Physics 129: Particle Physics
Lecture 6: Probability and Statistics

September 15, 2020

- Suggested Reading:
  - PDG Reviews on Probability and Statistics listed under *Mathematical Tools*
  - Slides from Kyle Cramer's 2020 Hadron Collder Summer School lectures:

    https://indico.fnal.gov/event/43762/contributions/192663/attachments/132852/163529/lecture1.pdf

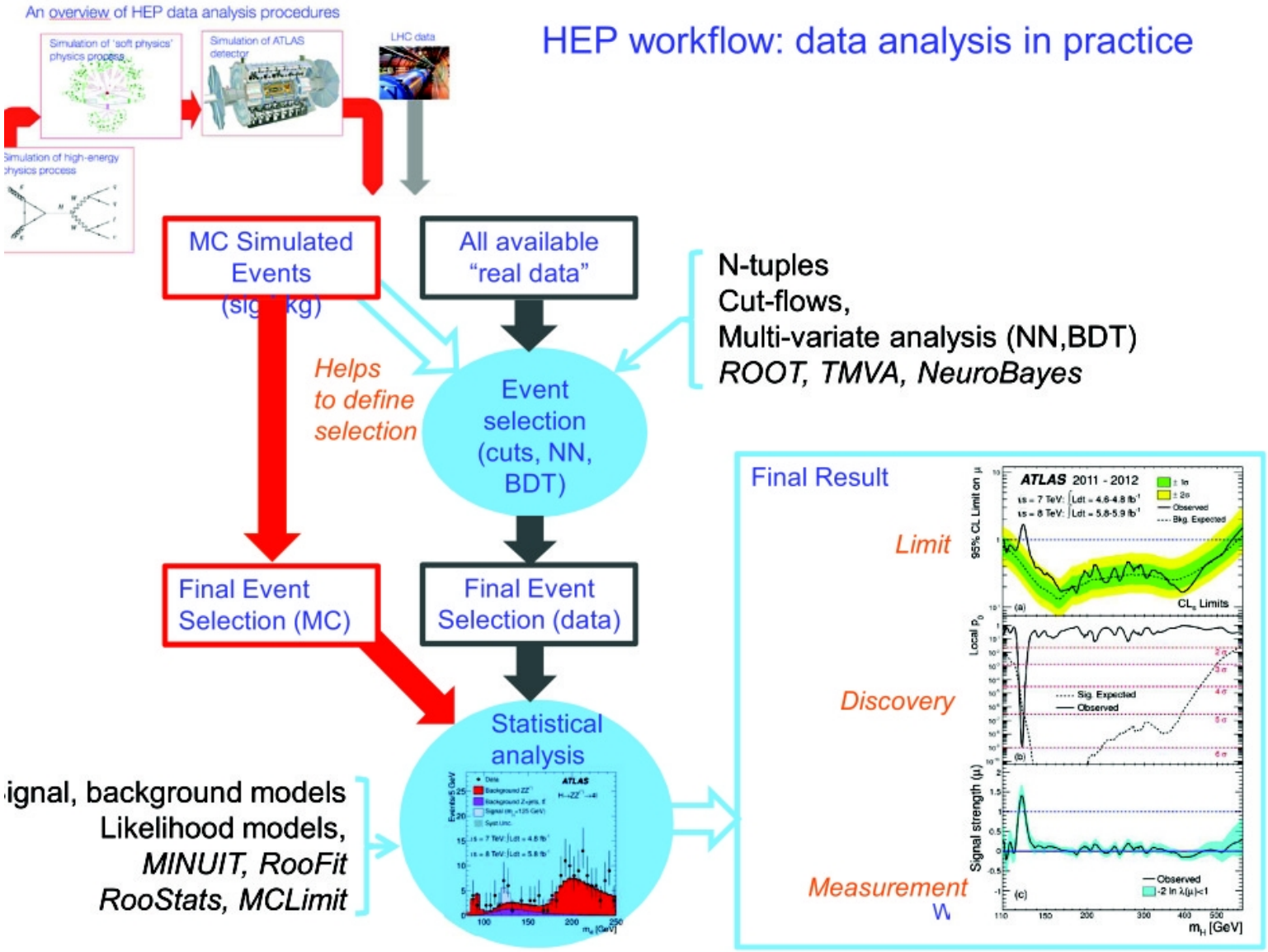
    https://indico.fnal.gov/event/43762/contributions/192672/attachments/132731/163512/HCPSS-stats-lectures-2020.pdf

Introduction

- Physics is based on experimental measurements
- Must understand precision and accuracy of these measurements
- Must also determine whether data is consistent with our theory and whether new physics could be hiding in the data

Statistics provides the tools to do this

How particle physicists analyze data



W. Verkerke, 2014 European Particle Physics Summer School

Probability: Basic Definitions and Axioms

- Probability P is a real-valued function defined by axioms:
  1. For every subset A in S, $P(A) > 0$
  2. For disjoint subsets ($A \cap B = 0$), $P(A \cup B) = P(A) + P(B)$
  3. $P(S) = 1$
- Bayes Theorem:
  (Conditional Probability $P(A|B) \equiv$ prob of A given B)

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

- Law of Total Probability

$$P(B) = \sum_i P(B|A)P(A_i)$$

- Together these give:

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

Probability: Random variables and PDFs

- For continuous variable x, probability density function (pdf):
  - $f(x;\theta) \equiv$ prob that x lies between x and $x + dx$
  - θ represents one or more parameters defining the distribution (eg, for a radioactive decay, θ is the lifetime)
    - Won't always explicitly write θ in our eq
- Cumulative probability distribution (cdf)

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

  Probability that $x < a$.
- For discrete variables, replace integral with sum
- For any function $u(x)$, expectation value:

$$E[u(x)] \equiv \langle u(x) \rangle = \int_{-\infty}^{\infty} u(x)f(x)dx$$

PDF Moments: Mean and Variance

- Mean value:

$$\mu \equiv \int_{-\infty}^{\infty} x f(x) dx$$

- Variance:

$$\sigma^2 \equiv Var(x) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

σ is called the "standard deviation."

