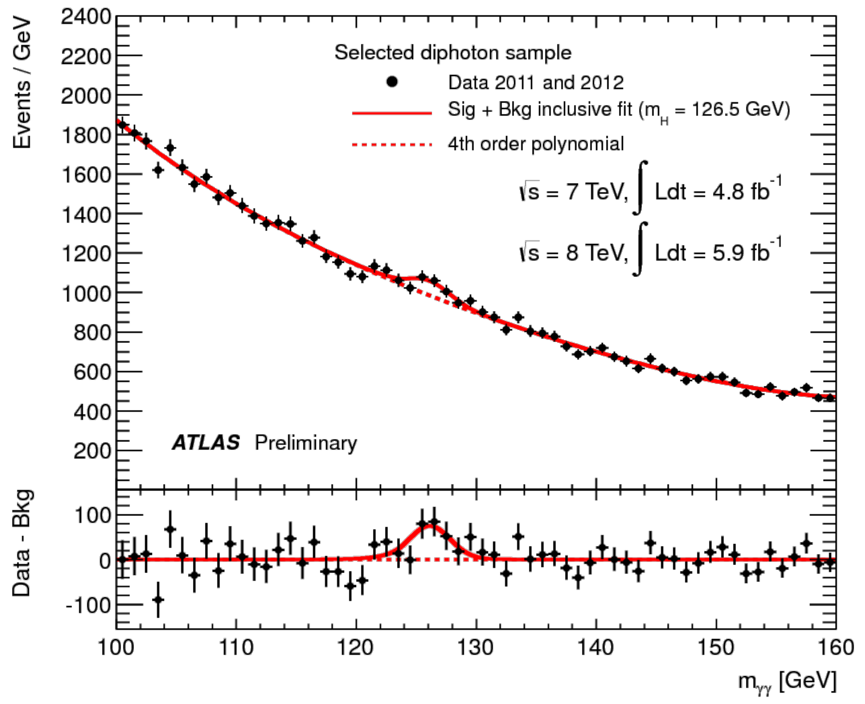
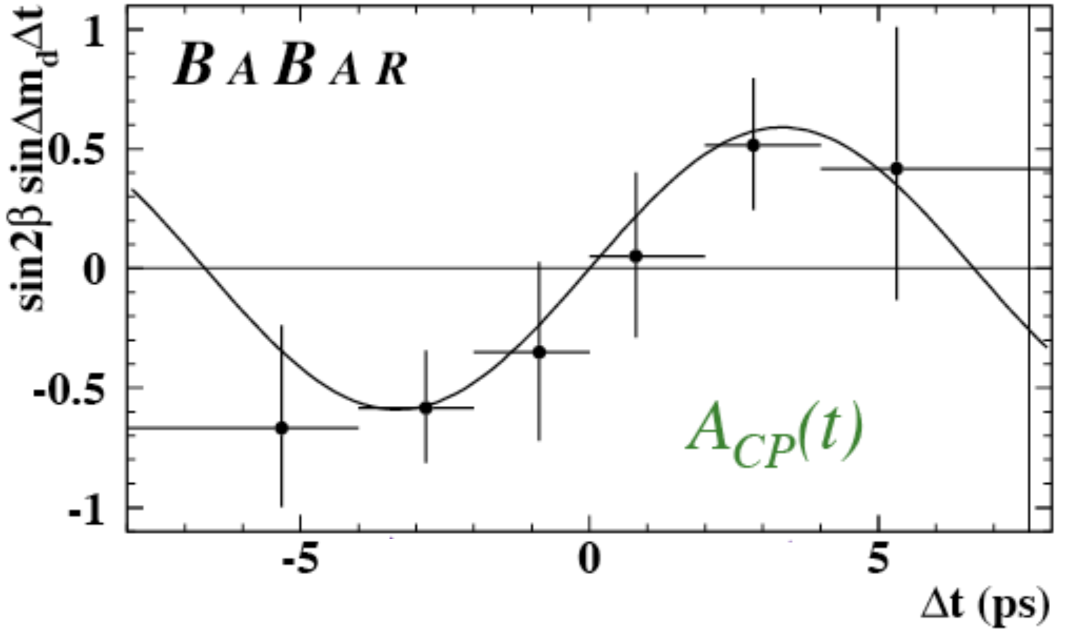
Probability and Statistics

"There are three kinds of lies: lies, damned lies, and statistics."
– Mark Twain, allegedly after Benjamin Disraeli



A refresher

Physics 290E, Spring 2022



A Statistics Refresher

- Intro
- Definitions: results of the experiments
  - ✓ Random variables, probability, PDFs
- Interpreting results
  - ✓ Point estimators
  - ✓ Max likelihood, least squares fits
- Hypothesis testing, confidence limits
- Systematics (time permitting)

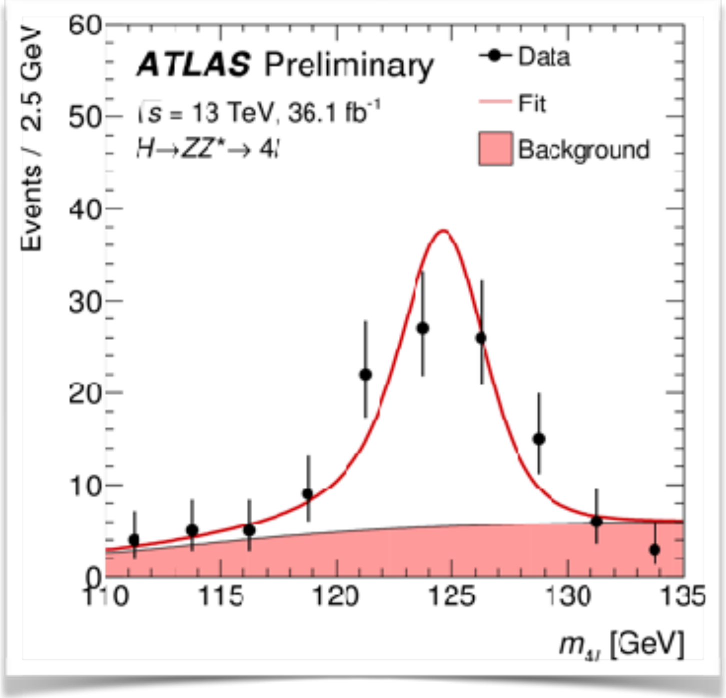
A Statistics Refresher

- Intro
- Definitions: results of the experiments
  - ✓ Random variables, probability, PDFs
- Interpreting results
  - ✓ Point estimators
  - ✓ Max likelihood, least squares fits
- Hypothesis testing, confidence limits
- Systematics (time permitting)

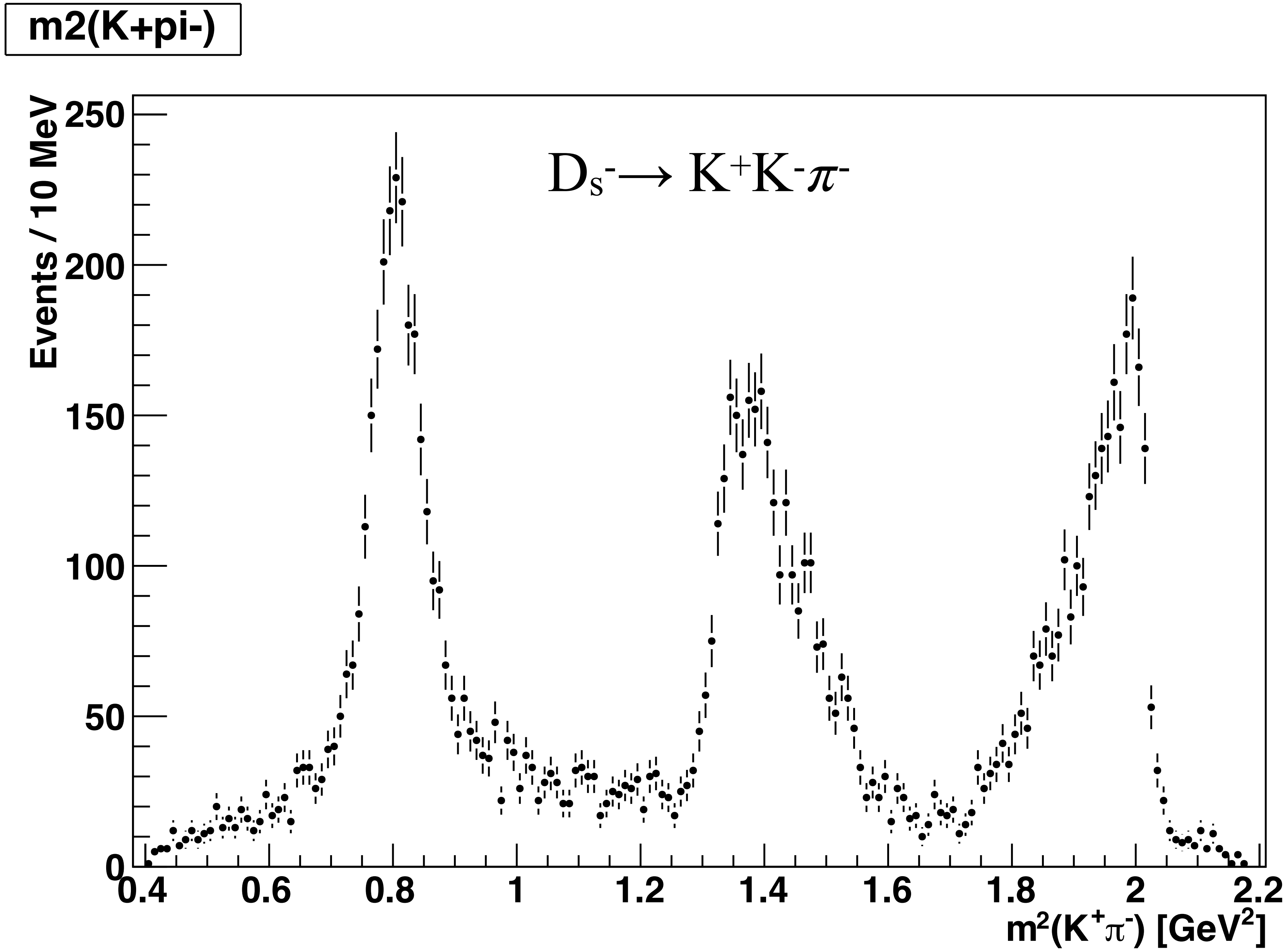
Fell free to yell if you know this and it is boring. Yell louder if I should slow down

Describing the Data

- Data: results of the measurements
  - In physics, we mostly deal with *quantitative data*, i.e. set of numbers
- Interpretation of the data:
  - Range of values of a physical observable
    - ☞ $G_N=(6.67430\pm0.00015)*10^{-11}\ m^3*kg^{-1}*s^{-2}$
  - Consistency with an expectation
    - ☞ Did we discover a new effect ?
  - Relationship between observables
    - ☞ What is the underlying set of parameters that control the process ?



