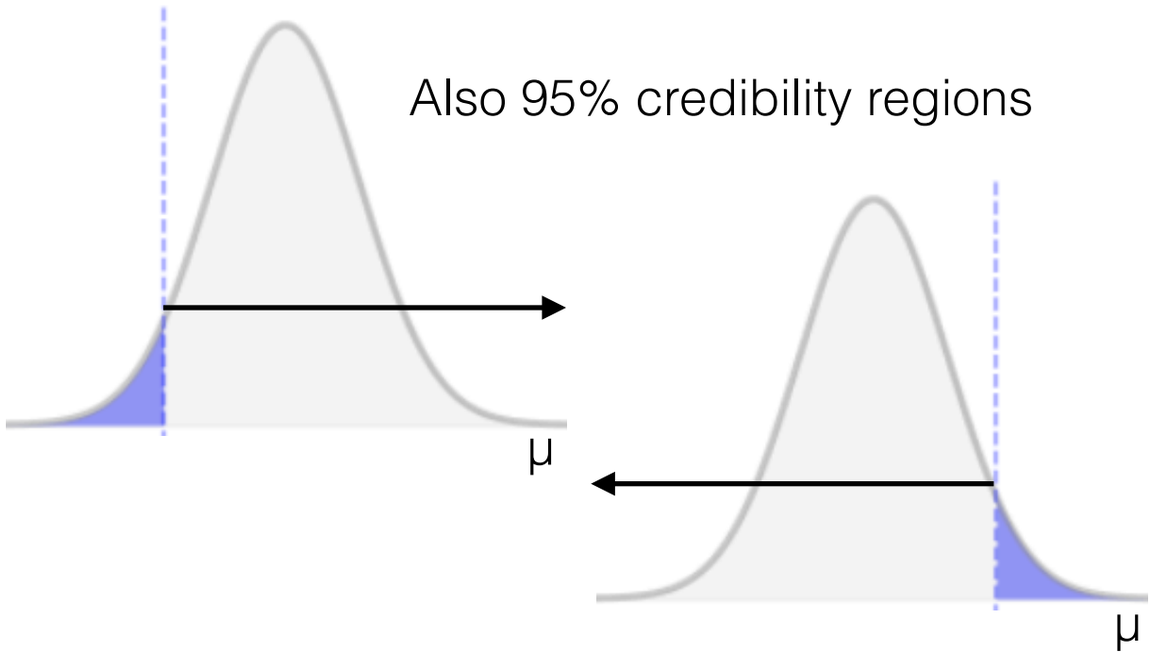
$$p(\mu|x) = \frac{p(x|\mu) * p(\mu)}{\int_{\mu'} p(x|\mu')p(\mu')d\mu'}$$