These basic definitions are used essentially everywhere. If we know the pdf, we know how to determine the mean and σ

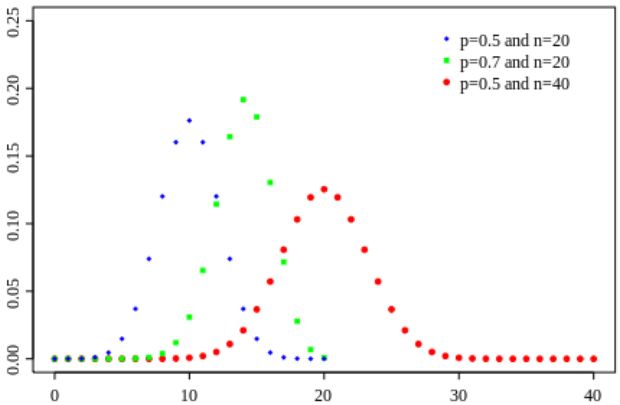
Binomial Distribution [Discrete]

- Random process with two possible outcomes
- p =Prob of outcome #1, $q = 1 - p$ =Prob of outcome #2
- In n trials, prob of getting outcome #1 exactly k times is:

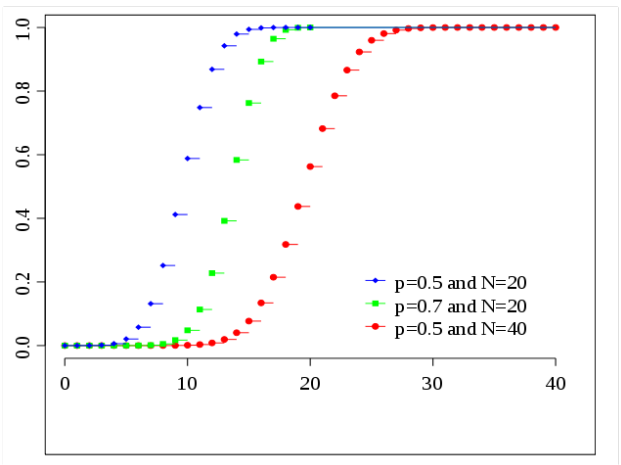
$$f(k; n, p) = \binom{n}{k} p^k q^{n-k} \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $\mu = np$; $\sigma^2 = npq$

Binomial PDF



Binomial Cumulative DF



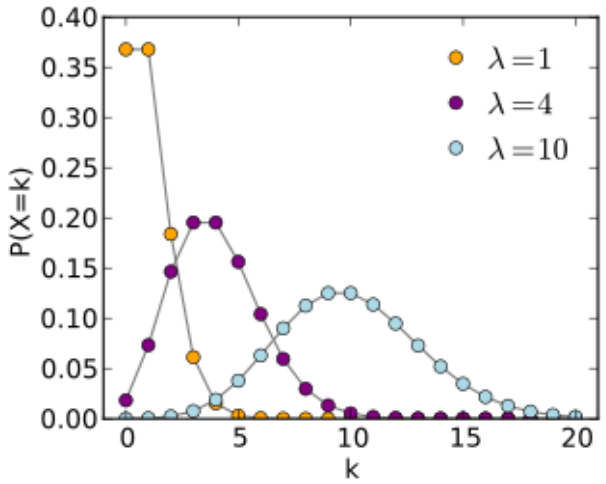
Poisson Distribution [Discrete]

- Prob of finding exactly k events in a fixed the interval of time or space if the events occur with a known constant mean rate λ and independently of the time since the last event

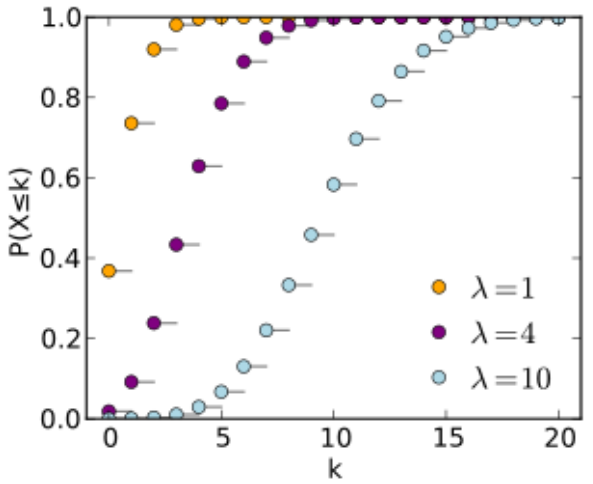
$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

  - $\mu = \lambda$; $\sigma^2 = \lambda$
  - For large λ, approaches a Gaussian
  - For small λ, distribution not symmetric

Poisson PDF



Poisson Cumulative DF



Normal (Gaussian) Distribution [Continuous]

Theorem (Central Limit Theorem)

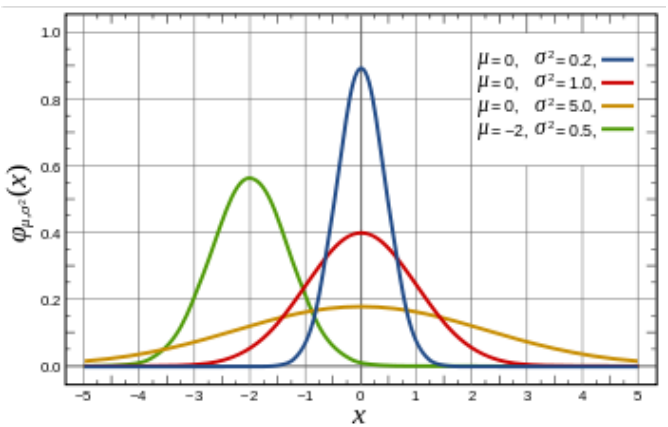
Given random sample $(x_1, x_2, ...x_n)$ *drawn from a pdf with mean* μ *and variance* σ*, the mean of* x *over* n *measurements:*

$$\langle x \rangle \equiv \frac{1}{n} \sum_{1}^{n} x_i$$

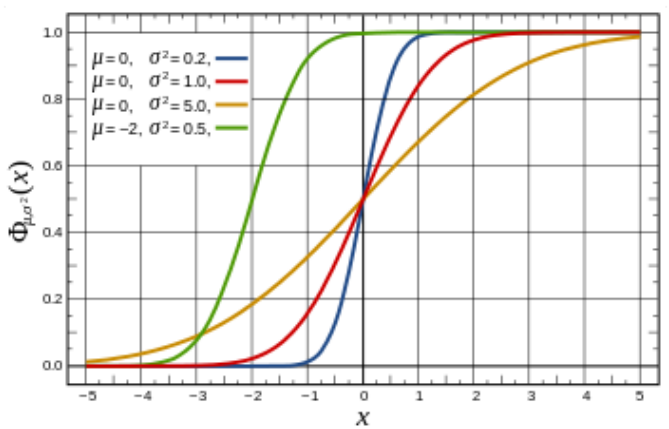
approaches a normal distribution as $n \to \infty$ *independent of pdf*