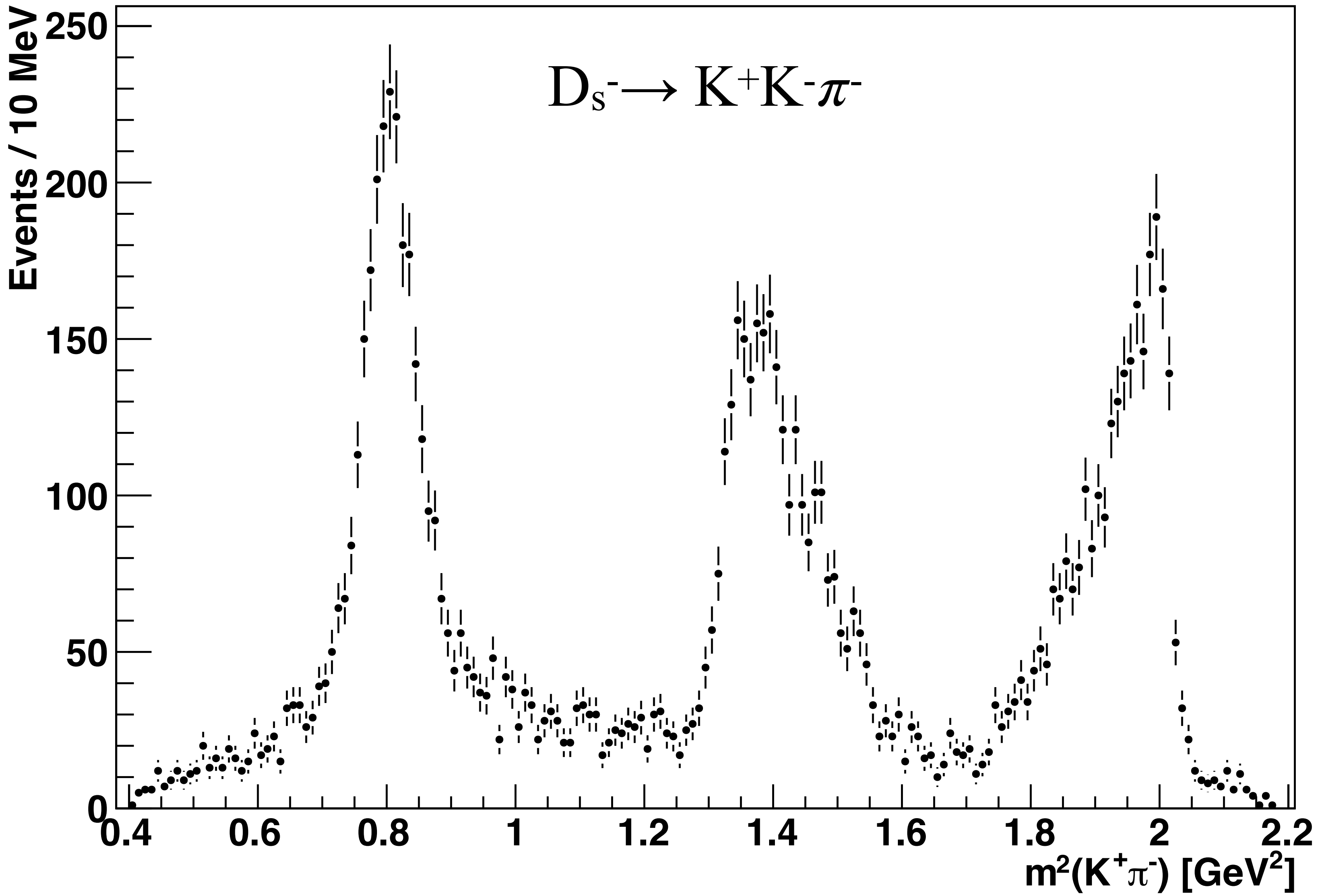
Example #1: Discovering Particles



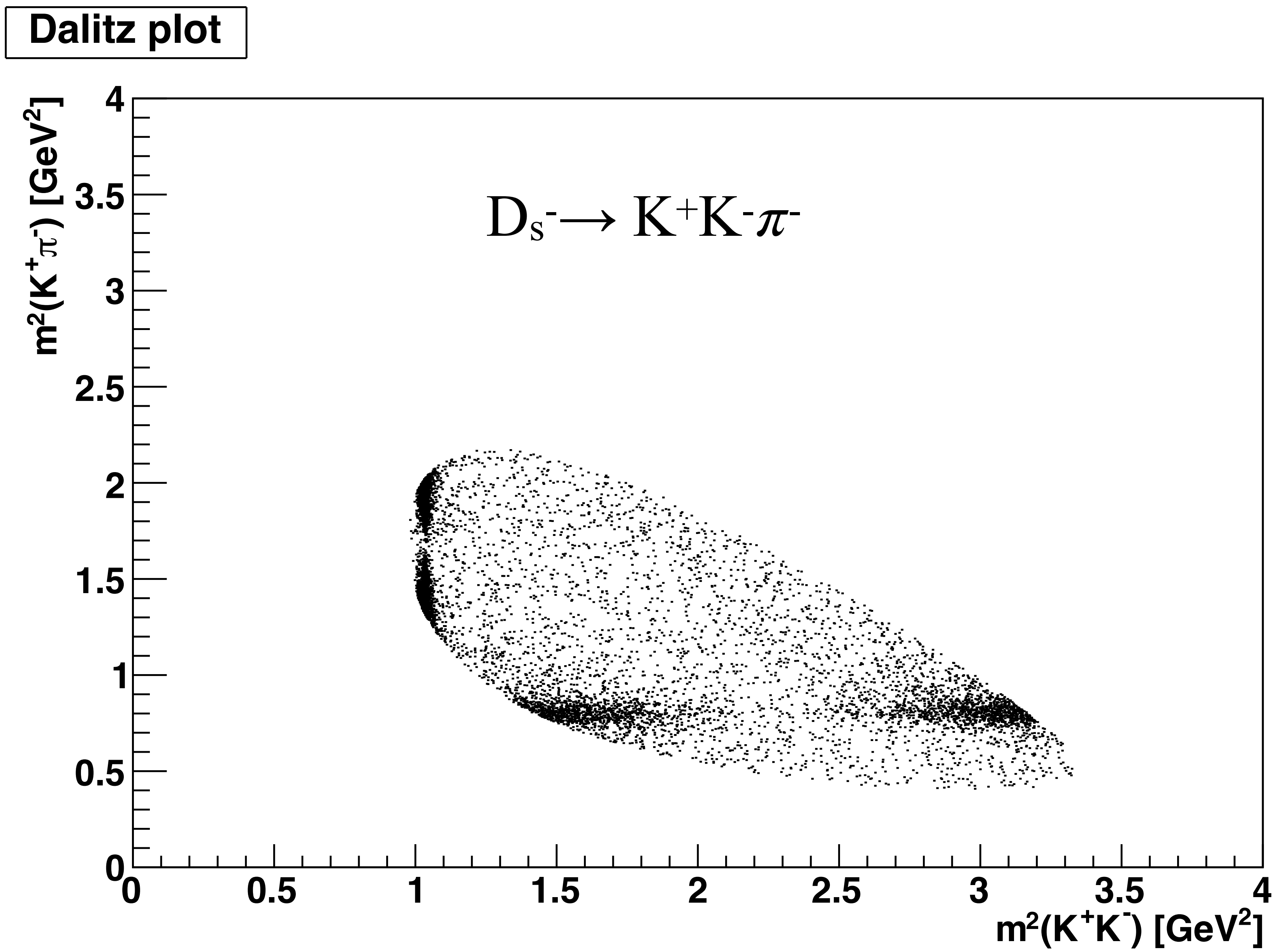
Example #1: Discovering Particles

m2(K+pi-)

How resonances are being produced ?



Example #1: Discovering Particles



Uncertainty and Error

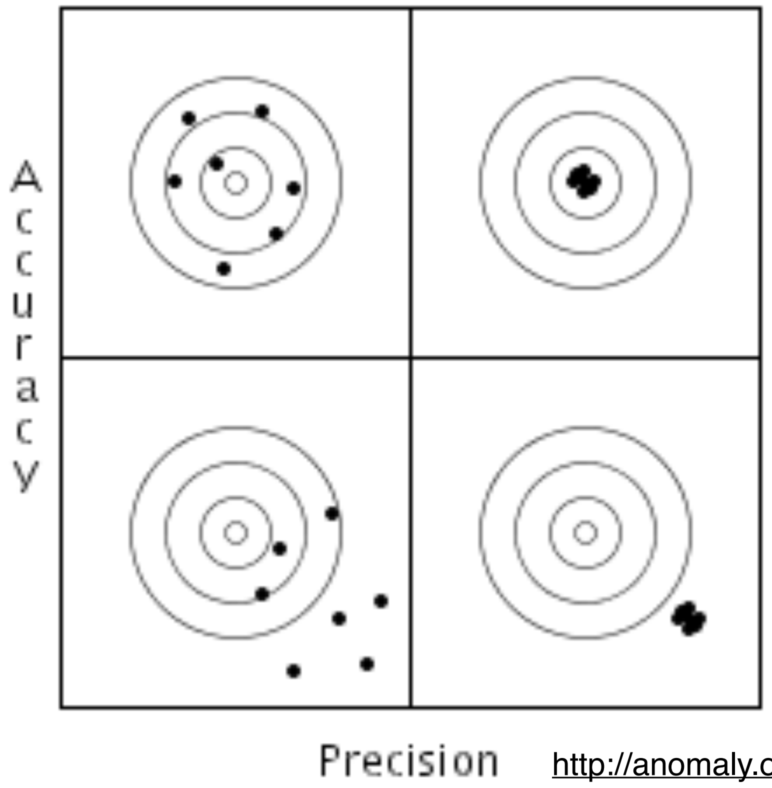
Uncertainty and Error

- In physics, the words "uncertainty" and "error" are used interchangeably to describe how far a particular measurement is expected to deviate from the true value — *typically*
  - ☞ Use symbol σ for the "error"
  - ☞ Formal definition is probabilistic: 68% chance to find the experimental result within $\pm 1\sigma$ of the true value (frequentist interpretation)
  - ☞ Though often interpreted as a range of possible true values (Bayesian interpretation)
  - ☞ We'll come back to the differences between Bayesian and Frequentist statistical approaches later

Uncertainty and Error

- How do we define what is typical ?
  - Underlying assumption: our experiment is one *sample* of a *population* of similar measurements
    - ☞ Derive the value of σ from the properties of the population
  - Implicit assumption: our experiment is mistake-free, i.e. all similar experiments would return similar results

Precision vs Accuracy



http://anomaly.org/wade/blog/2006/01/accuracy_and_precision.html

- Precision: spread of the data around the average value. Typically associated with statistical uncertainty
- Accuracy: deviation of the average value from true value. (bias) Typically associated with systematic uncertainty
- Bad data: “outliers”. Data inconsistent with distribution (e.g. mistakes)

Golden Rules

- When reporting results of a measurement, ALWAYS report its uncertainty
  - And round off values to 1-2 digits of uncertainty:
    - ☞Rule of thumb: 1 digit if the last digit is > 4, 2 digits otherwise
    - ☞ $x = 3.142 \pm 0.024$
    - ☞ $y = 3.1 \pm 0.6$
- Uncertainty can come from the spread in the data and/or precision of the instrument
  - ☞ “Half of last digit” rule of thumb
  - ☞ Statistically correct: $\sigma_{instrument}$ = last digit/sqrt(12)

Probability: Definitions

- For numerical data, probabilistic description is often most convenient (and quantitative)
- Let's define probability now
  - Formally, it is a quantity that defined by Kolmogorov axioms:

1. For every subset A in S, $P(A) \geq 0$;
2. For disjoint subsets ($i.e.$, $A \cap B = \emptyset$), $P(A \cup B) = P(A) + P(B)$;
3. $P(S) = 1$.

Two Interpretations

- "Frequentist" interpretation:
  - Probability is a limiting frequency a given outcome is reported when *experiments* are repeated an infinite number of times
    - ☞ Measurable parameters are represented by "estimators" with assigned confidence levels (CL). CL measures a probability an estimator would fall in a certain range, given a true value of a parameter. No probability is assigned to constants of nature.
- "Bayesian" interpretation:
  - More general: define probability as a *degree of belief* that a given statement is true
    - ☞ E.g. that the true value of parameter x is in interval $[a,b]$
    - ☞ This is somewhat subjective, but follows how most humans think

Frequentist Probability

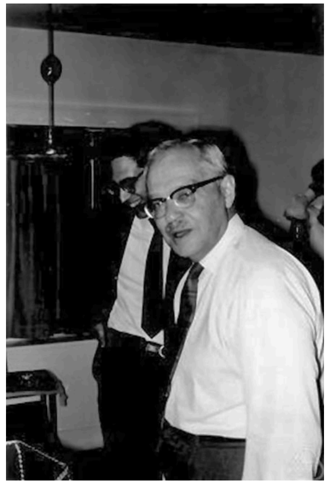
- Defs:
  - Let $\mathcal{S}$ be set of all possible outcomes of a measurement
  - Any subset $\mathcal{A}$ with only one element (single outcome) is *elementary outcome*
  - Define

$$P(A) = \lim_{N\to\infty} (\text{\# of occurrences of A in N trials})/N$$

Assume outcomes are (in principle) repeatable
Confidence in a measurement grows with N

Frequentist statistics is appropriate (and often argued for) in situations where measurements can be reproducibly repeated, so that validity of approach can be tested (e.g. particle physics)

John von Neumann



Jerzy Neyman

Bayes Theorem

Conditional probability of A *given* B

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

The Reverend Thomas Bayes
(1701-1761)

Interpreted within Bayesian statistics as

$$P(\text{theory}|\text{data}) \propto P(\text{data}|\text{theory})P(\text{theory})$$

Posterior probability

Likelihood
(result of the measurement)

Prior probability
(initial prejudice)

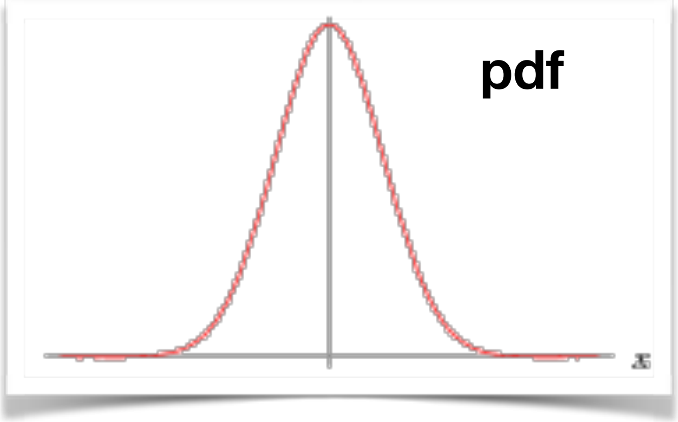
- Allows one to interpret a single experiment as a measure of (subjective) probability that a given hypothesis is correct (e.g. that some fundamental constant is in some range).
- Requires assigning some probability interpretation to prior knowledge. Often useful when *nuisance parameters* (e.g. some parameters of the *theory)* have uncertainties, or when *data* are near a physical boundary. Thus Bayesian Inference is becoming increasingly popular (even in particle physics).
  - But there is an issue of subjectivity in assigning “priors”.