- Ranges are not unique

- Most common choices: central, one-sided



Also 95% credibility regions

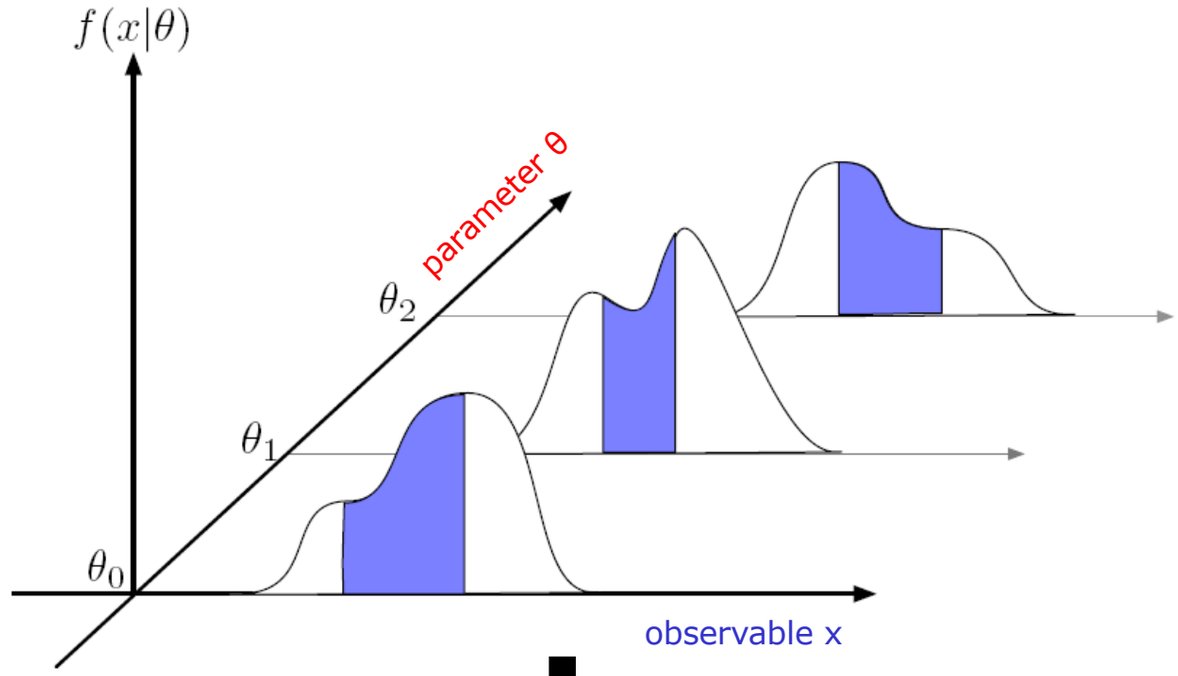


Frequentist Confidence Interval

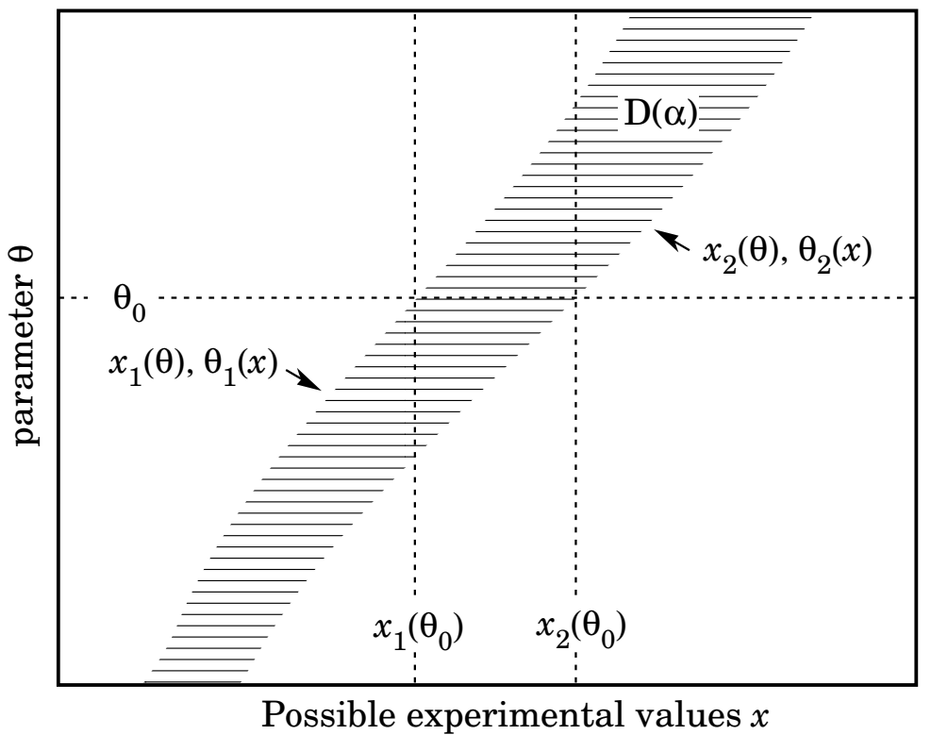
- Confidence interval determined following Neyman construction
- Similarly to Bayesian, interval has given probability (ex. 68%, 95%) and type (central, one-sided)
- For each value of μ, can determine range [x_1,x_2] such that

$$\int_{x_1}^{x_2} p(x|\mu)dx \geq 1-\alpha$$

- where $1-\alpha$ is the probability, and [x_1,x_2] are chosen based on the interval type

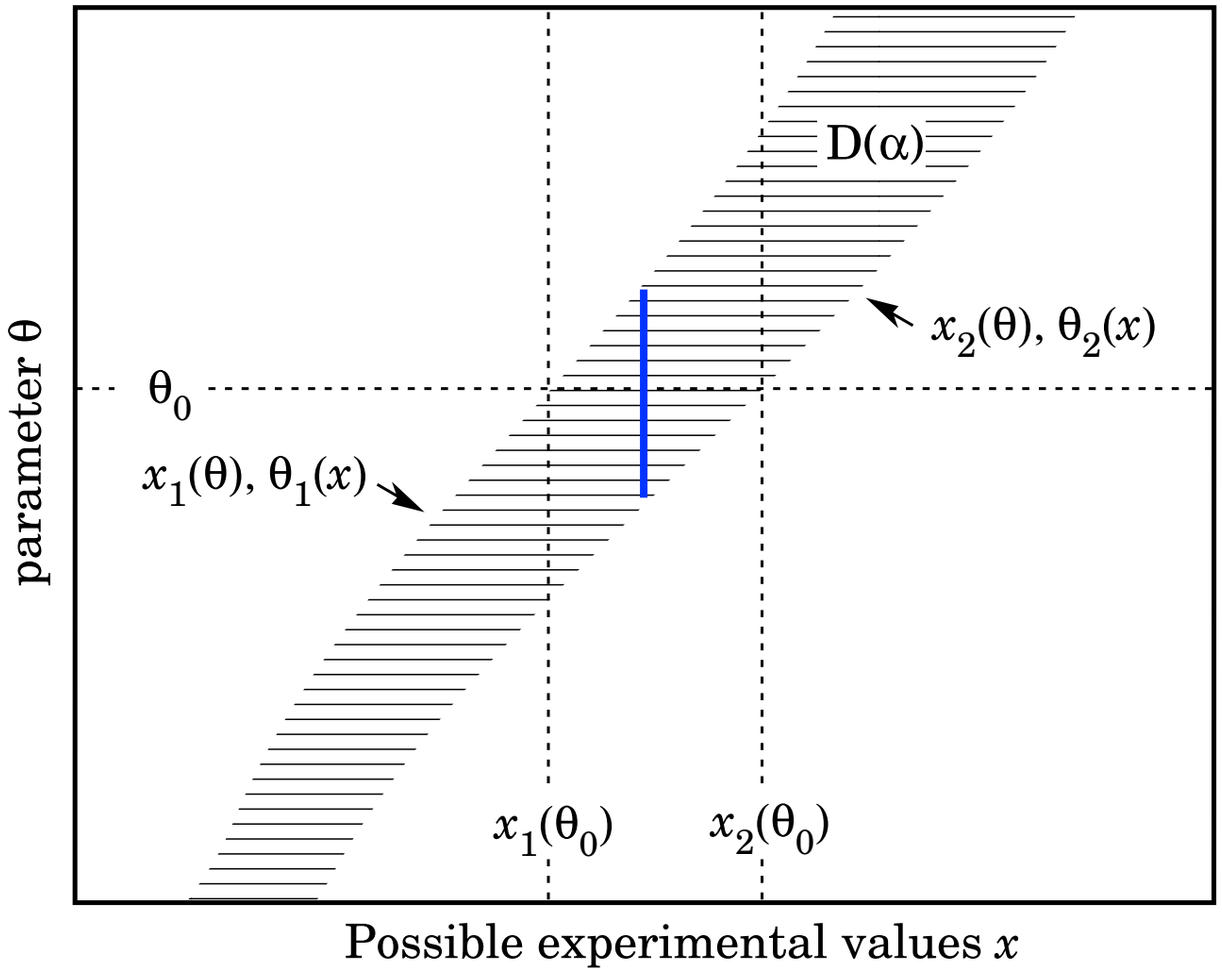


Wouter Verkerke, NIKHEF



Frequentist Confidence Interval (cont.)