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian PDF



Gaussian Cumulative DF

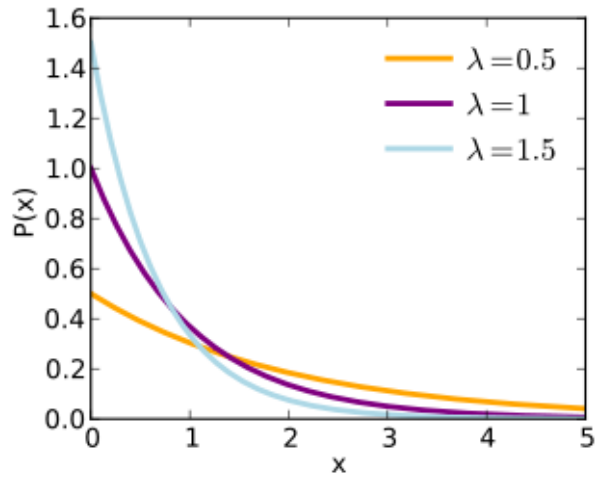


Exponential Distribution [Continuous]

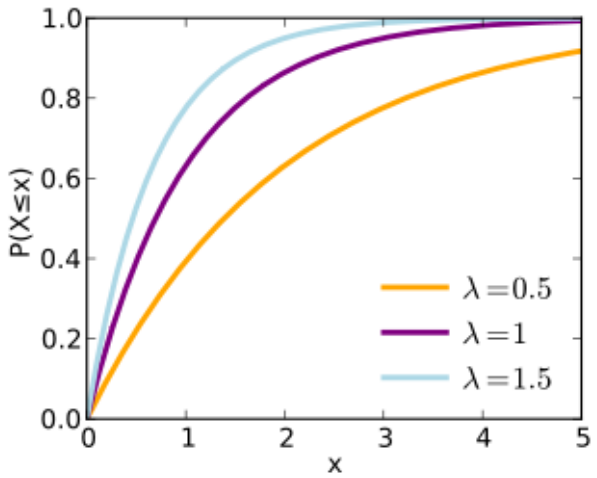
Number of events lost per unit length or time is proportional to number of events

$$\begin{aligned} f(x;\lambda) &= \lambda e^{-\lambda x} \\ \mu &= \frac{1}{\lambda};\ \sigma^2 = \frac{1}{\lambda^2} \end{aligned}$$

Exponential PDF



Exponential Cumulative DF



Statistical Estimators

- One aim of statistical analysis: estimate true value of one or more parameters from experimental data and understand the uncertainty on that measurement
- Important characteristics a good estimator are:
  - Consistency: As sample size $\to \infty$, estimate converges to true value
    - Bias $\equiv$ difference between expectation value of estimator and true value of parameter
  - Robustness: Estimator doesn't change much if true pdf differs slightly from assumed pdf (eg tails in distributions)
- We also want to know the uncertainty on our estimate (how far might the true parameter be from our estimate due to statistical fluctuations in the ensemble of measurements)

The Method of Least Squares

- Assume our measurements are made with high enough statistics that we can assume we are in the Gaussian regime
- We want to find the best estimates of the parameters of function that describes the data
- Do this by minimizing the scatter of data from fit function, taking into account uncertainties on data points
- Scatter defined in terms of χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}$$

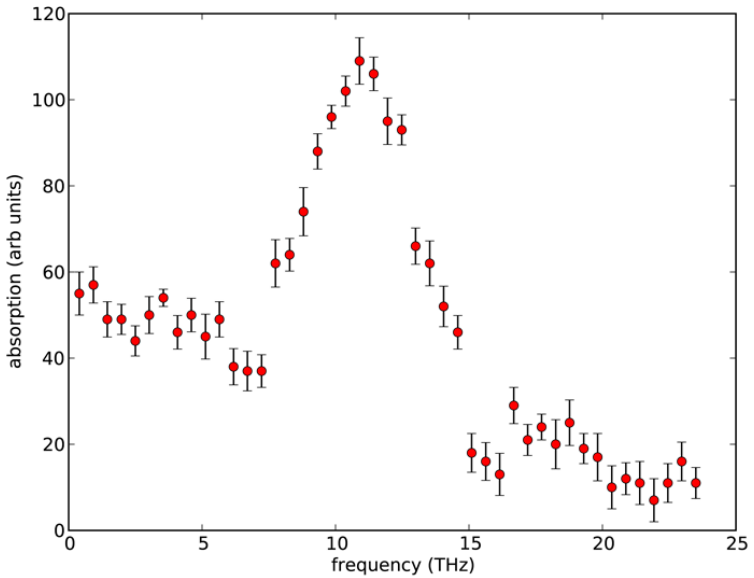
- We can write the χ^2 in terms of our observables

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - F(x_i, \theta))^2}{\sigma_i^2}$$

- Minimize χ^2 with respect to θ (or multiple θ_i)

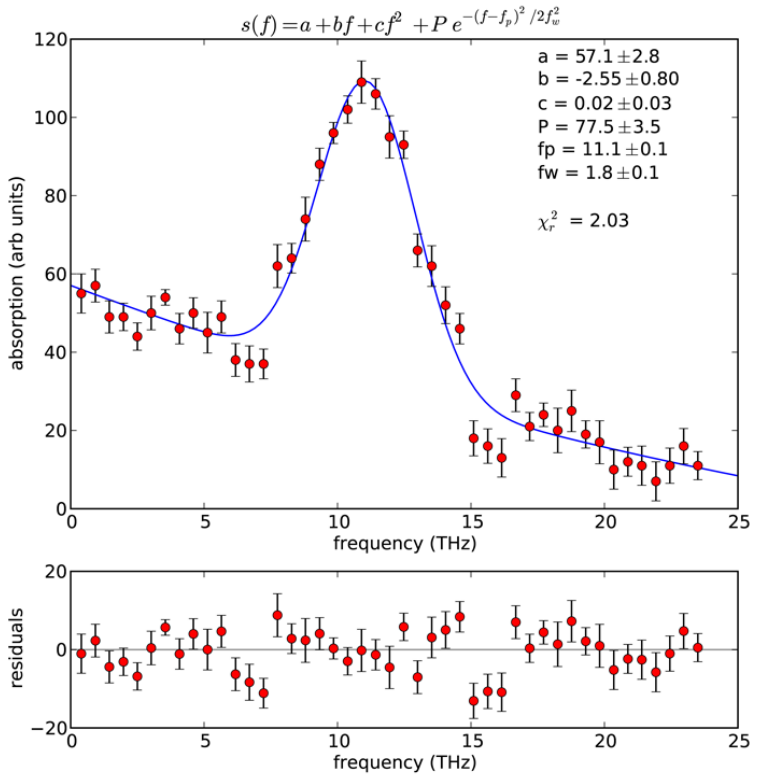
Example of a χ^2 Fit

The Data



Fit to:

$$s(f) = a + bf + cf^2 + Pe^{-(f-f_p)^2/2f_2^2}$$



- 44 data points
- 6 parameters in the fit
- RH plot shows the best fit function and below that the "residuals"
- Number of degrees of freedom: $N_{dof} = N_{data\ points} - N_{fit\ params}$
- $\chi_r^2 = \chi^2/N_{dof}$ (we'll come back to this on pg 20)
- See: https://physics.nyu.edu/pine/pymanual/html/chap8/chap8_fitting.html for the python code used to perform this fit

Likelihood Function

- Likelihood $\mathcal{L}(x;\theta)$ (often just written as $\mathcal{L}(\theta)$):

  Given a specific measurement x and a model with free parameters θ_i, what is the probability that the observed data x would be produced with the specified values of the parameters θ_i

  - Likelihood is a tool for summarizing the data's evidence about unknown parameters.
  - To determine likelihood, must know both the theory and the values of any parameters the theory depends on

- If we have an ensemble of measurements, overall likelihood obtained from product of the likelihoods for the measurements

$$\mathcal{L}(x;\theta) = \prod_{i=1}^{n} \mathcal{L}_i$$

Here θ represents one or more parameters of the theory

Log Likelihood

- To estimate parameter(s) θ, maximize the likelihood
- Usual technique to find maximum: set derivative equal to zero
- Easier to maximize $\ln \mathcal{L}$

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial \theta} &= \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} \mathcal{L}_i \\ &= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln \mathcal{L}_i \\ &= 0 \end{aligned}$$

- Maximum of likelihood function corresponds to value of θ that minimizes $-\sum_i \ln \mathcal{L}_i$
- If several θ_i can minimize with respect to each
  - We'll come back to correlations in a few minutes

Poisson example of likelihood

- Count how many events we see in a specific time interval Δt
- Repeat the counting experiment N independent times (called "trials")
  - The results of these trials are the measurements n_i
- Likelihood function for observing n_i if true mean is μ

$$\mathcal{L}(n_i;\mu) = \frac{e^{-\mu}(\mu)_i^n}{n_i!}$$

Product over N measurements:

$$\begin{aligned}
\mathcal{L}(data;\mu) &= \prod_{i=1}^{N} \frac{e^{-\mu}(\mu)^{n_i}}{n_i!} \\
\ln \mathcal{L} &= \sum_i (-\mu + n_i \ln \mu - \ln(n_i!)) \\
&= -N\mu + \left(\sum_i n_i\right) \ln \mu + constant \\
\frac{\partial \ln \mathcal{L}}{d\mu}|_{\hat{\mu}=\mu} &= -N + \frac{\sum_i n_i}{\mu} = 0 \\
\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N} n_i
\end{aligned}$$

As expected, the best estimator is the mean value

Gaussian example of likelihood

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Now take derivative of the log likelihood:

$$\begin{aligned}
\frac{\partial}{d\mu}\left(\ln \mathcal{L}\right)|_{\hat{\mu}=\mu} &= \frac{\partial}{d\mu}\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} + const\right) \\
&= -\sum_i \frac{(x_i-\mu)}{\sigma^2}|_{\mu=\hat{\mu}} = 0 \\
\Rightarrow \quad \hat{\mu} &= \frac{1}{N}\sum_i x_i
\end{aligned}$$

- Warning: The unbiased estimator for σ is

$$\hat{\sigma} = \frac{1}{N-1}\sum_i (x_i-\mu)^2$$

I'll leave this as an exercise for the student!

Binned vs unbinned likelihood functions

- Likelihood formalism works for any well behaved pdf
- The product of the likelihood is a product over measurements
- We can define what we mean by a measurement
- Example: Measure the lifetime of particle of a given species from an ensemble of such particles produced at time $t = 0$ that decay at time t:

$$f(t) = \frac{1}{\tau} e^{-t/\tau}$$

  Two ways to construct a likelihood:

  1. For each decay i measure t_i and take the product of all measured times to get $\mathcal{L}$ (unbinned likelihood)
  2. Make a histogram of the number of decays in bins of time. Now, the measurement is the number of decays in each bin i (binned likelihood)

- The first is a bettter choice for low statistics samples, the second can be faster when the stats are very high

Connecting the Log Likelihood to the χ^2

- From page 17, for Gaussian case

$$\ln \mathcal{L} = -\sum_i \frac{(x_i - \mu)^2}{2\sigma^2} + const$$

- Compare this to

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}$$

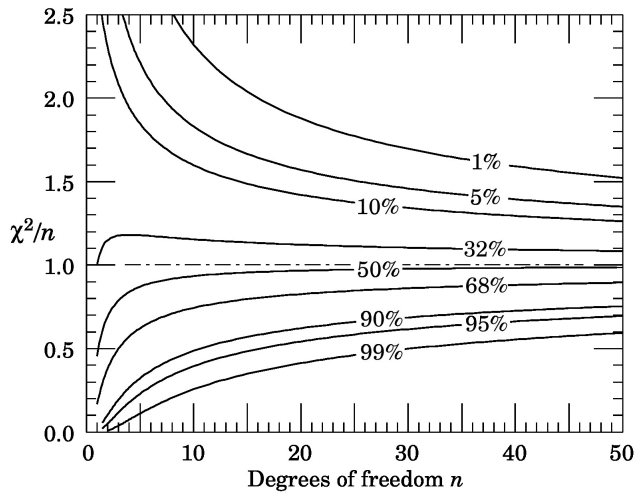
- By inspection, for the case of a Gaussian distribtuion

$$\chi^2 = -2 \ln \mathcal{L}$$

- However: The likelihood formulation works for all pdf's and is therefore more general!