Random Variables

- Random variable: a numerical outcome of a (repeatable) measurement
- Characterized by a Probability Density Function

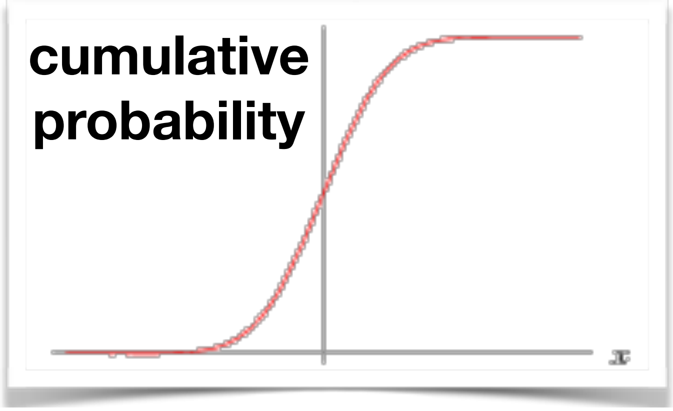
$$dP(x \in [x, x + dx]) = f(x; \theta)dx$$

  - ❑ Depends on a set of parameters θ
    - ☞ C.f. quantum mechanics



- Cumulative distribution (CDF):

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



Expectation Values

Expectation value of function *u(x)*:

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

Moments of a random variable *x*:

$$\alpha_n \equiv E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$$ n-th moment

$$m_n \equiv E[(x-\alpha_1)^n] = \int_{-\infty}^{\infty} (x-\alpha_1)^n f(x)\, dx$$ n-th central moment

Special moments:

$$\mu \equiv \alpha_1 \ ,$$ Mean

$$\sigma^2 \equiv V[x] \equiv m_2 = \alpha_2 - \mu^2$$ Variance

Common PDFs

Distribution	Probability density function f (variable; parameters)	Characteristic function $\phi(u)$	Mean	Variance σ^2
Uniform	$f(x;a,b) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$	$\dfrac{e^{ibu}-e^{iau}}{(b-a)iu}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Binomial	$f(r;N,p) = \dfrac{N!}{r!(N-r)!} p^r q^{N-r}$ $r = 0,1,2,\ldots,N\ ; \quad 0 \le p \le 1\ ; \quad q = 1-p$	$(q+pe^{iu})^N$	Np	Npq
Poisson	$f(n;\nu) = \dfrac{\nu^n e^{-\nu}}{n!}\ ; \quad n = 0,1,2,\ldots\ ; \quad \nu > 0$	$\exp[\nu(e^{iu}-1)]$	ν	ν
Normal (Gaussian)	$f(x;\mu,\sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp(-(x-\mu)^2/2\sigma^2)$ $-\infty < x < \infty\ ; \quad -\infty < \mu < \infty\ ; \quad \sigma > 0$	$\exp(i\mu u - \frac{1}{2}\sigma^2 u^2)$	μ	σ^2
Multivariate Gaussian	$f(\boldsymbol{x};\boldsymbol{\mu},V) = \dfrac{1}{(2\pi)^{n/2}\sqrt{\lvert V \rvert}}$ $\times \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T V^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$ $-\infty < x_j < \infty; \quad -\infty < \mu_j < \infty; \quad \lvert V \rvert > 0$	$\exp\left[i\boldsymbol{\mu}\cdot\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T V \boldsymbol{u}\right]$	$\boldsymbol{\mu}$	V_{jk}
χ^2	$f(z;n) = \dfrac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)}\ ; \quad z \ge 0$	$(1-2iu)^{-n/2}$	n	$2n$

Ex: Binomial Distribution

- Two outcomes of an experiment
  - E.g. Pass and Fail
    - Define probability of Pass to be *p*
    - Probability of Fail is $q=1-p$
- Draw *N* samples
- Define *r* to be the number of Passes (out of *N*)
- Key properties:

$$\langle r \rangle = pN$$

$$V[r] = Npq = Np(1-p)$$

Example: Measure Efficiency

- Generate a sample of N events
- Apply selection; suppose n_{pass} events passed
- Estimate

$$\hat{\epsilon} = \frac{n_{\text{pass}}}{N}$$

$$\sigma(\hat{\epsilon}) = \sqrt{V[\hat{\epsilon}]} = \sqrt{\frac{V[n_{\text{pass}}]}{N^2}} = \sqrt{\frac{\hat{\epsilon}(1-\hat{\epsilon})}{N}}$$

$$\sigma(\hat{\epsilon}) \neq \frac{\sqrt{n_{\text{pass}}}}{N} = \sqrt{\frac{\hat{\epsilon}}{N}}$$

Example: Measure Efficiency

- Generate a sample of N events
- Apply selection; suppose n_{pass} events passed
- Estimate

$$\hat{\epsilon} = \frac{n_{\text{pass}}}{N}$$

$$\sigma(\hat{\epsilon}) = \sqrt{V[\hat{\epsilon}]} = \sqrt{\frac{V[n_{\text{pass}}]}{N^2}} = \sqrt{\frac{\hat{\epsilon}(1-\hat{\epsilon})}{N}}$$

$$\sigma(\hat{\epsilon}) \neq \frac{\sqrt{n_{\text{pass}}}}{N} = \sqrt{\frac{\hat{\epsilon}}{N}}$$

What happens when n_{pass} or n_{fail}=0 ?

Central Limit Theorem

- Let x_1, x_2, …, x_N be independent random variables
  - ☞ Each belongs to a distribution of with a well-defined mean $\langle x_i \rangle$ and variance $V[x_i]$
- Define

$$x \equiv \lim_{N\to\infty} \sum_{i=1}^{N} x_i$$

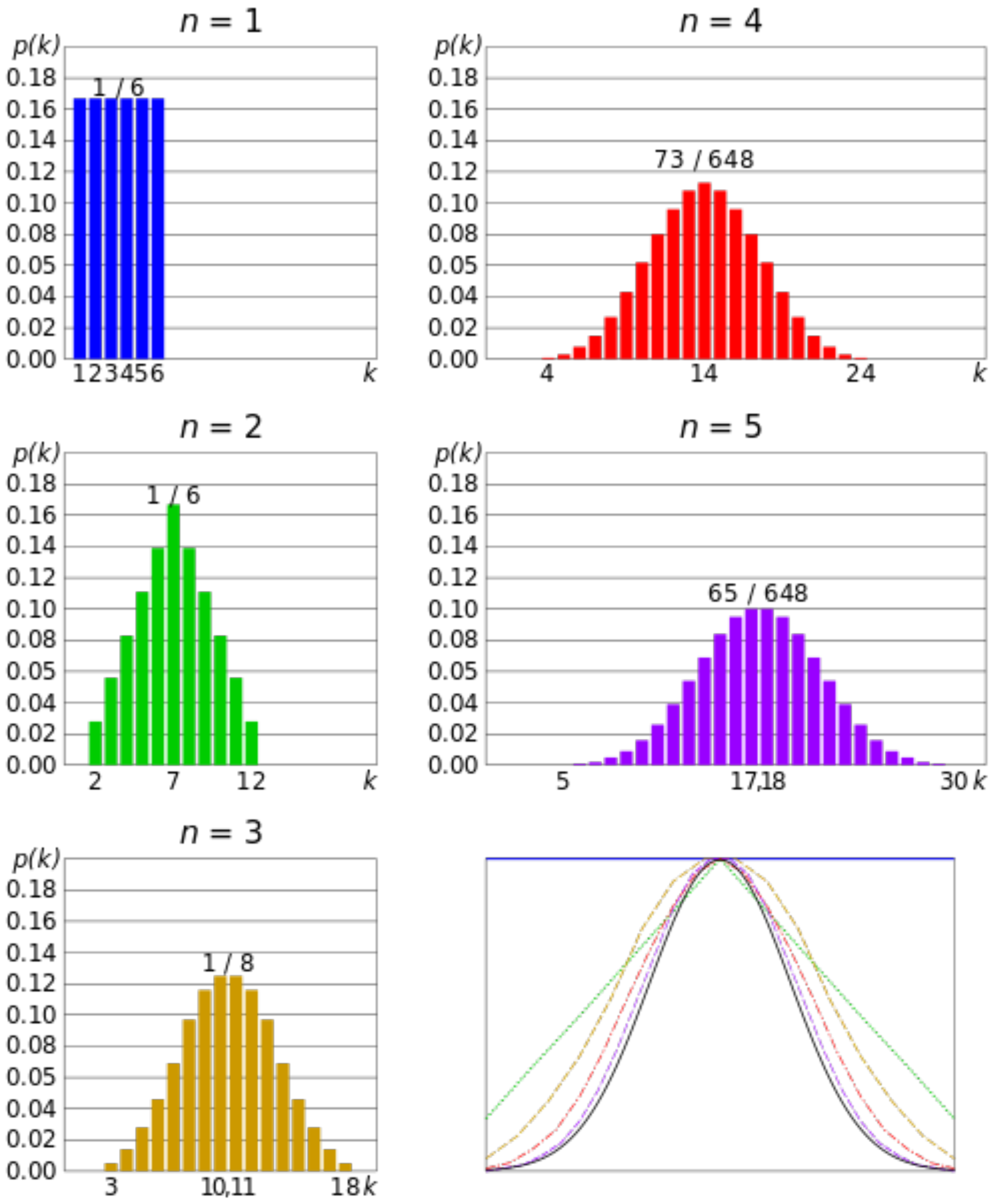
- Theorem: x is Gaussian-distributed with

$$f(x) = g(x; \mu_x, \sigma_x)$$

$$\mu_x = \sum_{i=1}^{N} \langle x_i \rangle$$

$$\sigma_x^2 = \sum_{i=1}^{N} V[x_i]$$

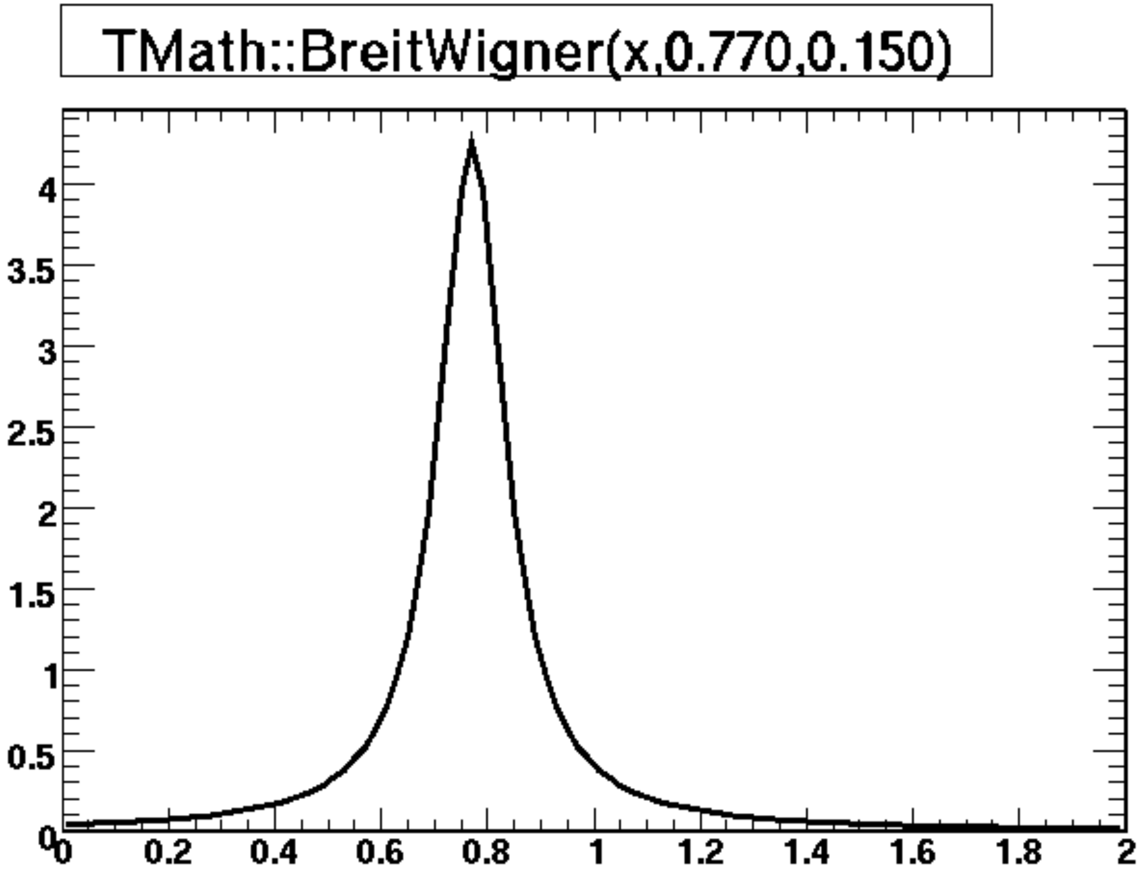
Central Limit Theorem



http://en.wikipedia.org/wiki/File:Dice_sum_central_limit_theorem.svg

Cauchy (Breit-Wigner) PDF

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x0}{\gamma}\right)^2\right]} = \frac{1}{\pi}\left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right]$$



Undefined variance

(central limit theorem does not apply)

Inverse Sqrt Law

- Suppose x_i are drawn from the same distribution with mean μ_x and variance $V[x]$
- Mean of N samples

$$\langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i$$

follows Gaussian distribution:

$$f(\langle x \rangle) = g(\langle x \rangle; \mu, \sigma)$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\langle x_i \rangle \equiv \mu_x$$

$$\sigma^2 = \frac{1}{N^2}\sum_{i=1}^{N} V[x_i] \equiv \frac{V[x]}{N} \quad \rightarrow \quad \sigma(\langle x \rangle) = \sqrt{\frac{V[x])}{N}}$$

"Inverse sqrt law"

Point Estimation

- Standard problem: set of values $x_1, x_2, \ldots, x_n$ described by PDF

Typical goal: estimate the true value of one or more parameters from the experimental data, and understand their uncertainties

- Point estimation: want to construct

$$f(x) \equiv f(x_n; \theta)$$

data — parameter(s)

☞ Estimator of parameter θ

$$\hat{\theta} = \theta(x_1, x_2, \ldots, x_n)$$

Estimator Properties

- Consistency
  - ☞ Approaches true value asymptotically for *infinite* dataset
- Bias
  - ☞ Difference wrt true value for *finite* dataset
- Efficiency
  - ☞ Variance of the estimator (compared to others)
- Sufficiency
  - ☞ Dependence on true value
- Robustness
  - ☞ Sensitivity to bad data, e.g. outliers
- Others: physicality, tractable-ness, etc.
- No “ideal” recipe, what is best depends on the problem

Basic Estimators

- Estimators for mean and variance
- Shape of the PDF (fitting):
  - Maximum likelihood
    - ☞ Most efficient, but may be biased
    - ☞ Goodness of fit is not readily available
  - Least Chisquared
    - ☞ ML for gaussian-distributed data
    - ☞ Convenient for binned data, analytic solutions for linear functions
    - ☞ Automatic goodness-of-fit measure
    - ☞ Be careful of gaussian approximations (e.g. when Poisson becomes Gaussian)

Mean and Variance from a Sample

Estimators:

(equally weighted data)

$$\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{N>0}$$

$$\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \widehat{\mu})^2 \qquad \text{N>1}$$

Variances of these estimators:

$$V[\hat{\mu}] = \frac{\sigma^2}{N} \ \text{ i.e. } \ \sigma[\hat{\mu}] = \sigma/\sqrt{N}$$

$$V\left[\widehat{\sigma^2}\right] = \frac{1}{N}\left(m_4 - \frac{N-3}{N-1}\sigma^4\right)$$

$$\rightarrow \quad \sigma[\hat{\sigma}] = \sigma/\sqrt{2N}$$ for Gaussian distribution of x and large N

Sample Mean and Variance, Weighted

Estimators:

(unequally weighted data)

$$\hat{\mu} = \sum_i w_i x_i, \text{ where} \sum_i w_i \equiv 1$$ N>0

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

The standard case is a collection of points with unequal error bars σ_i.