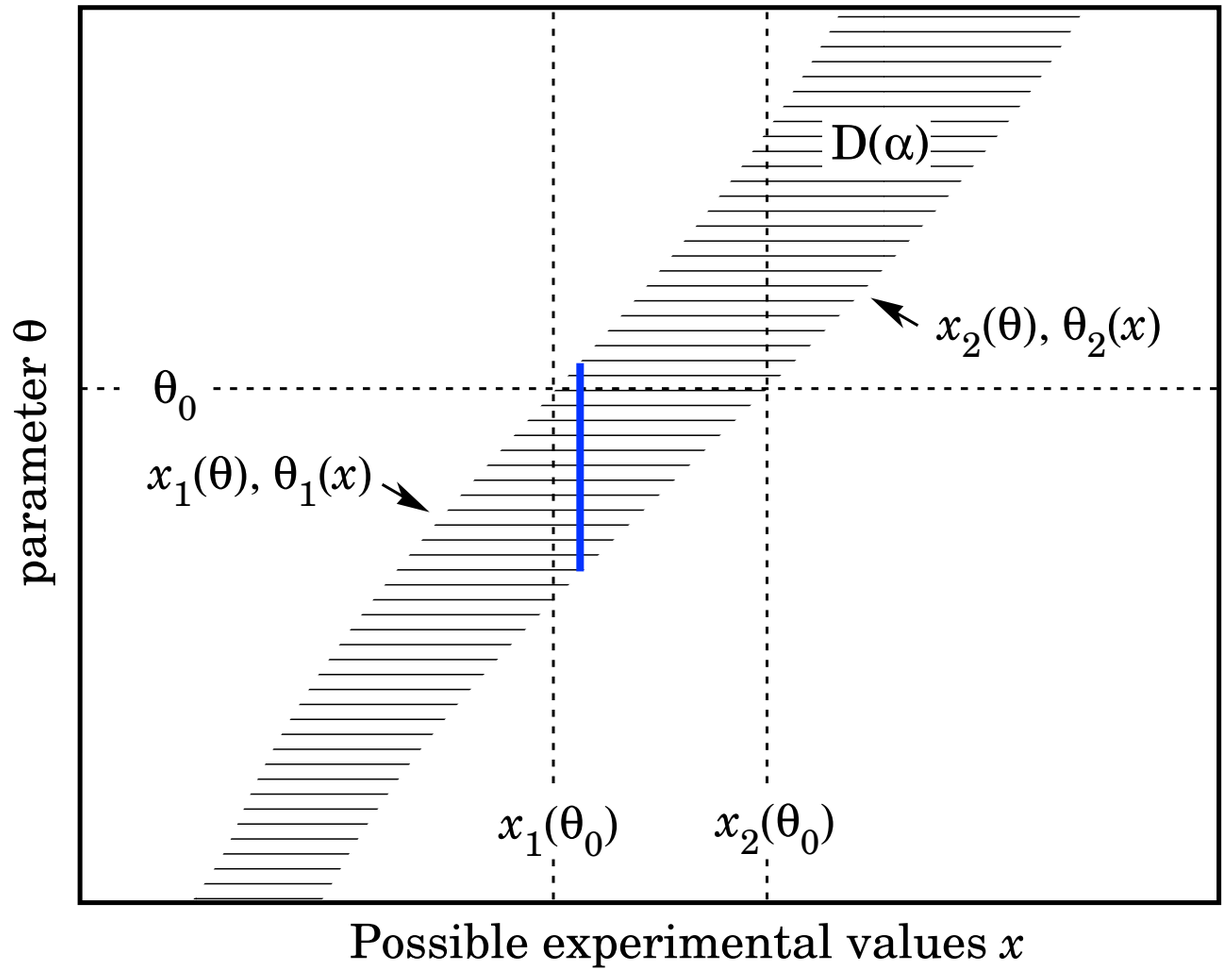
- Define parameter range $[\mu_{lo}(x), \mu_{hi}(x)]$ as a function of measurement x
- Say true value of parameter is μ
  - $\mu \in [\mu_{lo}(x), \mu_{hi}(x)]$ iff $x \in [x_1(\mu), x_2(\mu)]$
- If experiment is repeated many times, probability that true value of parameter μ falls in confidence interval is



$$P(\mu_{lo}(x) < \mu < \mu_{hi}(x)) = P(x_1(\mu) < x < x_2(\mu))$$
$$= 1 - \alpha$$

Frequentist Confidence Interval (cont.)