The Uncertainty on the estimate of θ

- χ^2 calculates distance squared (in units of σ between measured distribution and prediction of the model
- "Expect" $\chi^2/N \sim 1$ if model is good
- Probability that χ^2/N is larger than a specific value as a function of n:



- From Gaussian case, can relate $-2\ln\mathcal{L}$ to χ^2
- Uncertainty on parameter θ estimated by by find values where $-2\ln\mathcal{L}$ increase by 1 unit

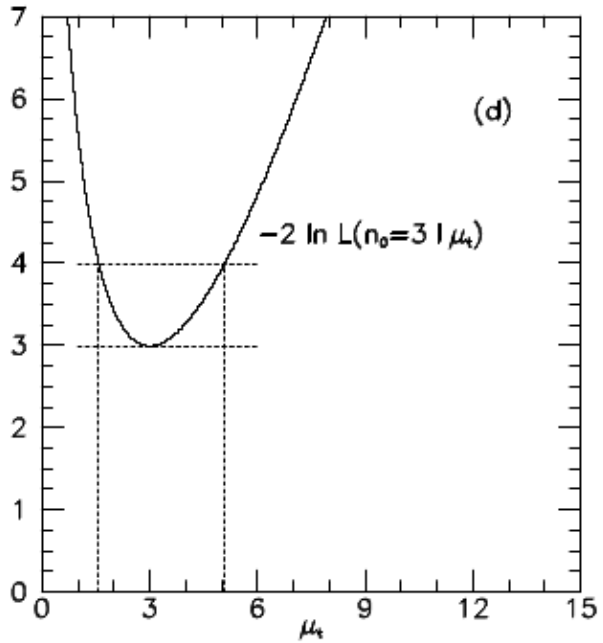


Figure from R. Cousins,
Am. J. Phys. 63 398 (1995)

Correlated Variables

- Often variables we fit for are not independent
- When doing minimization, correlations must be taken into account
- Reminder: variance is:

$$\sigma^2 \equiv Var(x) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

- Define covariance $cov[x, y]$ as

$$cov[x, y] == \int_{-\infty}^{\infty} xy f(x, y) dx dy - \mu_x \mu_y$$

- If x and y are uncorrelated, independent variables, then

$$cov[x, y] = 0 \;\; \text{for } x \neq y \text{ (uncorrelated)}$$

Expanding to N variables

- Each measurement is an ensemble of N quantities x_1 to x_N
- Covariance matrix

$$V_{i\,j} = cov(x_i, x_j) \equiv \langle \langle x_i - \mu_i \rangle \langle x_j - \mu_j \rangle \rangle$$

  is an $N \times N$ matrix
- For uncorrelated variables, it is diagonal
- A related quantity is the *correlation coefficient*:

$$\rho_{i\,j} = V_{i\,j} / \sqrt{V_{i\,i} V_{j\,j}} \equiv V_{i\,j} / \sigma_i \sigma_y$$

  It can be shown that: $-1 \leq \rho_{i\,j} < 1$

The covariance matrix (Gaussian example)

- If x and y are independent variables

$$G(x,y|\mu_x,\sigma_x,\mu_y,\sigma_y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}$$

$$\frac{\partial^2}{d\mu_x^2}(\ln\mathcal{L}) = -\sum_i \frac{1}{\sigma_x^2}$$

  Second derivative wrt μ proportional to $\frac{1}{\sigma^2}$

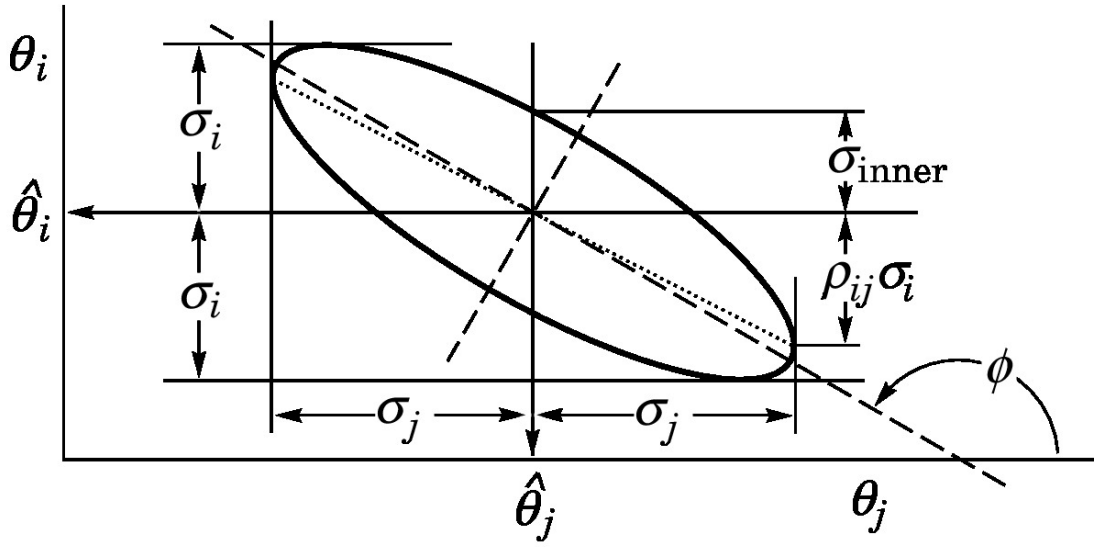
- Now remove assumption that x and y are uncorrelated
- Inverse of the Covariance matrix for a set of Maximum Likelihood estimateors defined by

$$\left\langle \hat{V}^{-1} \right\rangle_{ij} = -\frac{\partial^2 \ln\mathcal{L}}{\partial\mu_i \partial\mu_j}$$

- For binned likelihood in region of large N, where likelihood can be reduced to a χ^2

$$\left\langle \hat{V}^{-1} \right\rangle = \frac{1}{2}\frac{\partial^2\chi^2}{\partial\mu_i\partial\mu_j}$$

Effect of Correlated Uncertainties



- Standard error ellipse for two parameters with a negative correlation
- Slope related to correlation coefficient $d\theta_i/d\theta_j$
  - The θ parameters here correspond to the μ parameters on the previous page
- Correlation matrix typically determined from data numerically during fitting procedure

Propagation of Errors

- Good description found on wikipedia:
  http://en.wikipedia.org/wiki/Propagation_of_uncertainty
- Basic expression is

$$\sigma_f^2 = \left(\frac{\partial f}{\partial \alpha}\right)^2 \sigma_\alpha^2 + \left(\frac{\partial f}{\partial \beta}\right)^2 \sigma_\beta^2 + 2\frac{\partial f}{\partial \alpha}\frac{\partial f}{\partial \beta} COV_{\alpha\beta}$$