In this case, the most efficient estimator would use

$$w_i = \frac{1/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$

You can then show that the variance of the mean is

$$V[\hat{\mu}] = \frac{1}{\sum_i 1/\sigma_i^2} \qquad \text{i.e.} \qquad \sigma[\hat{\mu}] = \frac{1}{\sqrt{\sum_i 1/\sigma_i^2}}$$

Maximum Likelihood Estimators

Define likelihood for N independent measurements x_i:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i; \boldsymbol{\theta})$$

→ max to determine estimators of θ

This leads to a system of (generally nonlinear) equations for parameters θ:

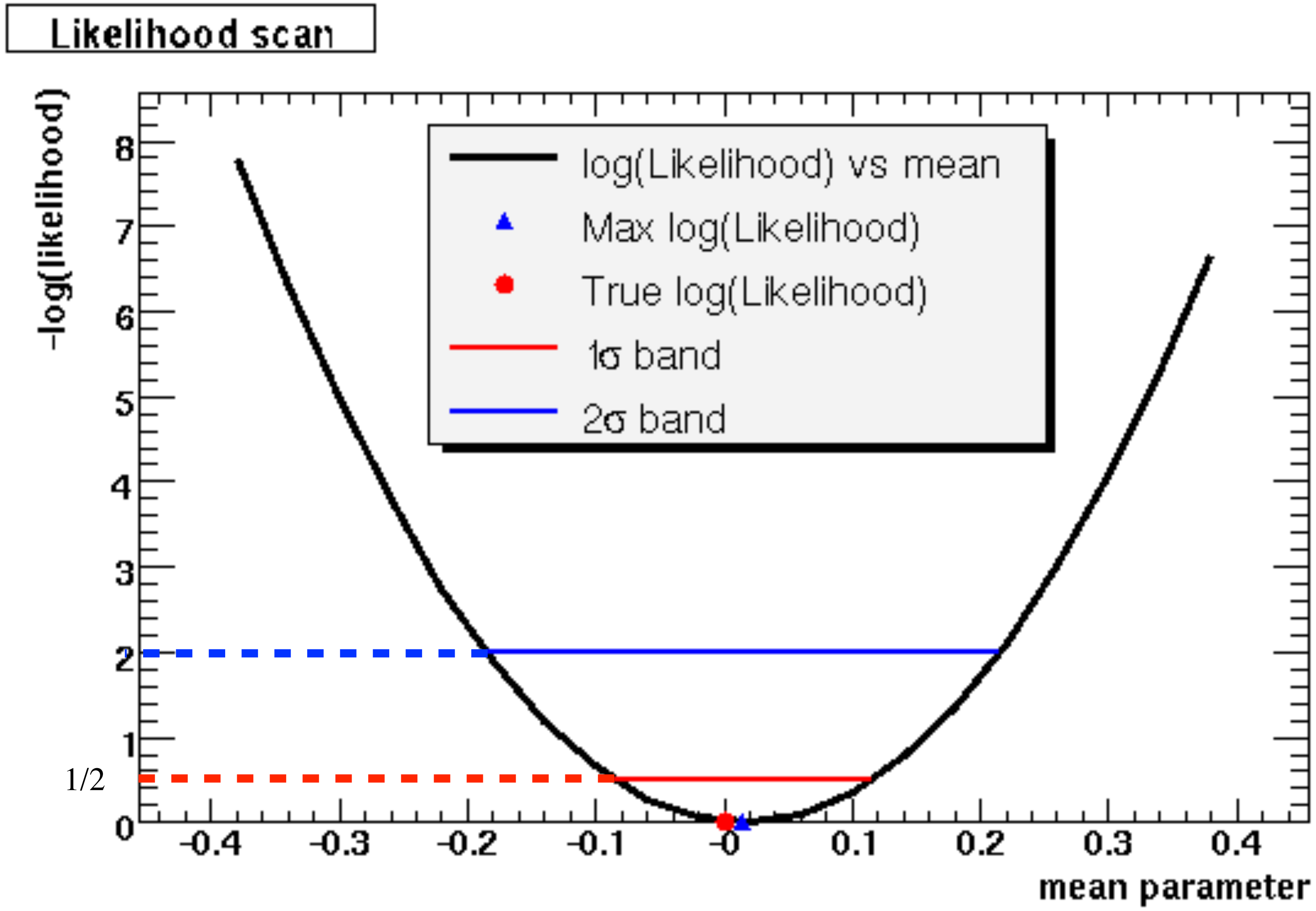
$$\frac{\partial \ln L}{\partial \theta_i} = 0 \ , \qquad i = 1, \ldots, n$$

Solutions of these equations (often done numerically) determine estimators $\hat{\theta}$. Their covariance matrix is given by

$$(\widehat{V}^{-1})_{ij} = - \left. \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right|_{\widehat{\theta}}$$

Maximum likelihood method has a nice property that (in the limit of infinite statistics) it produces unbiased estimators with smallest possible variance. But beware of small statistics samples ! ML fits are implemented in many statistical packages (ROOT, Python, MATLAB). Can be applied to binned or unbinned data

Error Intervals From Likelihood Ratio



Least Squares Estimators

For a set of Gaussian-distributed variables y_i, define:

$$\chi^2(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + \text{ constant } = \sum_{i=1}^{N} \frac{(y_i - F(x_i;\boldsymbol{\theta}))^2}{\sigma_i^2}$$

Estimators:

$$L(\theta) \to \max; \Rightarrow \chi^2(\theta) \to \min$$

In particular, if the function F is linear in parameters θ, LS estimators are found by solving a system of linear equations analytically:

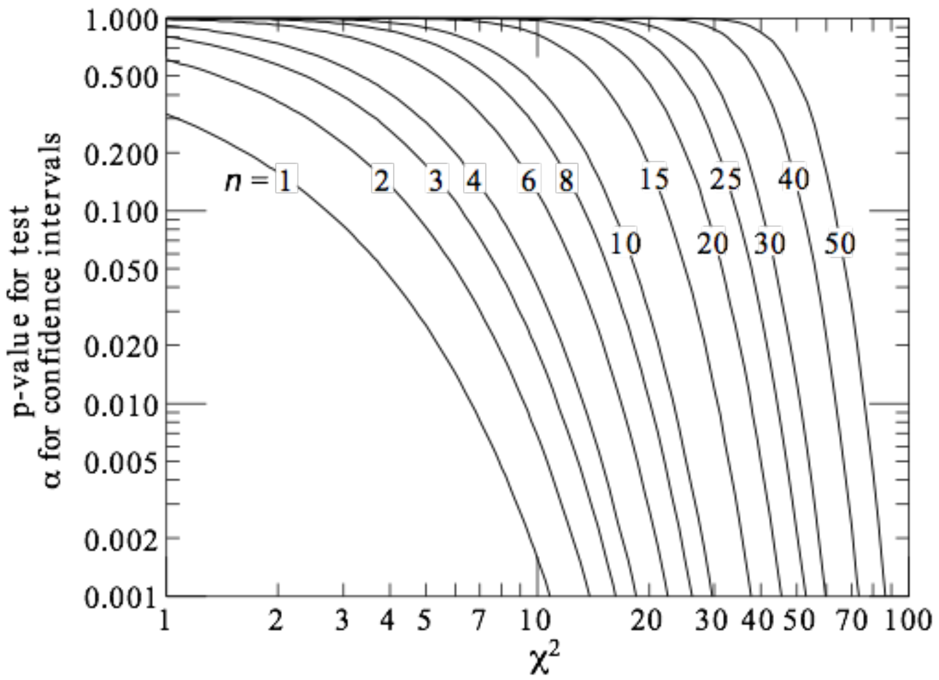
$$F(x_i;\boldsymbol{\theta}) = \sum_{j=1}^{m} \theta_j h_j(x_i) \longrightarrow \widehat{\boldsymbol{\theta}} = (H^T V^{-1} H)^{-1} H^T V^{-1} \boldsymbol{y} \equiv D\boldsymbol{y}$$

Least-squares fits are typically done on binned data, and implemented in most statistical packages (SciPy, ROOT, MATLAB, even Excel)

Example: chi-squared p-values

One advantage of a χ^2 fit is that the value of the minimum χ^2 can be interpreted as a measure of goodness-of-fit, iff errors on each data point are known, and the "noise" (distribution of data around their expected values) are Gaussian

In the plot below, n= number of degrees of freedom = $N_{\text{data points}}$ - $N_{\text{parameters}}$

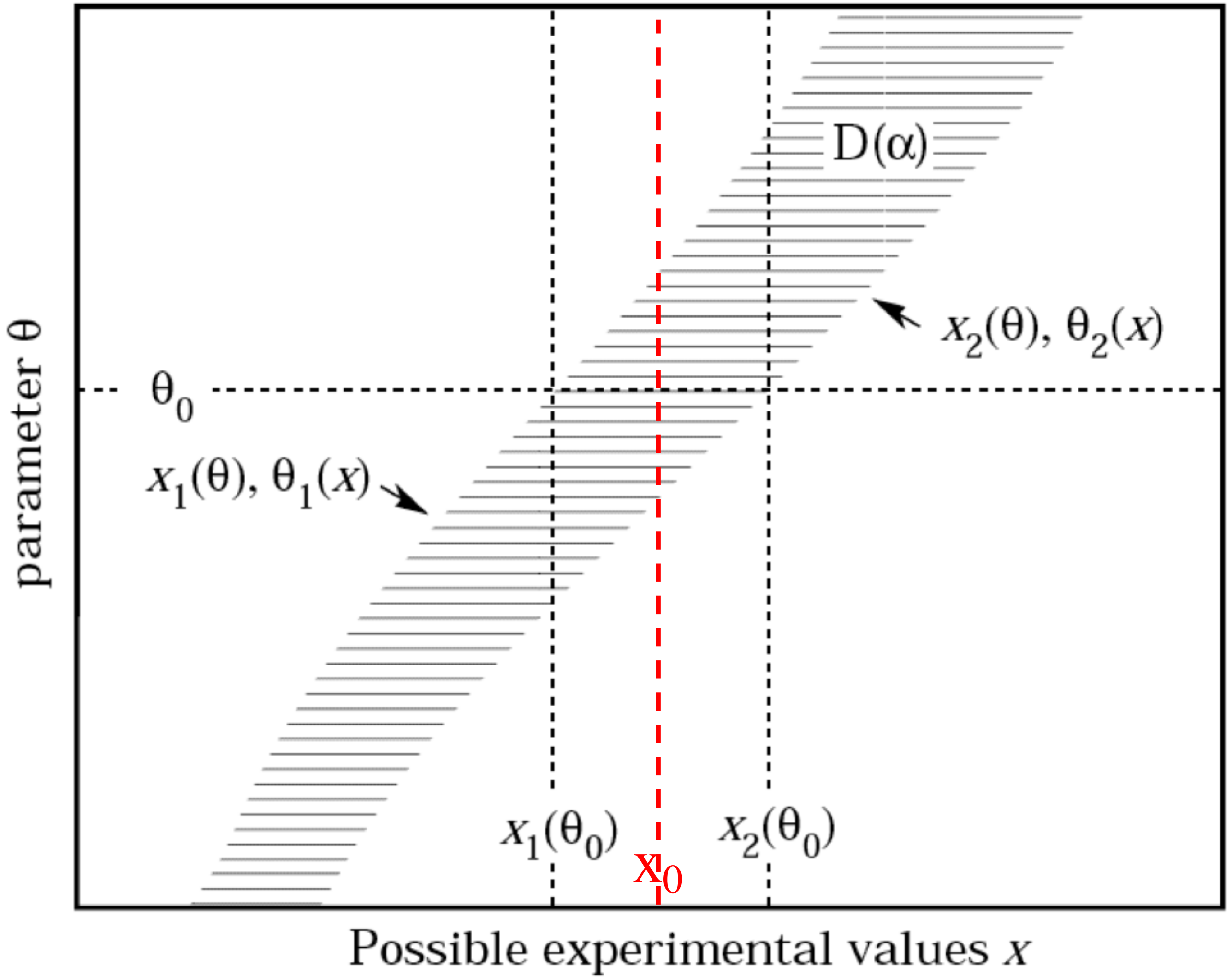


For a "good fit", expect χ^2 to be close to number of degrees of freedom = $N_{\text{data points}}$ - $N_{\text{parameters}}$

Confidence Limits

- Frequentist approach: confidence belts
  - ☞ Define

$$P(x_1 < x < x_2; \theta) = 1 - \alpha = \int_{x_1}^{x_2} f(x;\theta)\, dx$$



Caveats: interval not unique. Problems near a physical boundary. Use central intervals (equal area on both sides) or decide based on likelihood ratio (e.g. Feldman-Cousins)

Bayesian Approach

- Likelihood function + prior -> posterior for parameter

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int L(\boldsymbol{x}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\,d\boldsymbol{\theta}'}$$