- Define parameter range $[\mu_{lo}(x), \mu_{hi}(x)]$ as a function of measurement x
- Say true value of parameter is μ
  - $\mu \in [\mu_{lo}(x), \mu_{hi}(x)]$ iff $x \in [x_1(\mu), x_2(\mu)]$
- If experiment is repeated many times, probability that true value of parameter μ falls in confidence interval is



$$P(\mu_{lo}(x) < \mu < \mu_{hi}(x)) = P(x_1(\mu) < x < x_2(\mu))$$
$$= 1 - \alpha$$

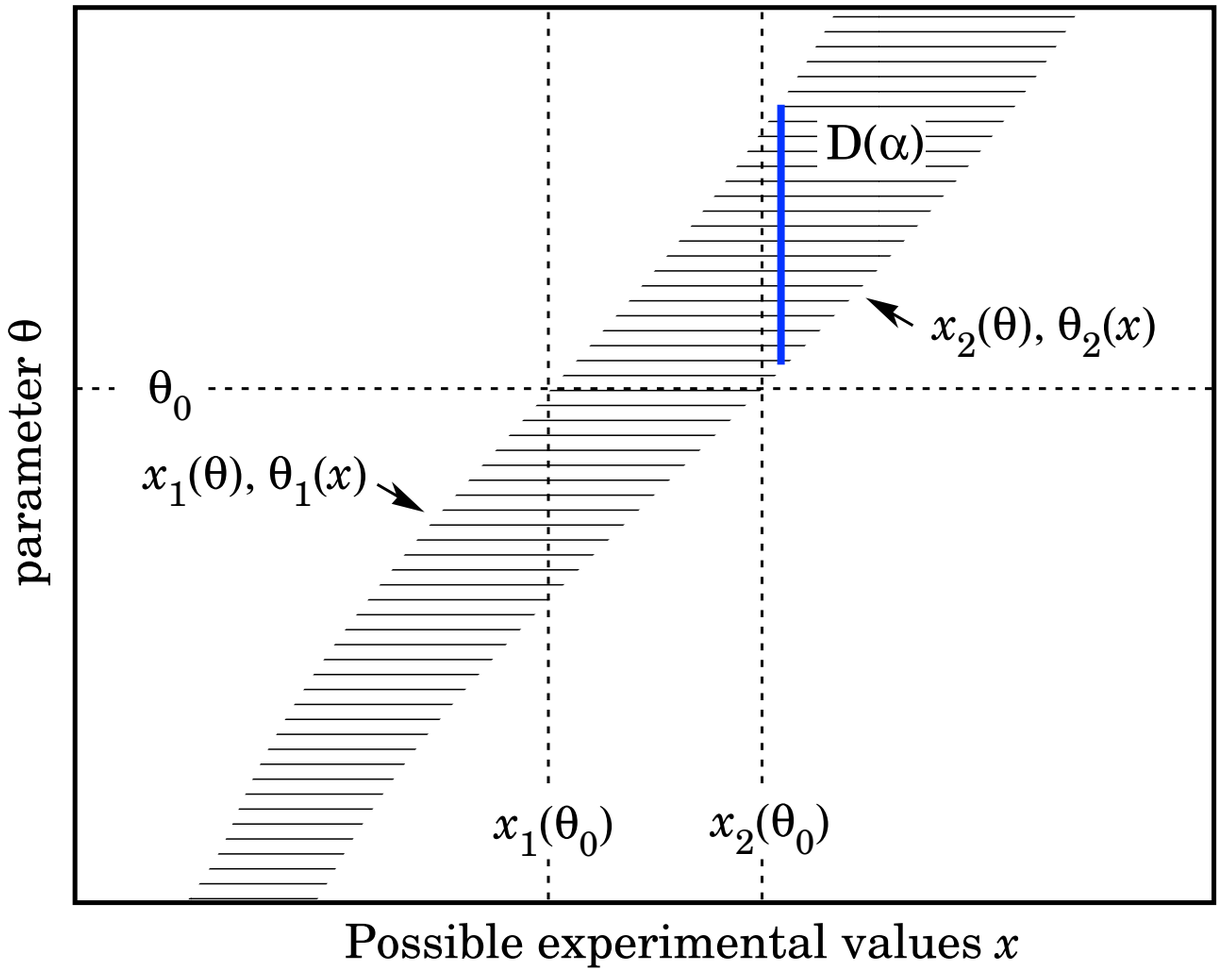
Frequentist Confidence Interval (cont.)

- Define parameter range $[\mu_{lo}(x), \mu_{hi}(x)]$ as a function of measurement x

- Say true value of parameter is μ

  - $\mu \in [\mu_{lo}(x), \mu_{hi}(x)]$ iff $x \in [x_1(\mu), x_2(\mu)]$

- If experiment is repeated many times, probability that true value of parameter μ falls in confidence interval is



$$P(\mu_{lo}(x) < \mu < \mu_{hi}(x)) = P(x_1(\mu) < x < x_2(\mu))$$
$$= 1 - \alpha$$

Reduced Likelihoods

- Previous results assume likelihood depends solely on parameter of interest
- Realistically, likelihood depends on other POIs, nuisance parameters —> how to account for effects on confidence interval?
- Solution: reduce likelihood
  - Conditional likelihood: fix values of other parameters
  - Profiled likelihood: set nuisance parameters to ‘best fit value’, i.e. values which maximize likelihood for given value of POI
  - Marginalized likelihood: integrate likelihood over nuisance parameters

Confidence Intervals from Profile Likelihood Ratio

- Actual approach used by most fitting algorithms is actually a special case: construct intervals in profile likelihood ratio rather than observable x

- Profile likelihood ratio: profiled likelihood divided by max likelihood

$$\lambda_p(\mu) = \frac{L(\mu, \hat{\hat{\nu}}(\mu))}{L(\hat{\mu}, \hat{\nu})}$$

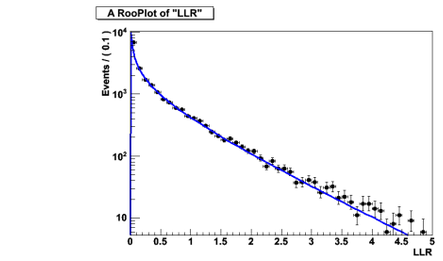
  - where $\hat{\mu}$ and $\hat{\nu}$ are the parameters which maximize the likelihood, and $\hat{\hat{\nu}}(\mu)$ are the nuisance parameters which maximize the likelihood for a given μ