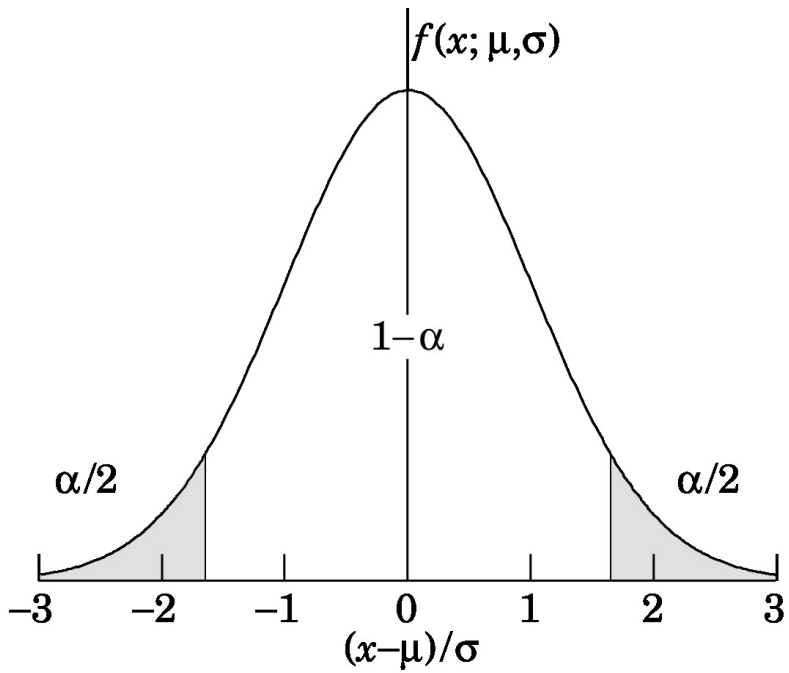
  for case where our model has two parameters α and β
- Extension to more dimensions usually expressed as a matrix
- In case of uncorrelated parameters, reduces to the usual expression you saw in lower division labs

Confidence Intervals

- Fraction of time result is not between x_ℓ and x_u is

$$1 - \alpha = \int_{x_\ell}^{x_u} P(x;\theta)dx$$

- Example for a Gaussian distribution



90% two-sided confidence interval, $\alpha = 0.1$; 5% of the integral shaded on each side.

Confidence Levels for Two Common Distributions

- Gaussian

Table 38.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

- Poisson

Table 38.3: Lower and upper (one-sided) limits for the mean μ of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	$\mu_{\rm lo}$	$\mu_{\rm up}$	$\mu_{\rm lo}$	$\mu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

Here α is fraction outside the region of integration

Goodness of Fit

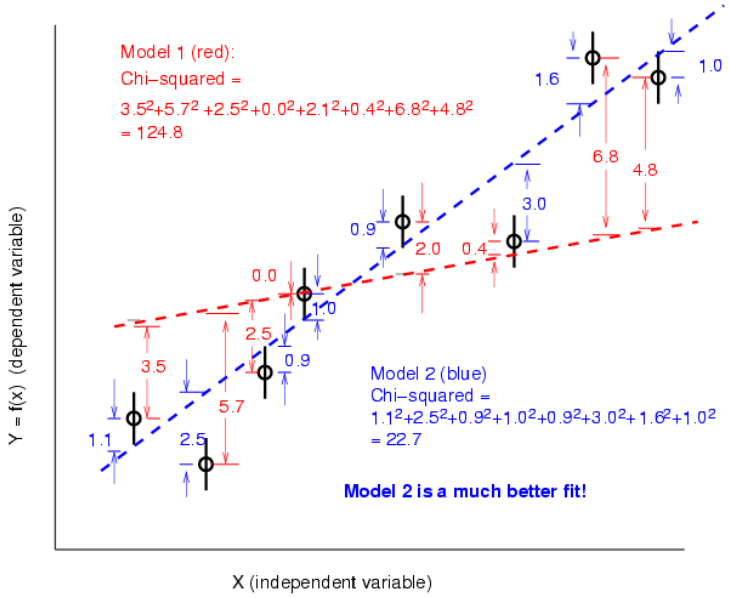
- But estimates of parameters and the uncertainties only makes sense if the model (pdf) used to determine the estimate is correct
- It's not trivial to determine whether a model is good
- Generically, we call parameter determination "fitting" the data
- Determination of whether a model is correct means asking whether the data is consistent with coming from the proposed pdf
  - This is called determining the "goodness of fit"

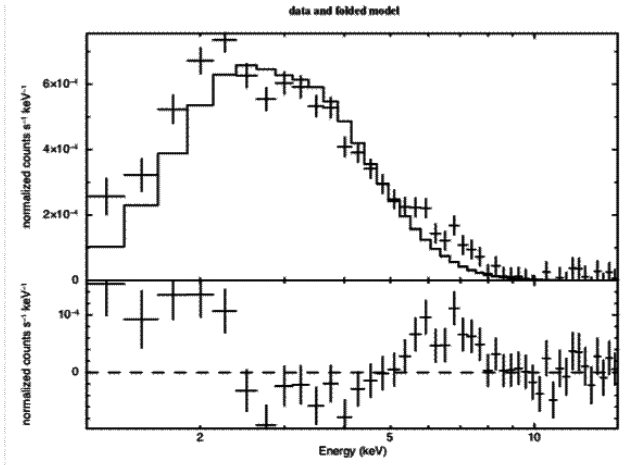
χ^2 Test of Goodness of fit

- Measures distance (in uncertainty space) between data and model

$$\chi^2 = \sum_i \frac{(y(x_i) - f_{model}(x_i))^2}{{\sigma_{y_i}}^2}$$

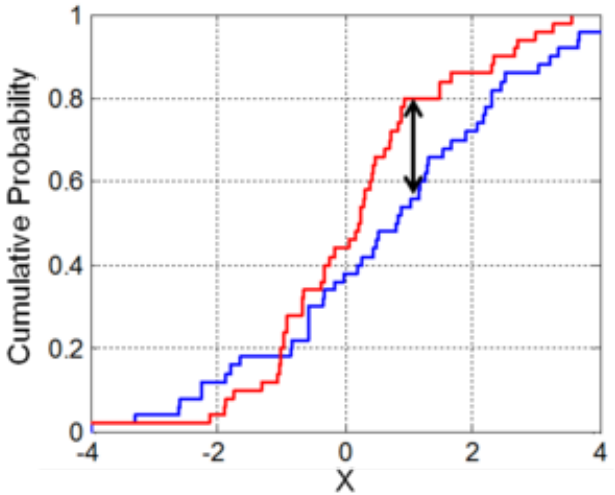
- Only works when uncertainties are symmetric (Gaussian limit)
- Insensitive to whether data is above or below the prediction
  - Not ideal for flagging systematic deviations in shape
  - Eg, for lower plot to the right, χ^2 test gives less discrepancy than you would notice by eye
- Use plot from page 20 to determine probability that data would have a χ^2 value at or larger than what we have measured