- Treat as PDF and integrate

$$1 - \alpha = \int_{\theta_{\mathrm{lo}}}^{\theta_{\mathrm{up}}} p(\theta|\boldsymbol{x})\,d\theta$$

- Caveat: choice of prior

Example

Ben Hooberman's thesis (UC Berkeley Ph.D. 2009)

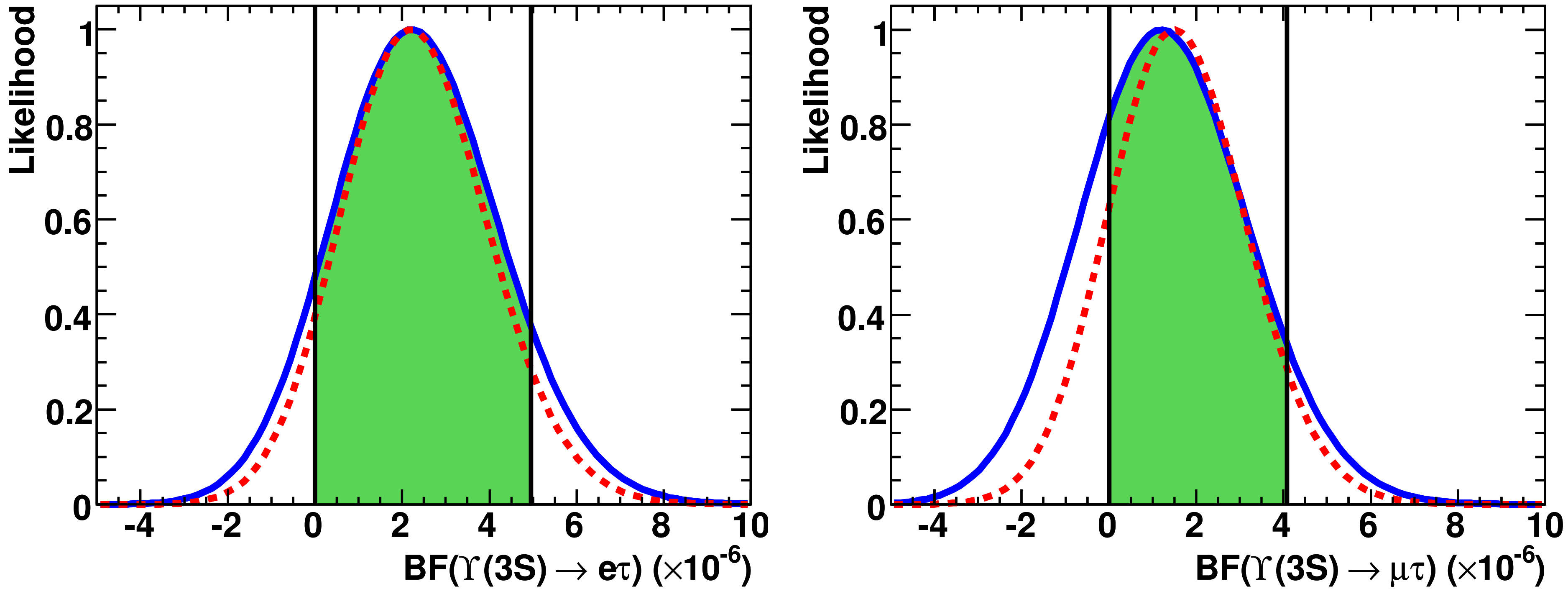
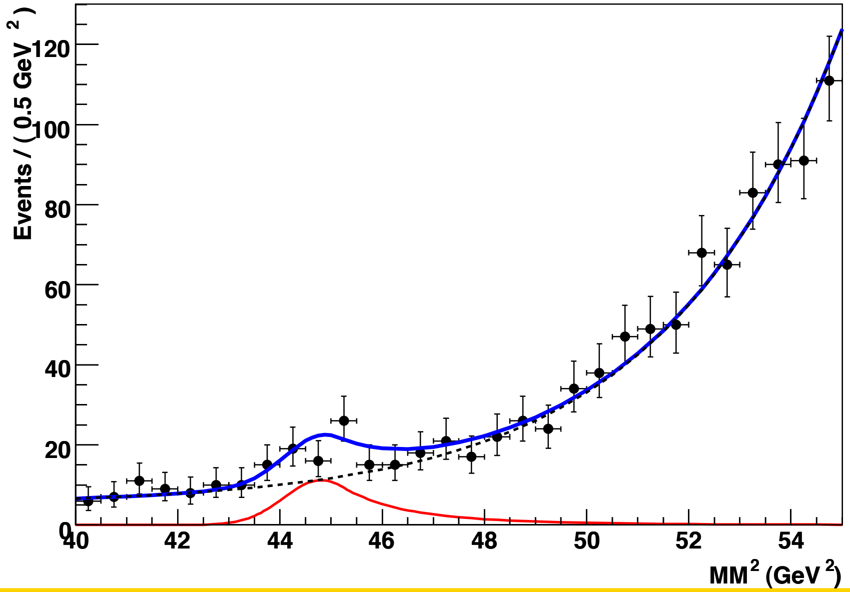


Figure 2.33: Likelihood as a function of the branching fractions $BF(\Upsilon(3S) \rightarrow e\tau)$ (left) and $BF(\Upsilon(3S) \rightarrow \mu\tau)$ (right) [60]. The dotted red curve includes statistical uncertainties only, the solid blue curve includes systematic uncertainties as well. The shaded green regions bounded by the vertical lines indicate 90% of the area under the physical ($BF > 0$) regions of the likelihood curves.

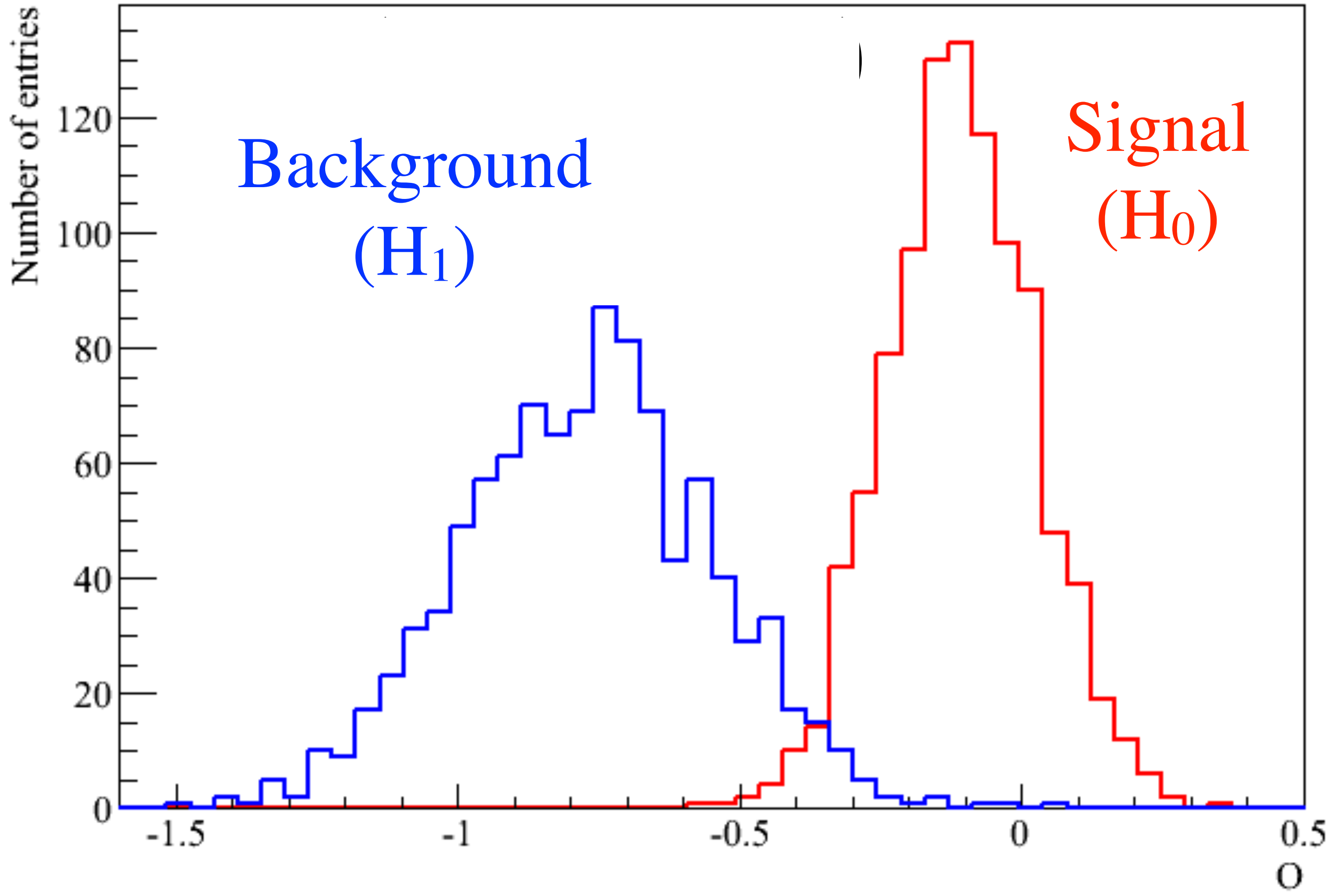
Hypothesis Testing

- Setting a confidence interval is a special case of a general problem of hypothesis testing
  - E.g. hypothesis is that *x* is within this interval
  - Or *x* belongs to a distribution
  - Hypothesis testing is a procedure for assigning a significance (confidence) level to a test
    - Generally involves computing quintiles of a distribution