- According to Wilks' theorem, $\lambda_p(\mu)$ is asymptotically normal

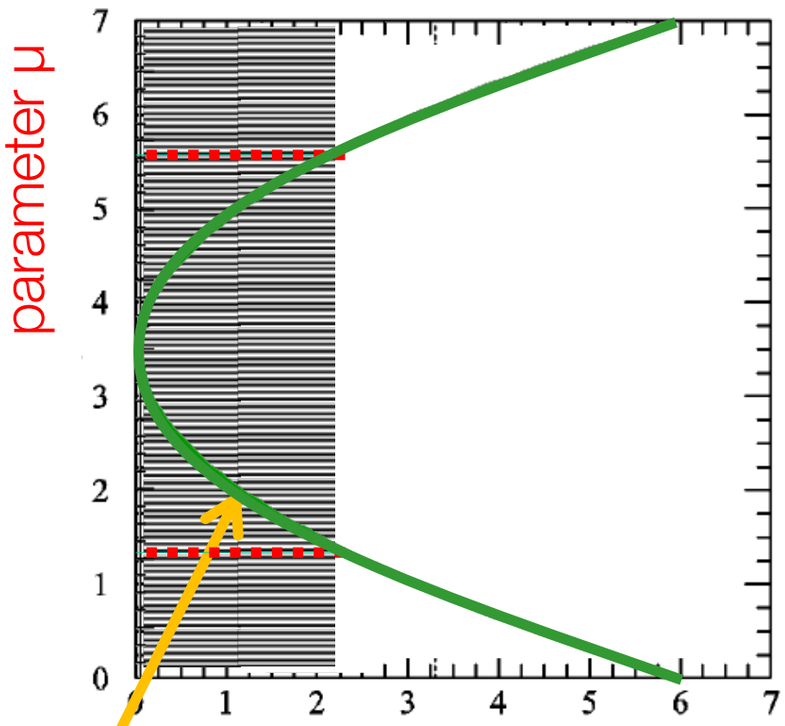
  - $-2ln(\lambda_p(\mu))$ approaches chi-square for sufficient stats

  - Independent of μ!

- Means we can compute single 'range' for given coverage, and determine confidence intervals from where likelihood ratio for each μ crosses range



$t_\mu(x,\mu)$



Minos / Hessian Calculations

- Parameter uncertainty estimates in RooFit / RooStats use one of two methods: Minos and Hessian.
  - Both use profile likelihood test statistic to define interval
- Minos method
  - Moving away from best fit value $\hat{\mu}_i$, calculate $-2ln(\lambda_p)$ for each value of μ_i (profiling other parameters)
    - Find two crossing points where $-2ln(\lambda_p)$
    - Repeat for remaining μ_i
- Hessian method
  - Calculate second derivatives of likelihood at best fit point
  - Use these to analytically calculate crossing points: $\mu_i^{\text{crossing}} - \hat{\mu} \propto \left(\frac{\partial^2 \mathcal{L}(\vec{\hat{\mu}})}{\partial \mu_i^2} \right)^{-2}$
  - Uncertainties are symmetric by construction

For well defined likelihood, these should agree

Hypothesis Testing

- Outside of parameter measurements, physics results are statements about compatibility of data with model
  - ‘No significant deviation was observed from the SM prediction’
  - ‘A new boson with mass 125 GeV was observed with >5σ significance’
- Base method is hypothesis test
  - Compare model hypothesis (H1) vs null hypothesis (H0)
    - Null hypothesis is generally full signal model minus process of interest
      - SM for BSM searches
- Outcomes: reject H0 in favor of H1, or accept H0
- What can go wrong?
  - Type I error: reject the null hypothesis when it is true
  - Type II error: accept null hypothesis when it is false

Hypothesis Testing (cont.)

- In HEP, we minimize type I errors (false positive)
  - False observation of NP —> wasted personpower and loss of public confidence
  - False rejection of NP —> Nobel Prize delayed
- Generally, require probability to falsely reject null hypothesis to be below some threshold **α**
  - Then minimize probability β to falsely reject signal hypothesis (conversely, maximize power = 1-β)

P(x,H1) / P(x,H0) > k_α

P(xIH0) < α

- Null hypothesis rejection in frequentist approach:
  - Define region of phase space w such that p(x∈w|H0) < **α**
  - If data is observed in w, reject null hypothesis
- Multiple regions w will exist — choose region with maximum power
  - Minimize β = P(x∉w,H1)
  - By Neyman-Pearson lemma, this is a fixed contour in the likelihood ratio, e.g x where P(x,H1) / P(x,H0) > k_α

$$P(x,H0) = \alpha$$

$$P(x,H1)/P(x,H0) > k_\alpha$$

Hypothesis Testing (cont.)

- In HEP, we minimize type I errors (false positive)
  - False observation of NP —> wasted personpower and loss of public confidence
  - False rejection of NP —> Nobel Prize delayed
- Generally, require probability to falsely reject null hypothesis to be below some threshold **α**
  - Then minimize probability β to falsely reject signal hypothesis (conversely, maximize power = 1-β)

P(x|H0) < α

- Null hypothesis rejection in frequentist approach:
  - Define region of phase space w such that p(x∈w|H0) < **α**
  - If data is observed in w, reject null hypothesis
- Multiple regions w will exist — choose region with maximum power
  - Minimize β = P(x∉w,H1)
  - By Neyman-Pearson lemma, this is a fixed contour in the likelihood ratio, e.g x where P(x,H1) / P(x,H0) > k_α

$$P(x,H0) = P(x,H0) = \alpha$$

$$P(x,H1)/P(x,H0) > k_\alpha$$

$$P(x,H1)/P(x,H0) < k_\alpha$$

Hypothesis Testing (cont.)