Shape Dependent Tests: Kolmogorov-Smirnov (KS)

- Test designed to determine consistency between two datasets
- Can either be two separate measurements (eg do women and men get the same number of colds per year) or a measurement and a prediction (eg real data and MC data)
  - Two samples not required to have same number of events
- No assumptions about the shape of the pdfs, only require that the statistical size of the samples be "large enough"
- For each sample, order data in measured variable (x) and calculate cumulative probability distribution (cpd)



- Measure difference between the two cpd's as function of x
- Identify the largest difference D
- Probability that the two distributions come from the same fundamental pdf:

$$P(D) = 2\sum_{j} = 1^{\infty}(-1)^{j-1}e^{-2j/D^2}$$

see http://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm for a good description

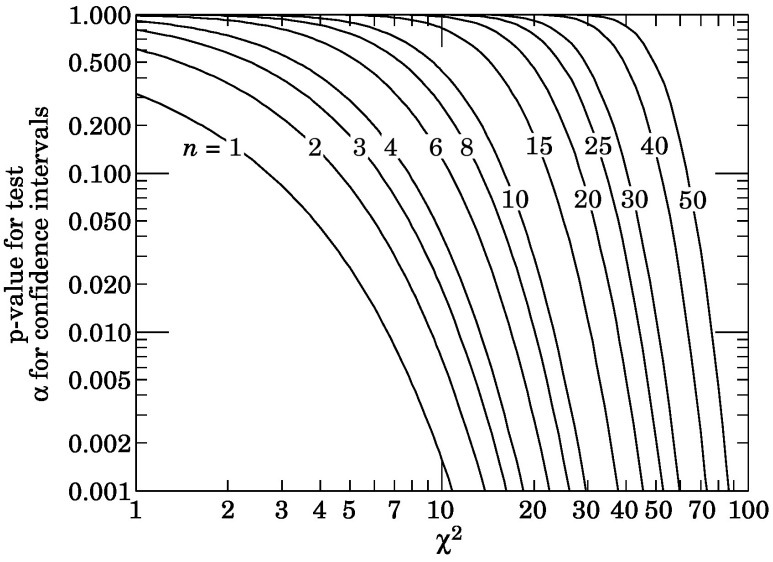
Introduction to Hypothesis Testing

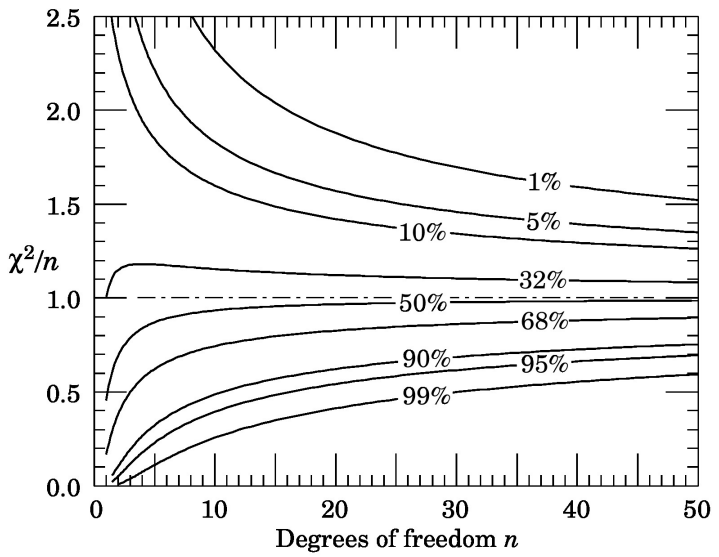
- So far, everything discussed geared to finding best value of parameters and uncertainy, under assumption that we know the pdf
- Nothing in our procedure tells us if data are consistent with hypothesis
- Need statistical tests of whether hypothesis is true
  - Significance tests: How likely is it that signal is just a fluctuation?
  - Goodness of fit tests: Is data consisten with coming from proposed hypothesis?
  - Exclusion tests: How big a signal could be hiding in our data?

Significance Tests

- Suppose we measure a value x_{meas} that is χ^2_{meas} from the prediction
  - How likely is it that we see a value x_{meas} that this far or further from the prediction?
- Suppose we measure a distribution $[x_i]$ of measurements
  - How consistent is our distribution with hypothesis?
- Can use our friend χ^2 to ans these questions:

$$P - value = \int_{\chi^2_{meas}}^{\infty} f(x; n_d) dx$$





The Likelihood Ratio

- Experiments typically have background in addition to signal
- How do we know if there is a significant signal "on top of" the background?
- Given two hypotheses H_B and H_{S+B}, ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{\mathcal{L}(\vec{N}|H_{S+B})}{\mathcal{L}(\vec{N}|H_B)}$$