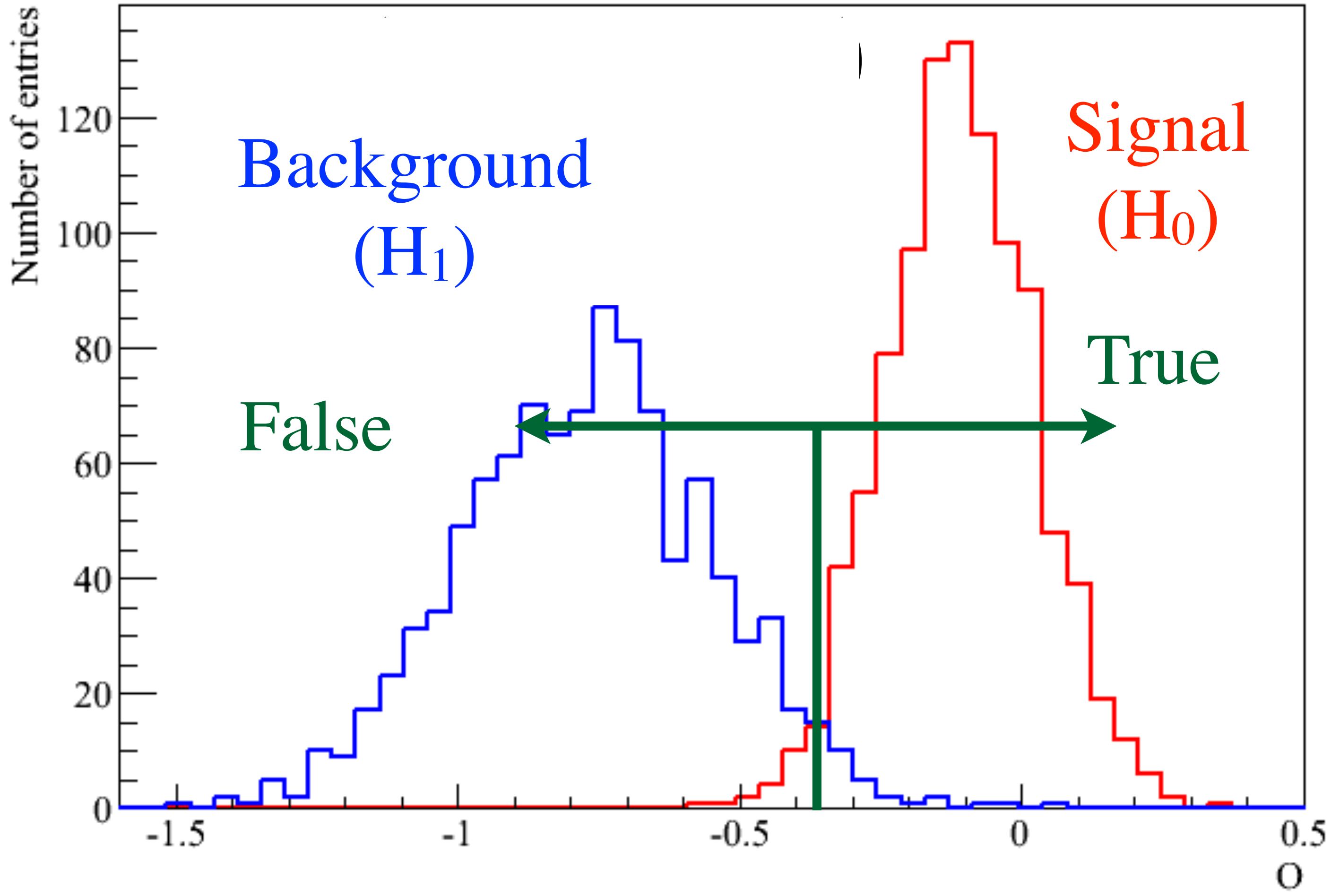


Luck of the Draw

Example

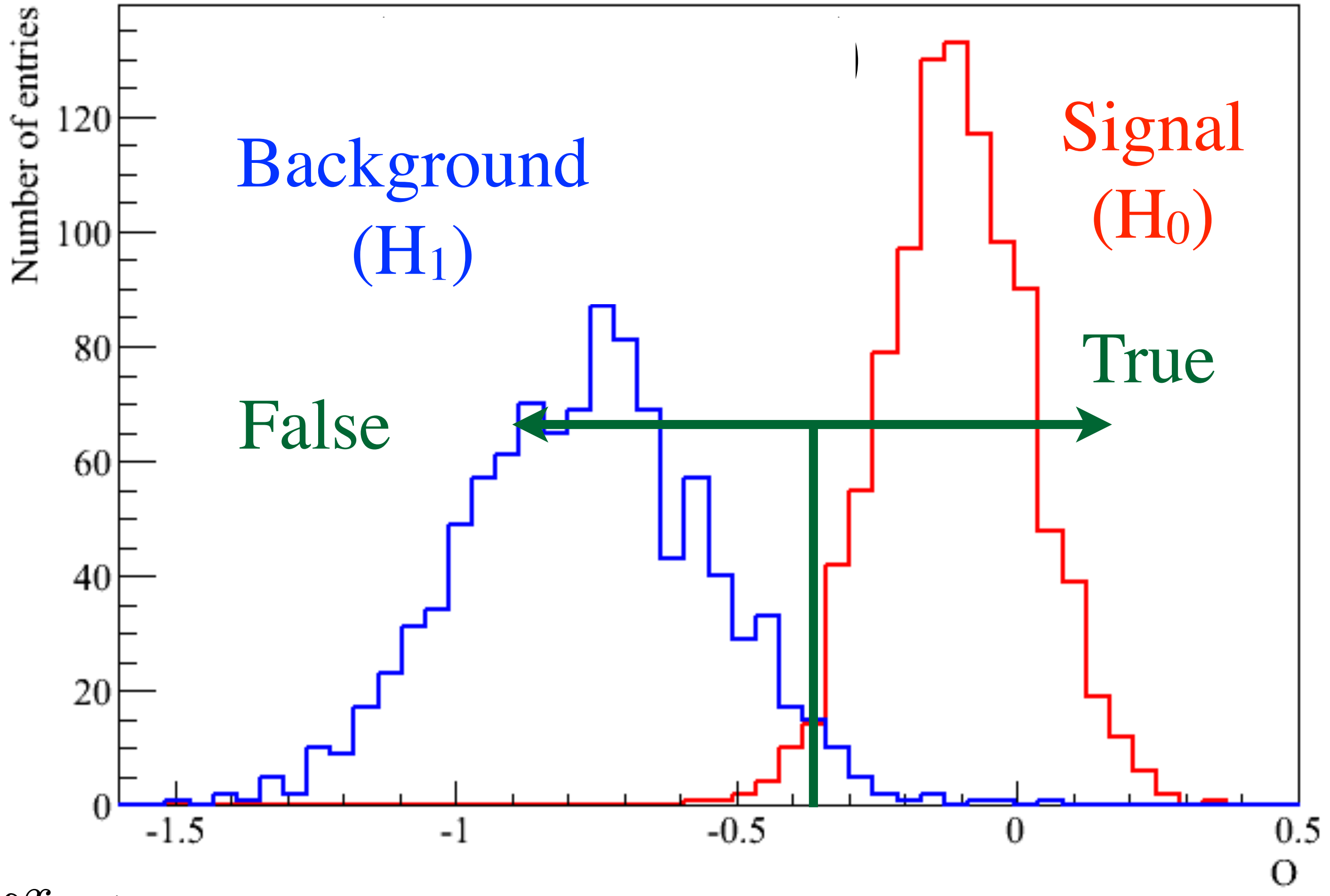

Number of entries
0
20
40
60
80
100
120
-1.5
-1
-0.5
0
0.5
O
Background
(H1)
Signal
(H0)

Example



Number of entries
120
100
80
60
40
20
0
-1.5
-1
-0.5
0
0.5
O
Background
(H1)
Signal
(H0)
False
True

Example



$$\alpha = \int_{-\infty}^{x_{cut}} f(x|\text{signal})dx$$

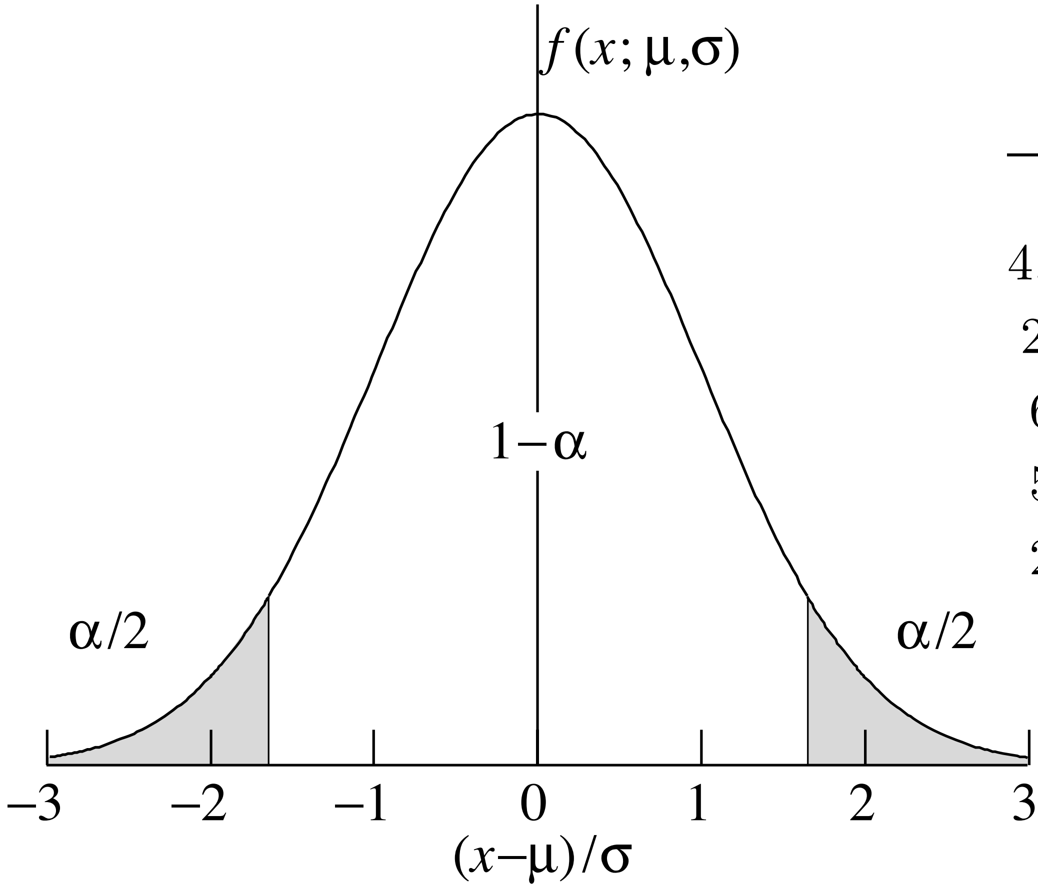
Type-I error (signal efficiency=1-α)

$$\beta = \int_{x_{cut}}^{+\infty} f(x|\text{bkg})dx$$

Type-II error (bkg misID=β)

Example: Gaussian distribution

$$1-\alpha = \frac{1}{\sqrt{2\pi}\sigma}\int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2}\,dx = \mathrm{erf}\left(\frac{\delta}{\sqrt{2}\,\sigma}\right)$$



α	δ
0.3173	1σ
4.55×10^{-2}	2σ
2.7×10^{-3}	3σ
6.3×10^{-5}	4σ
5.7×10^{-7}	5σ
2.0×10^{-9}	6σ

α	δ
0.2	1.28σ
0.1	1.64σ
0.05	1.96σ
0.01	2.58σ
0.001	3.29σ
10^{-4}	3.89σ

Neyman-Pearson Lemma

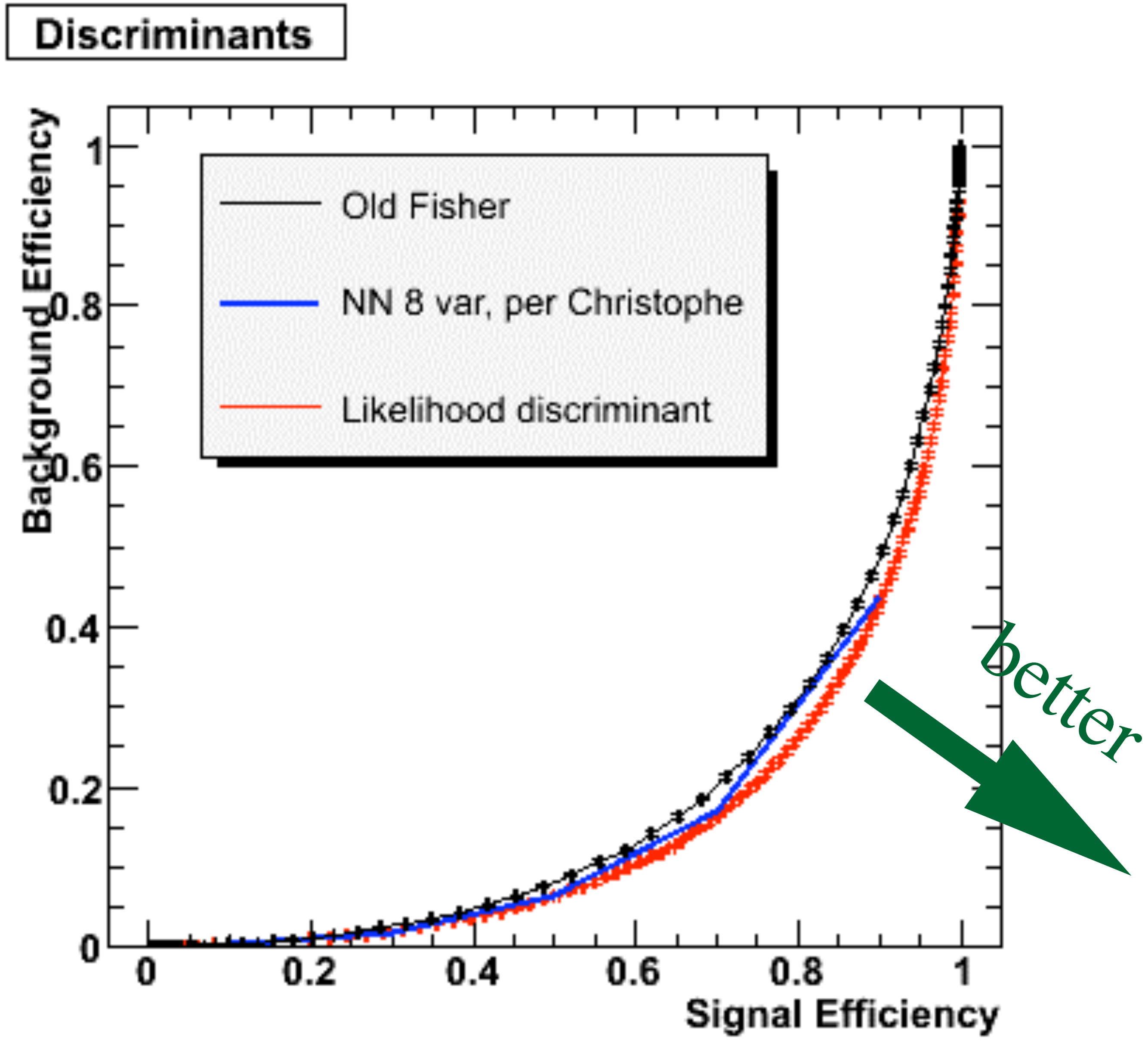
- Want to choose a cut such that α & β are as small as possible *at the same time*
  - Or maximize efficiency and purity:
    - ☞ ε=1-α→ max
    - ☞ β→min so

$$p = \frac{\epsilon_{\rm sig} N_{\rm sig}}{\epsilon_{\rm bkg} N_{\rm bkg} + \epsilon_{\rm sig} N_{\rm sig}} \to \max$$

- Neyman-Pearson Lemma:
  - Acceptance region giving the best rejection power (smallest β) for a given α is defined by the region

$$t = \frac{f(x|H_0)}{f(x|H_1)} > C(\alpha)$$

Example



Discriminants
Background Efficiency
Signal Efficiency
0
0.2
0.4
0.6
0.8
1
Old Fisher
NN 8 var, per Christophe
Likelihood discriminant
better

Single-var vs Multi-var Discriminants

- For a single variable, there is a 1-to-1 transformation between x_{cut} and α, and therefore t and x_{cut}
- Not so obvious for a multiple discriminating variables
  - N-P lemma says likelihood ratio is in theory the best discriminating variable
    - ☞ Assuming likelihood ratio is computed correctly (e.g. with correlations)
  - In practice, other techniques are computationally easier to implement
    - ☞ Machine learning !
    - ☞ Fisher, Neural networks, Boosted Decision Trees, etc
    - ☞ More to come

Goodness of Fit

- Standard problem: does fit agree with data ?
  - H_0: data belong to a given distribution
- Chi-squared test

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

→ $N_{dof} = N_{Points} - N_{parameters}$

(for good fit)

- Or, for a correlated set of points

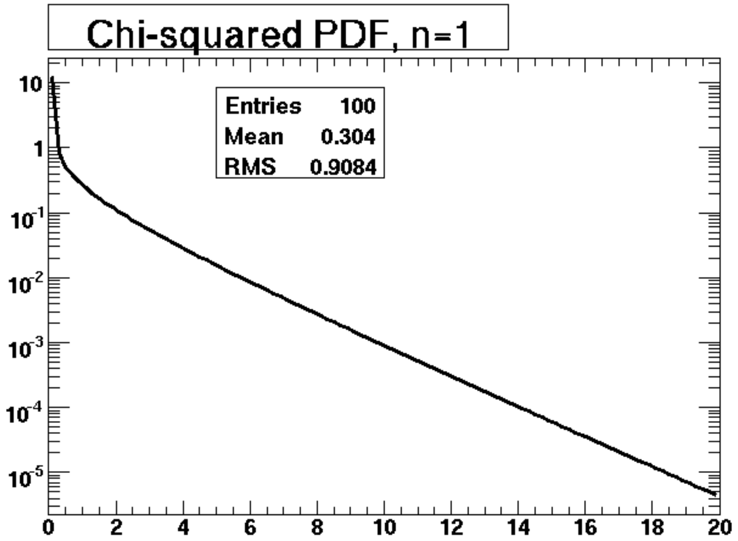
$$\chi^2 = (\vec{y} - \vec{f})^T V^{-1} (\vec{y} - \vec{f})$$