- In HEP, we minimize type I errors (false positive)
  - False observation of NP —> wasted personpower and loss of public confidence
  - False rejection of NP —> Nobel Prize delayed
- Generally, require probability to falsely reject null hypothesis to be below some threshold **α**
  - Then minimize probability β to falsely reject signal hypothesis (conversely, maximize power = 1-β)

- Null hypothesis rejection in frequentist approach:
  - Define region of phase space w such that $p(x \in w|H0) < \alpha$
  - If data is observed in w, reject null hypothesis
- Multiple regions w will exist — choose region with maximum power
  - Minimize $\beta = P(x \notin w, H1)$
  - By Neyman-Pearson lemma, this is a fixed contour in the likelihood ratio, e.g x where $P(x,H1) / P(x,H0) > k_\alpha$

$$P(x,H0) = P(x,H0) = P(x,H0) = \alpha$$

$$P(x,H1)/P(x,H0) > k_\alpha$$
$$P(x,H1)/P(x,H0) > k_\alpha$$

$$P(x,H1)/P(x,H0) < k_\alpha$$

$$P(x,H1) > P(x,H1)$$

p-value and Significance

- We may want to test an individual hypothesis, without reference to null hypothesis
- p-value
  - Define test statistic t which quantifies agreement between data and hypothesis H (ex. chi-square)
  - Calculate pdf for t: p(t|H)
  - p-value is probability for t to indicate worse agreement with hypothesis than observed in data
  - Can convert p-value into effective significance under gaussian assumption
    - Ex. p-value of 2.87E-07 —> effective significance of 5σ
  - What is best choice for test statistic?
    - Ideally, we want a test statistic which is independent of nuisance parameters
    - Return to profile likelihood ratio: $\lambda_p(\mu) = \frac{L(\mu, \hat{\hat{\nu}}(\mu))}{L(\hat{\mu}, \hat{\nu})}$

The Problem of Toys

- p-value calculation requires distribution of test statistic under given hypothesis to be known
- How to compute? Generally not analytic
  - Throw (lots of) toys
    - Sample full phase space — POI x nuisance parameter
  - However this is computationally intensive…
- Previously, we were able to use asymptotic approximation of profile likelihood ratio to simplify calculation of confidence intervals (Wilks' theorem)
- Is there an equivalent approximation which simplifies the p-value computation?
  - Yes — Wald's theorem

Asimov Toy

- Wald's theorem gives test statistic distribution for arbitrary POI

$$t_\mu = -2ln\lambda(\mu) = \frac{\left(\mu - \hat{\mu}\right)^2}{\sigma^2_{\hat{\mu}}} + O\left(1/\sqrt{N}\right)$$

- f(tμ|μ') follows non-central chi-square distribution with non-centrality parameter

$$\Lambda = \frac{(\mu - \mu')^2}{\sigma^2}$$

- Only missing piece is $\sigma^2_{\hat{\mu}}$

- Take from Asimov data set, i.e. 'representative event'

Significance Thresholds

- 5 sigma = ‘observation’
  - Very low probability (<0.0001%) of false positive, but achievable with current particle physics statistics
  - Arbitrary value — not always hard rule
    - Higgs discovery in 2012 had 5.1 (4.5) σ significance for ATLAS (CMS), but both claimed observation
- 3 sigma = ‘evidence’
  - Generally used for processes we expect to see, e.g. rare SM — sign that data agrees with model, even if we can’t claim observation yet