- See

  https://en.wikipedia.org/wiki/Likelihood-ratio_test

  for more details

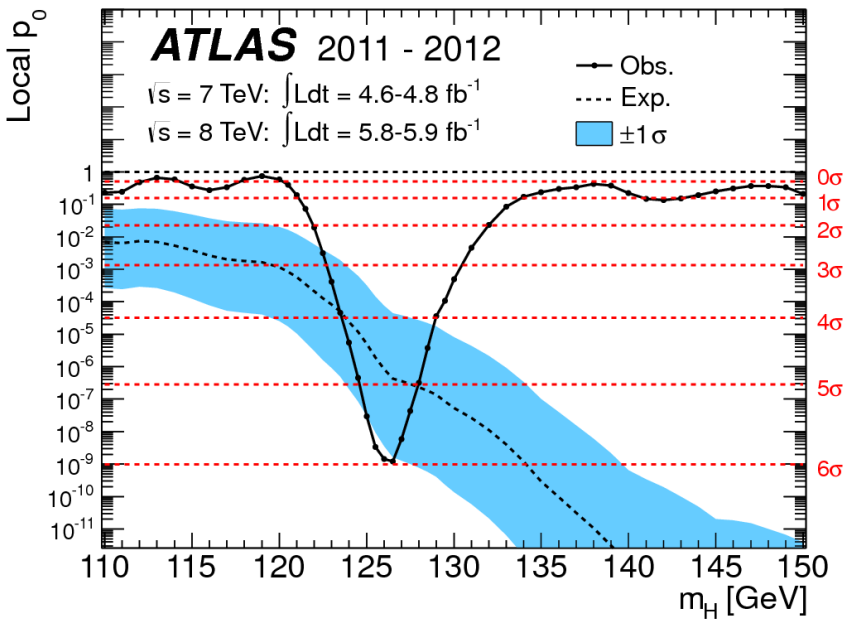
Hypothesis Testing: Searching for New Physics

- Decide which hypothesis is to be rejected (eg, if new physics is present): Call it the NULL hypothesis
  - Often this is the background only hypothesis
- Construct a test statistic such that large values of the statistic would cast doubt on the validity of the NULL hypthesis
- Decide on a threhold for rejecting the NULL hypothesis. Analyze the experimental data, compute the statistic and reject or accept NULL hypothesis depending on whether above or below the threshold
  - Eg a confidence limit
    - 95% c.l. exclusion means only 5% of an an ensemble of experiments with same statistics as yours would reject the NULL hypothesis

Two Approaches to Hypothesis Testing

- Fisher: Reject NULL hypothesis if test statistic large enough
  - Essentially a goodness-of-fit test
- Neyman: Compare NULL to an alternative hypothesis using a test statistic that depends on both. Reject NULL if alternative is preferred by specified amount.
- In both approaches, need to model the NULL hypothesis
- In Neyman approach, need also to model the alternative hypothesis
  - Often signal+background
- Neyman approach can also be used to distinguish between two choices of model
  - Eg: Use angular distribution of particle decays to distinguish between Spin=1 vs Spin=0 hypotheses)

Example: p-value and Higgs Discovery



- Local p-value vs m_H
- Dotted line is expected p-value for a SM Higgs with that mass
- Warning: "Look-elsewhere" effect
  - When asking how likely something is, must take into account how many places you looked!

Early CMS Higgs spin-parity test of 0^+ vs. 0^-

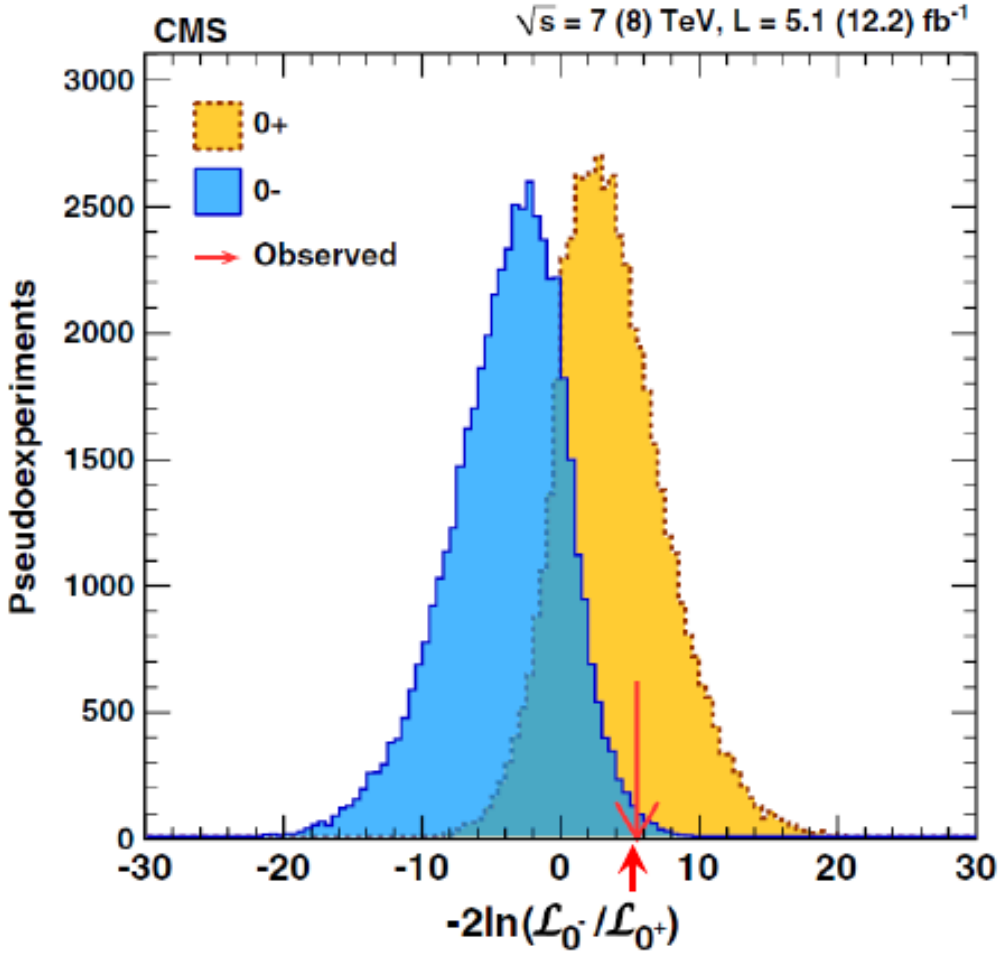


FIG. 3 (color online). Expected distribution of $-2\ln\mathcal{L}_{0^-}/\mathcal{L}_{0^+}$ under the pure pseudoscalar and pure scalar hypotheses (histograms). The arrow indicates the value determined from the observed data. **CMS, Phys. Rev. Lett. 110 (2013) 081803**

Paper reported (fixing typo):

1) $-2\ln(\mathcal{L}_{0^-}/\mathcal{L}_{0^+})$ = 5.5 favoring 0^+

2) for H_0: 0^-, p-value = 0.0072

3) for H_0: 0^+, p-value = 0.7

4) CL_s = (0.0072)/(1–0.7) = 0.024, "a more conservative value for judging whether the observed data are compatible with 0^- "

N.B. See backup for figure and pointer to paper by Demortier and Lyons discussing two p-values in simple-vs-simple case.