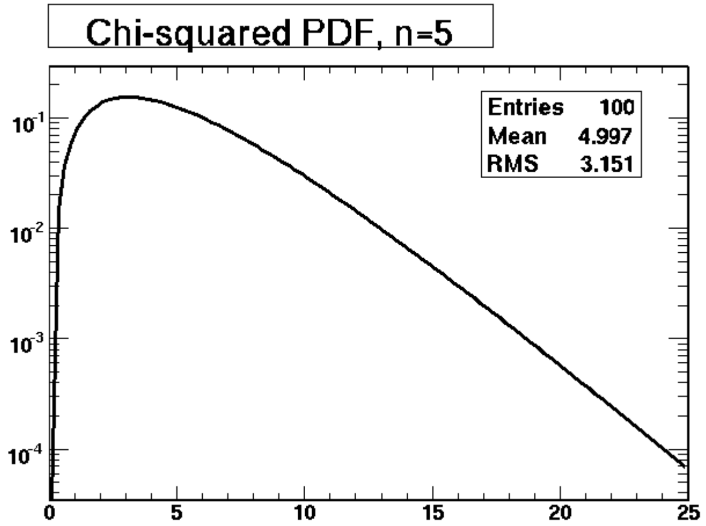
where

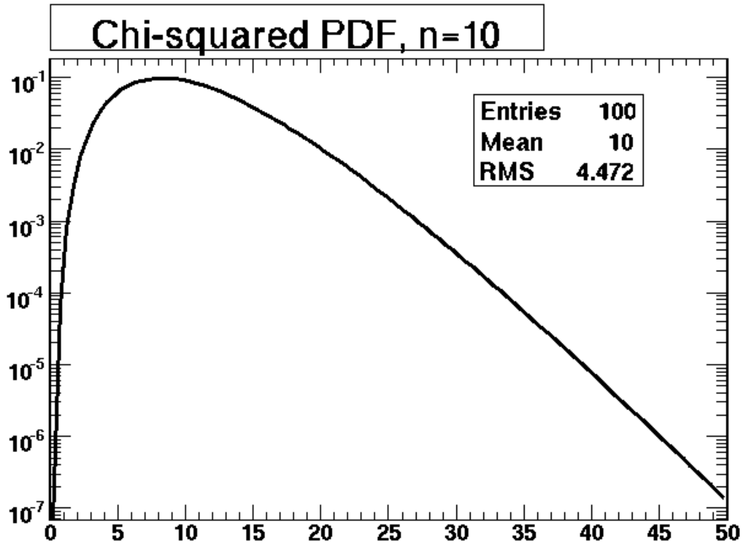
$$V_{ij} = \langle (y-f)_i (y-f)_j \rangle$$ (covariance matrix)