Beyond 5 sigma discovery

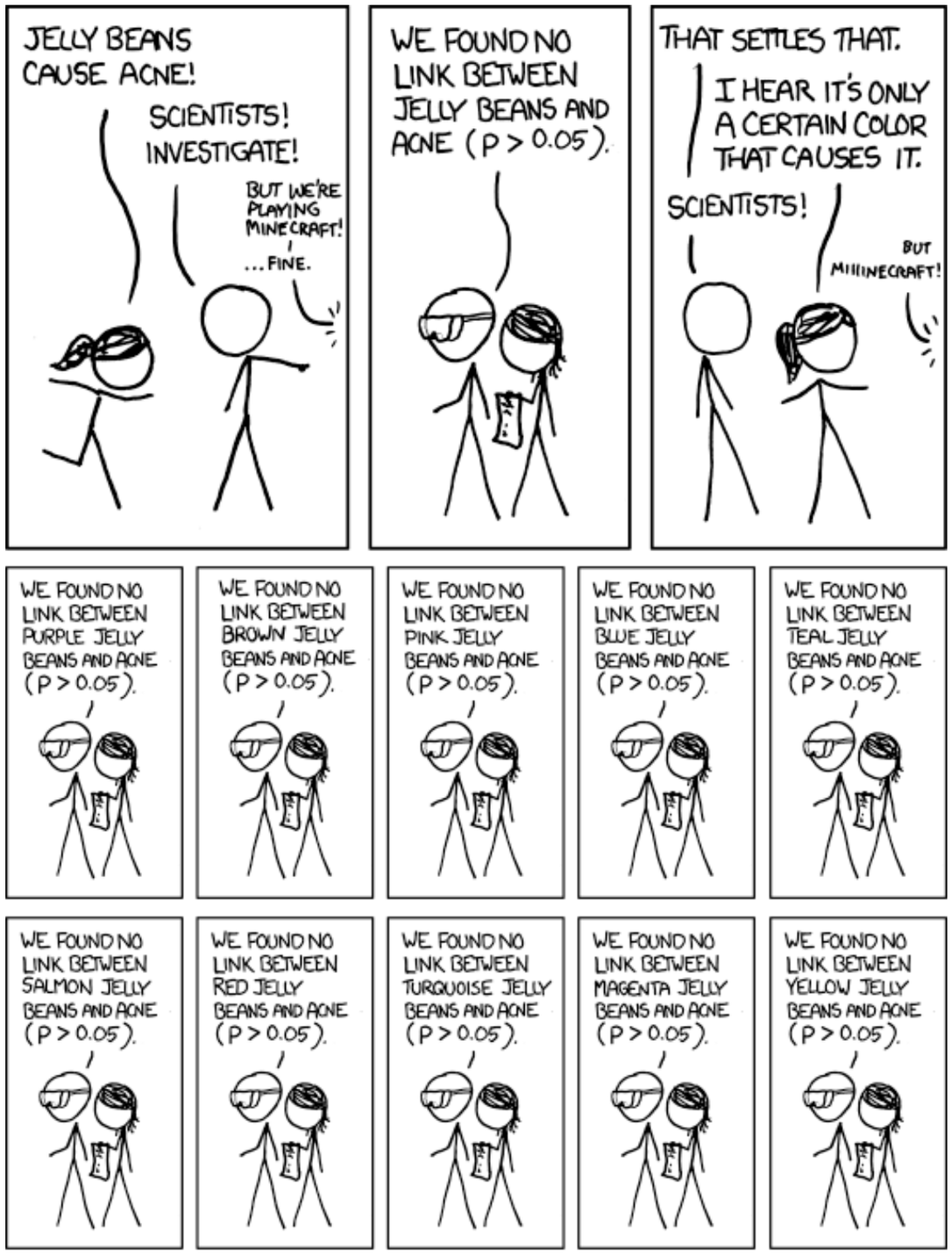
Search	Degree of surprise	Impact	LEE	Systematics	Number of σ
Higgs search	Medium	Very high	Mass	Medium	5
Single top	No	Low	No	No	3
SUSY	Yes	Very high	Very large	Yes	7
B_s oscillations	Medium/low	Medium	Δm	No	4
Neutrino oscillations	Medium	High	$sin^2(2\theta), \Delta m^2$	No	4
$B_s \to \mu\mu$	No	Low/Medium	No	Medium	3
Pentaquark	Yes	High/very high	M, decay mode	Medium	7
$(g-2)_\mu$ anomaly	Yes	High	No	Yes	4
H spin $\neq 0$	Yes	High	No	Medium	5
4^{th} generation q, l, ν	Yes	High	M, mode	No	6
$v_\nu > c$	Enormous	Enormous	No	Yes	>8
Dark matter (direct)	Medium	High	Medium	Yes	5
Dark energy	Yes	Very high	Strength	Yes	5
Grav waves	No	High	Enormous	Yes	7

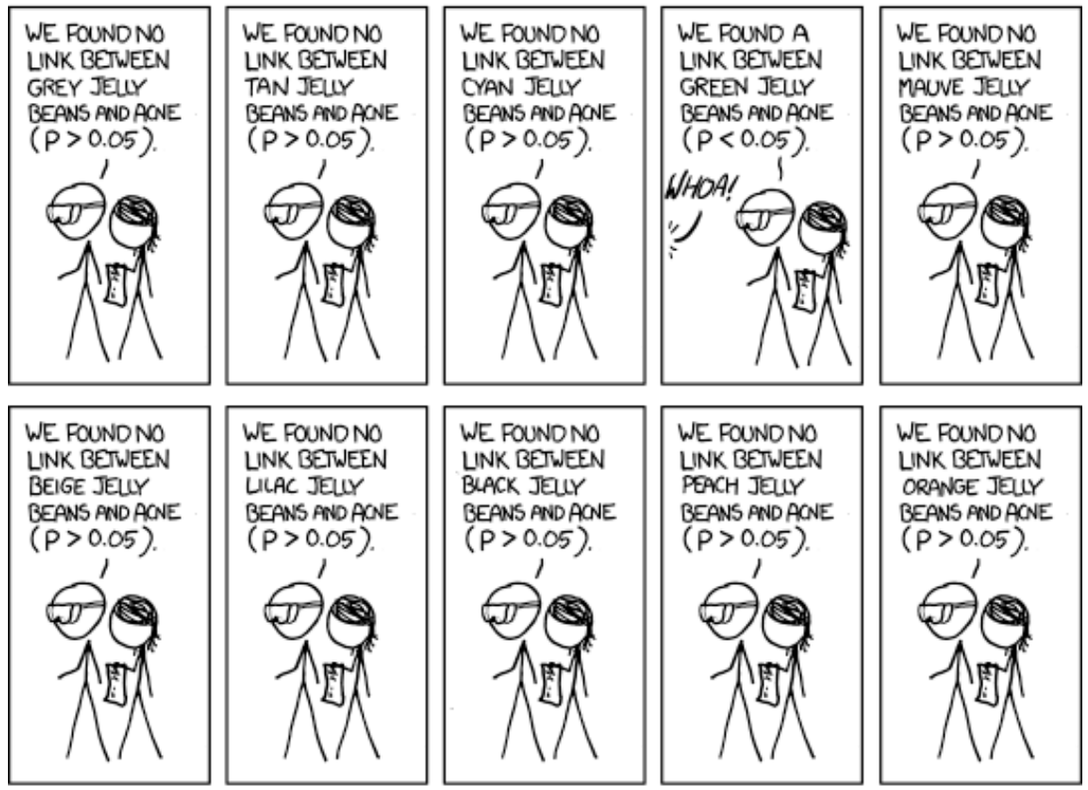
Louis Lyons: Discovering the Significance of 5 sigma

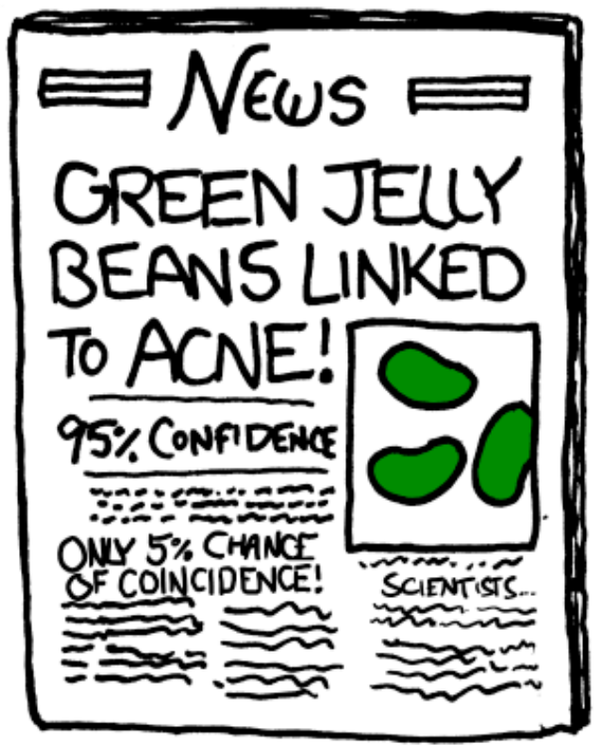
Local vs global significance

- For bump hunt searches, generally looking for a narrow resonance somewhere in a large mass range
  - Local significance: standard definition
  - Global significance: significance accounting for results in full mass range
- ‘Look-elsewhere effect’
  - The larger mass range you search, the more likely a statistical fluctuation will be present somewhere in that range
  - Proper treatment of significance includes where you *don’t* see the signal, as well as where you do
  - Global significance ~ local significance / degrees of freedom

Look-elsewhere effect







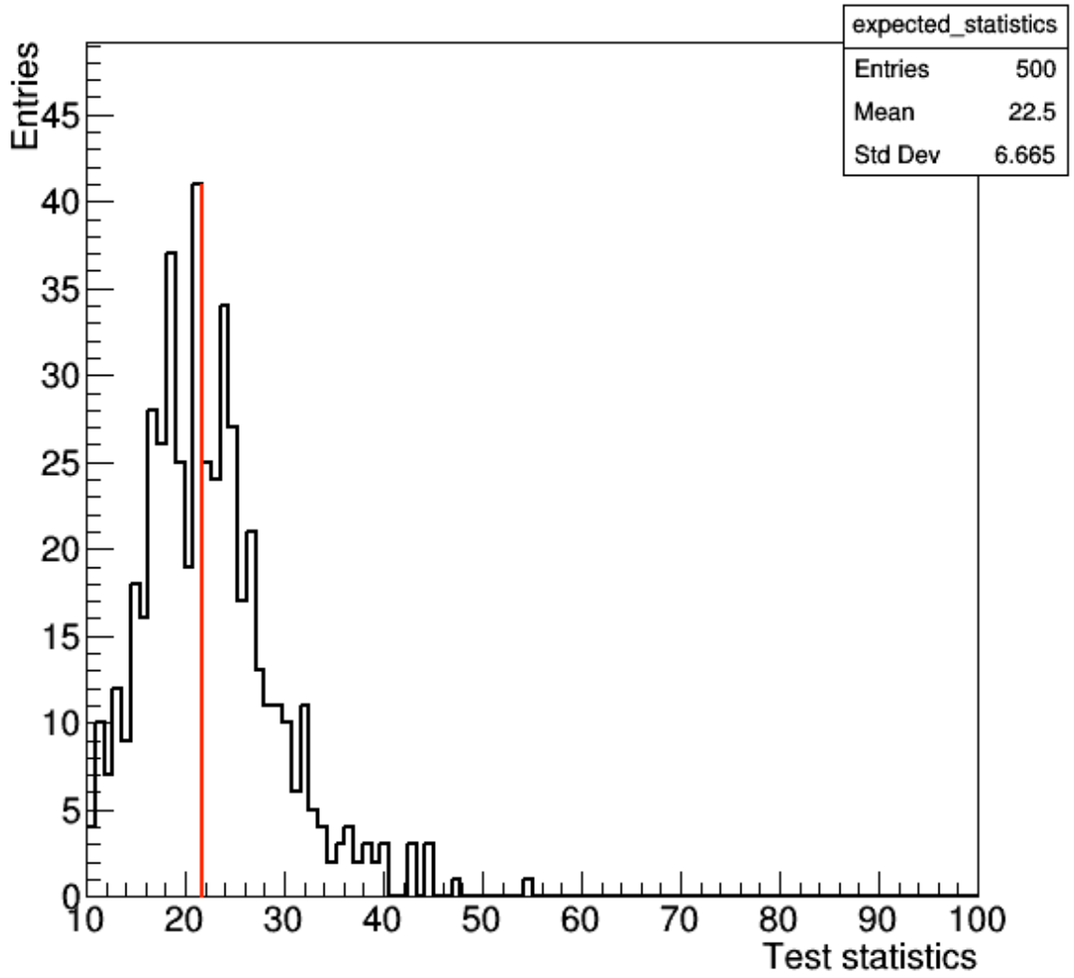
source: xkcd

Goodness Of Fit

- Sometimes, instead of accepting / rejecting hypothesis, want to simply measure how well hypothesis describes data
  - Is mismodeling covered by systematic uncertainties?
- Define goodness of fit test statistic as

$$log(\frac{L(H)}{L(H, saturated)})$$

  - where saturated model exactly matches observation
- Use toys (thrown from H) to generate test statistic distribution and measure p-value of data test statistic



source: combine documentation

Resources

- CMS Statistics Committee
  - cms-talk
  - Regular meetings: Monday odd weeks, 17:00 CERN time (indico)
  - twiki
- Other resources
  - PDG statistics review