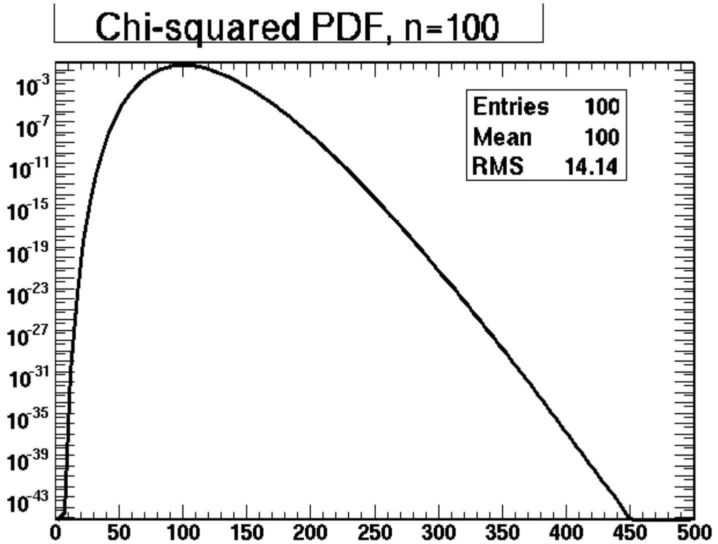
Chi-squared Distribution

$$f(z;n) = \frac{z^{n/2-1}e^{-z/2}}{2^{n/2}\Gamma(n/2)} \;; \qquad z \geq 0$$

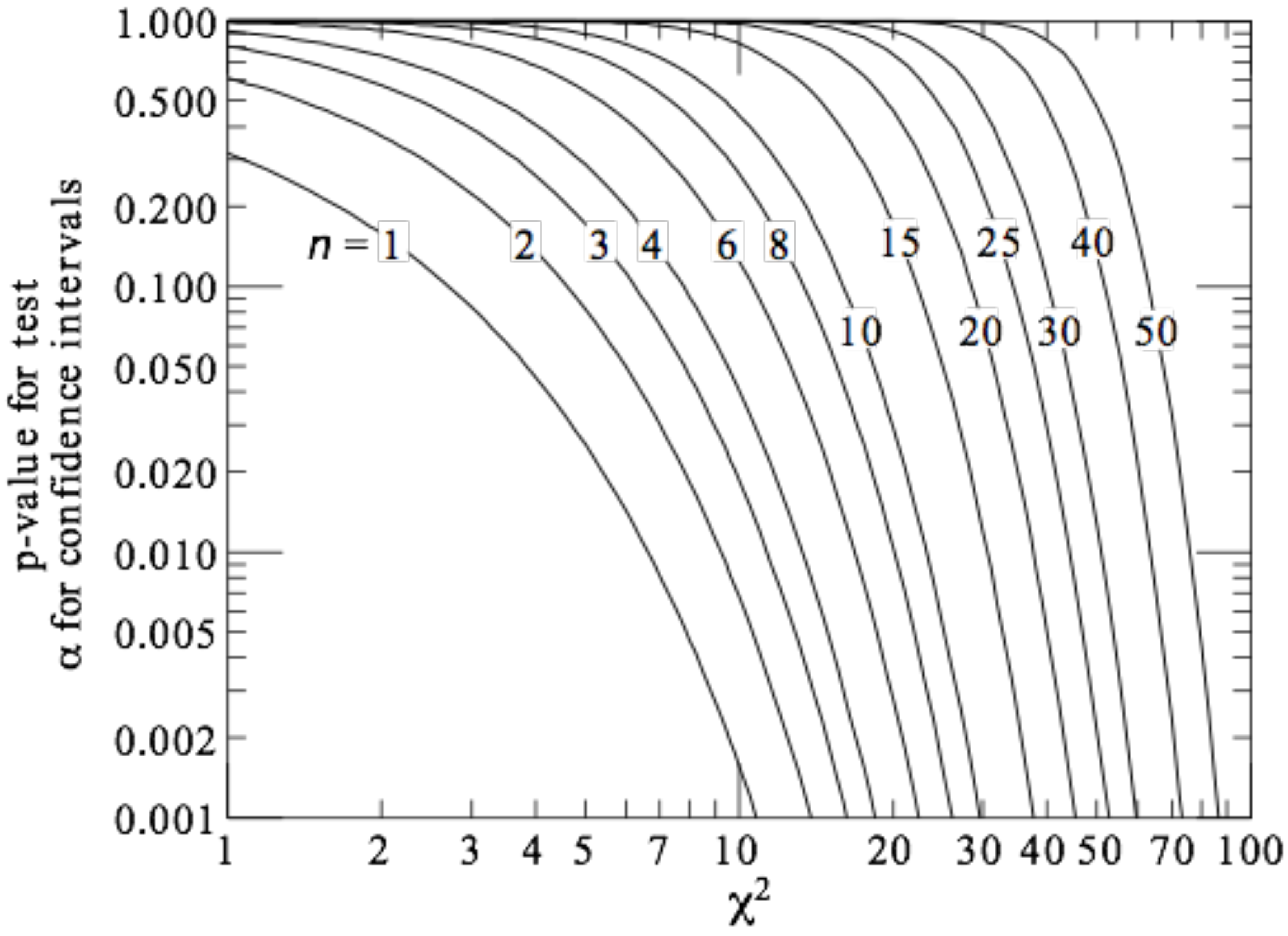




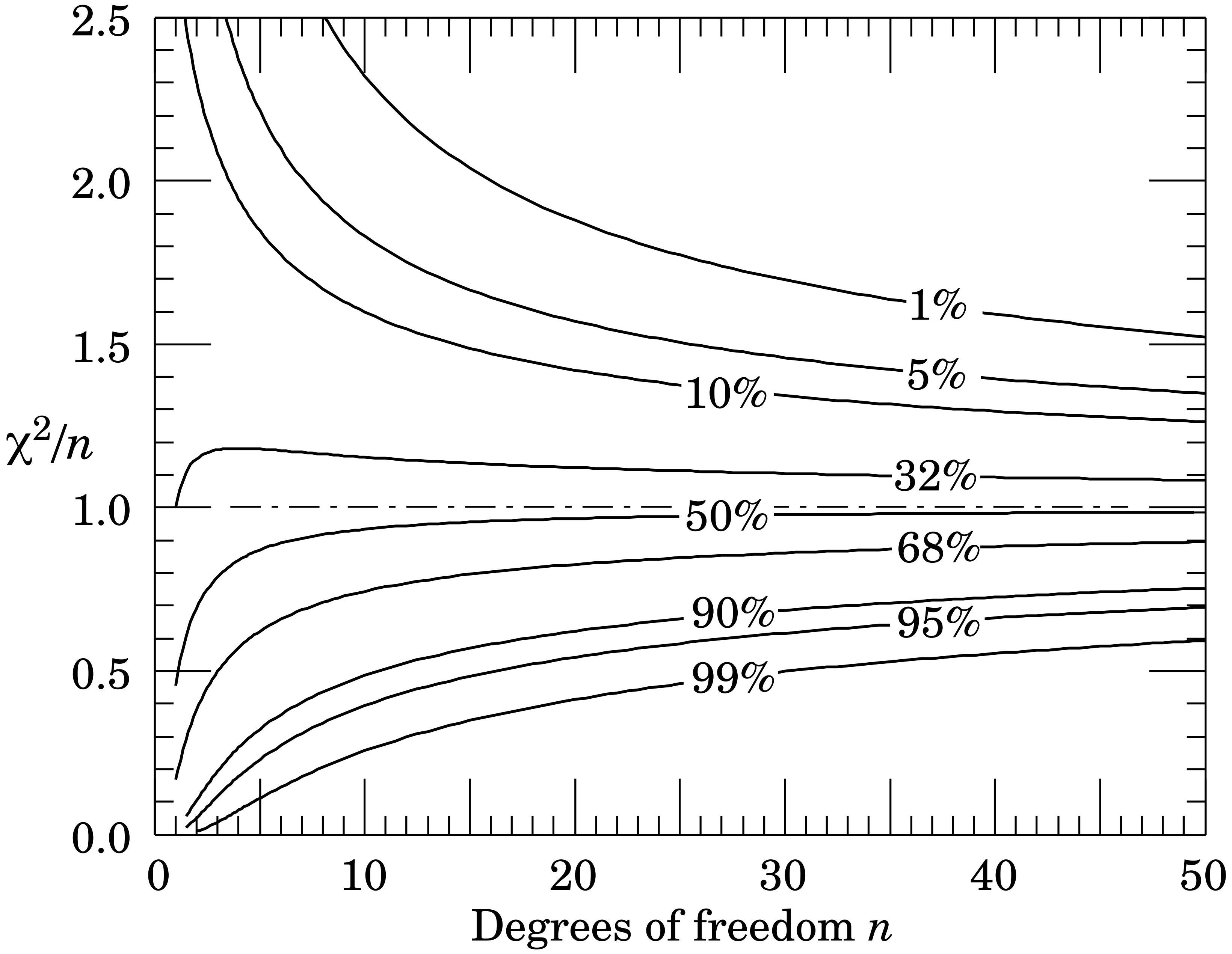




Example: chi-squared p-values



Chi-squared p-values



Kolmogorov-Smirnov Test

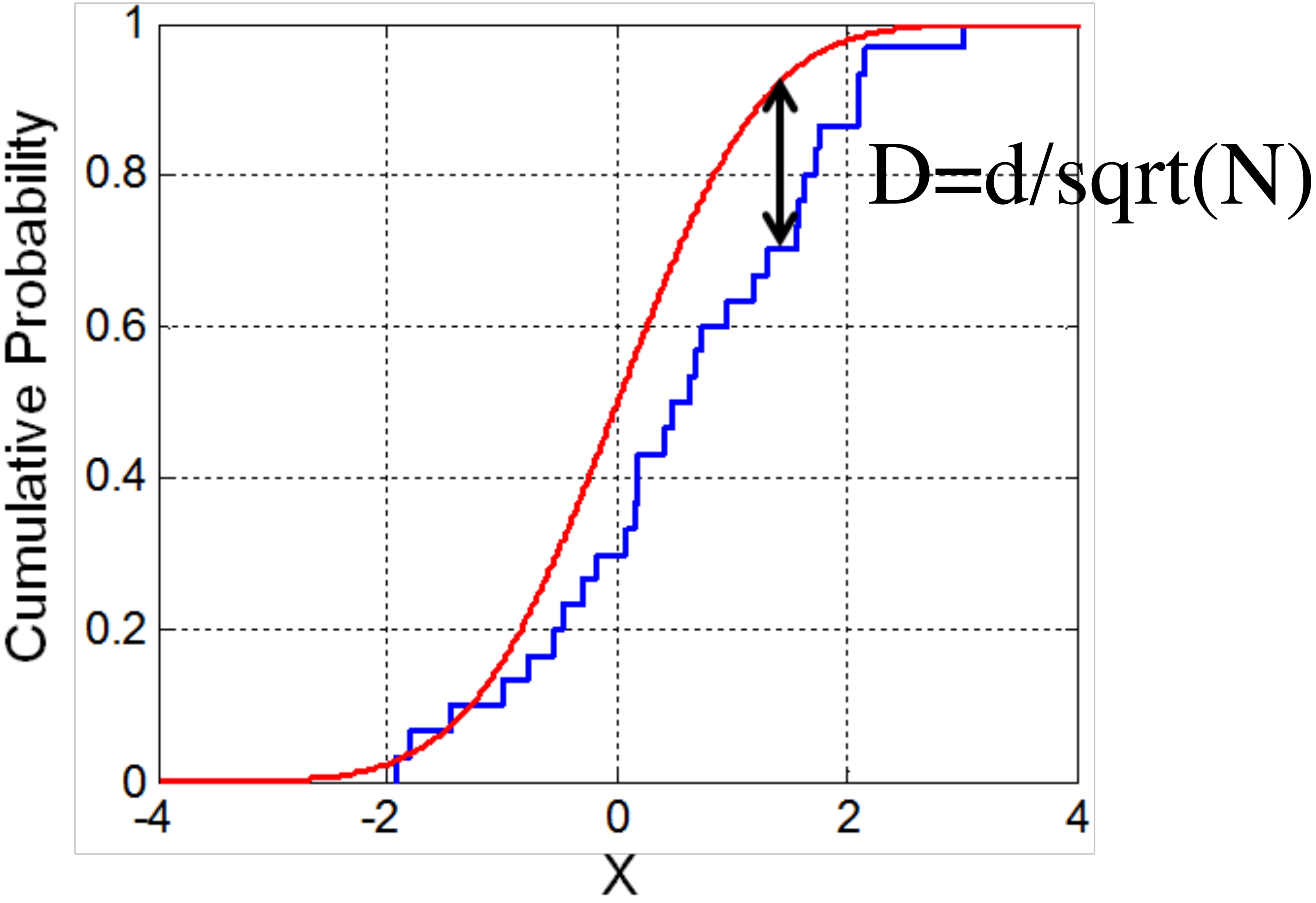
- Useful for small number of events to avoid binning
  - χ^2 only valid in Gaussian limit → many events/bin
- Form a cumulative distribution $\Sigma(\{x\})$ for each event in $\{x\}$
- Overlay CDF $F(x)$ computed from PDF $f(x)$
- Compute max deviation

$$d \equiv \max |\Sigma(x) - F(x)| \sqrt{N}$$

Test: $d > c(\alpha) \rightarrow$ reject H_0

α	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

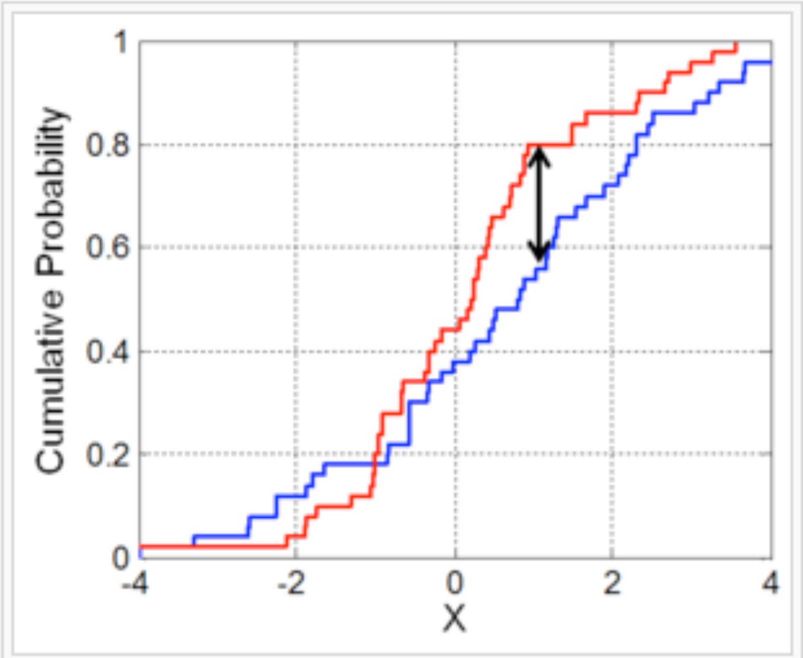
K-S Test



K-S Test with 2 Samples

- Can compare two CDF computed from two independent samples, without prior knowledge of an underlying CDF

$$d \equiv \max \left| \Sigma(x_1) - \Sigma(x_2) \right| \sqrt{\frac{N_1 N_2}{N_1 + N_2}}$$



Standard Problem

- We see a small peak on top of a background, and want to determine if we have made a discovery
  - ❑ Need to evaluate *significance* of observation
- Standard recipe: evaluate likelihood ratio of two hypotheses
  - ❑ (a) signal is present on top of background
  - ❑ (b) signal is absent
    - ☞ In other words, we want to know how likely it is for background B to fluctuate to observed value S+B
    - ☞ Practically, it means computing max likelihood (for S+B) and likelihood for S=0

Caveats

- Often report answer in terms of "gaussian sigmas":

$$\mathcal{S} = \sqrt{2(\log \mathcal{L}_{\max} - \log \mathcal{L}_0)}$$

- But have to confirm (with toy MC) that this significance truly corresponds to gaussian p-value

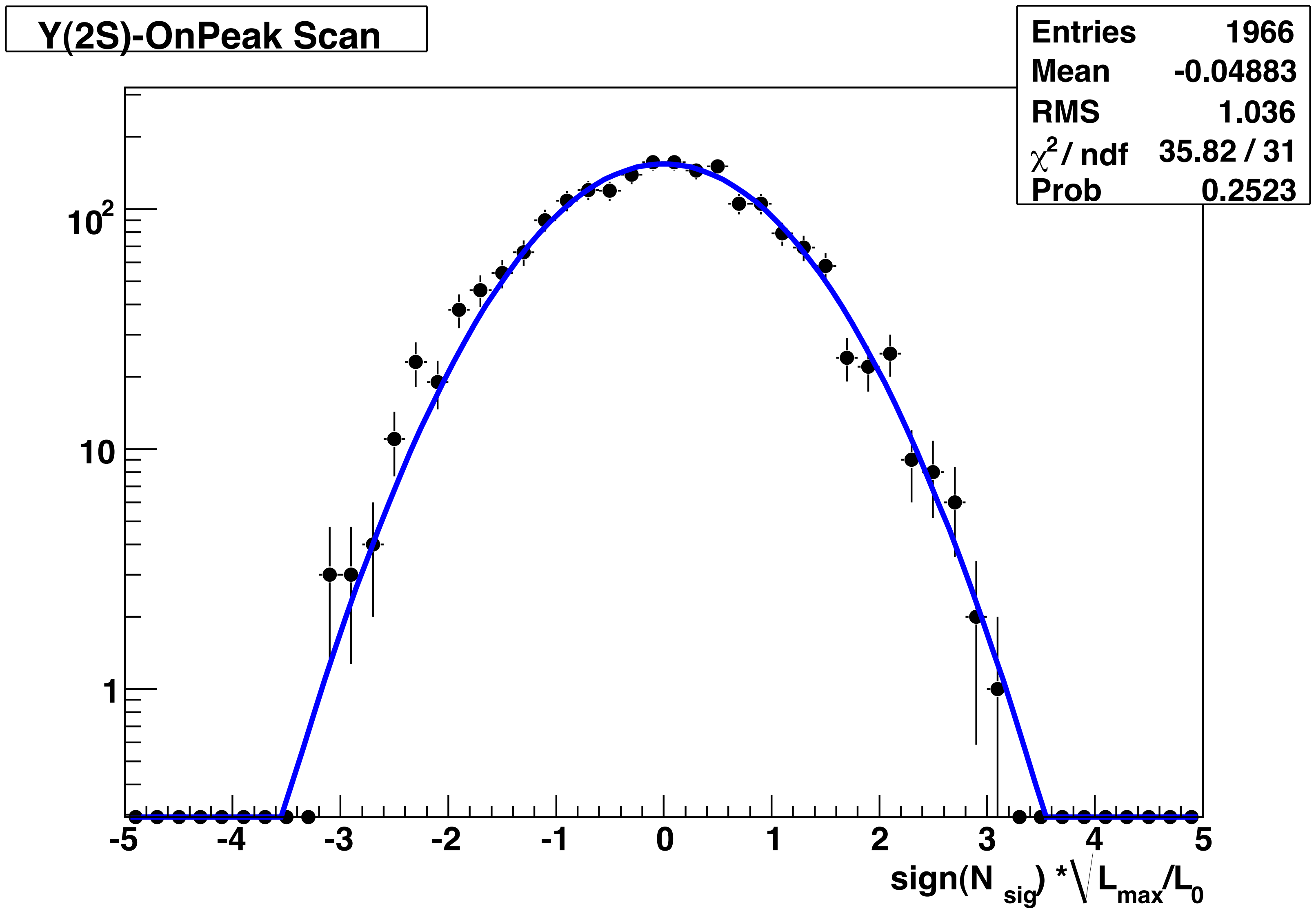
  ☞ Toy MC

- Another important issue: trial factor, or "look elsewhere" effect

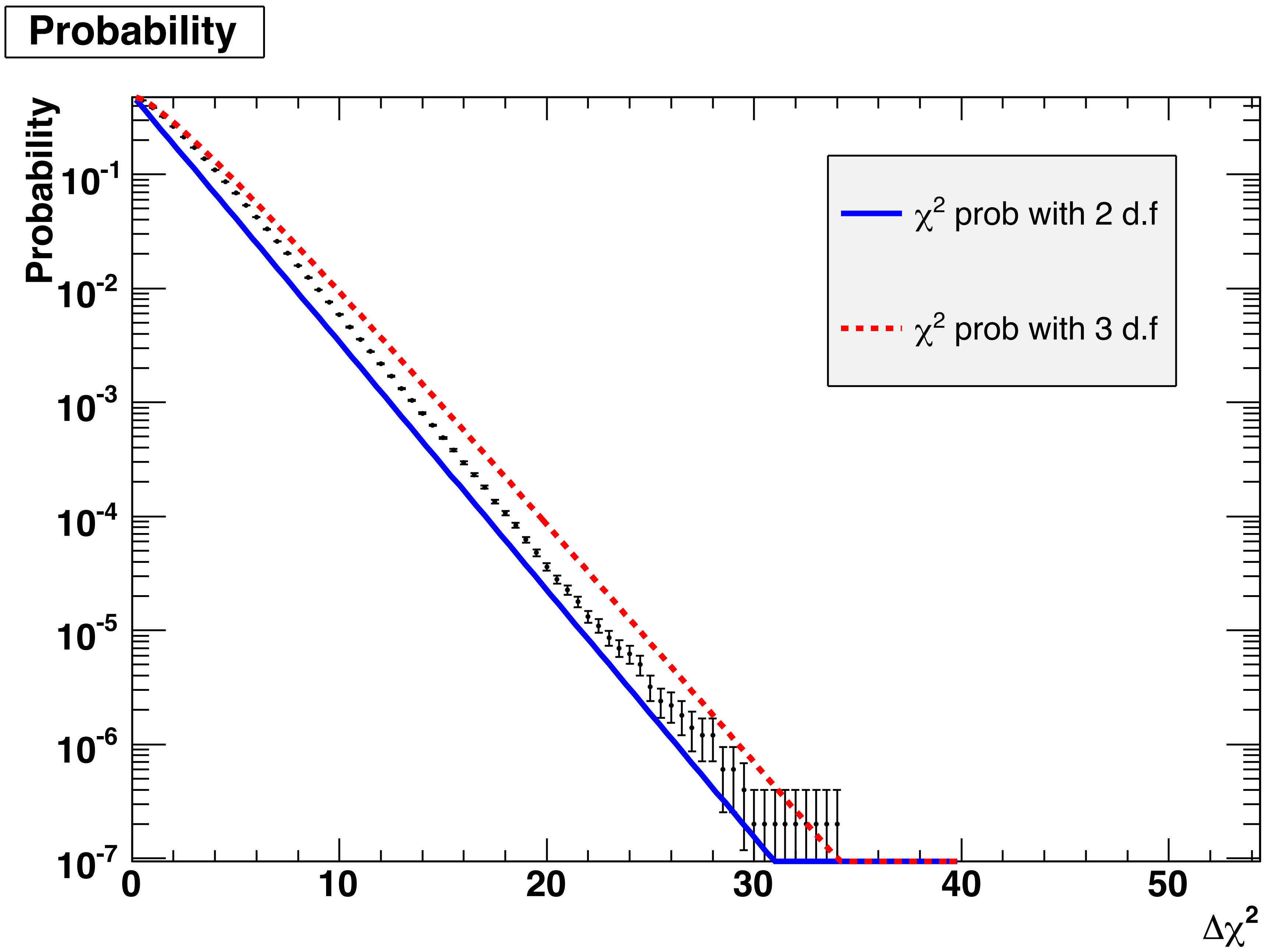
Trial Factors

- If we do not know a-priori where the signal is, significance of any peak is diluted by the number of *independent* windows we opened
  - Compute probability to observe a given fluctuation *anywhere* in the dataset
    - ☞ Naively, multiply the p-value by the number of independent trials
    - ☞ Better yet, estimate probability with toy Monte Carlo

Example: Search for Peak with Unknown Mean

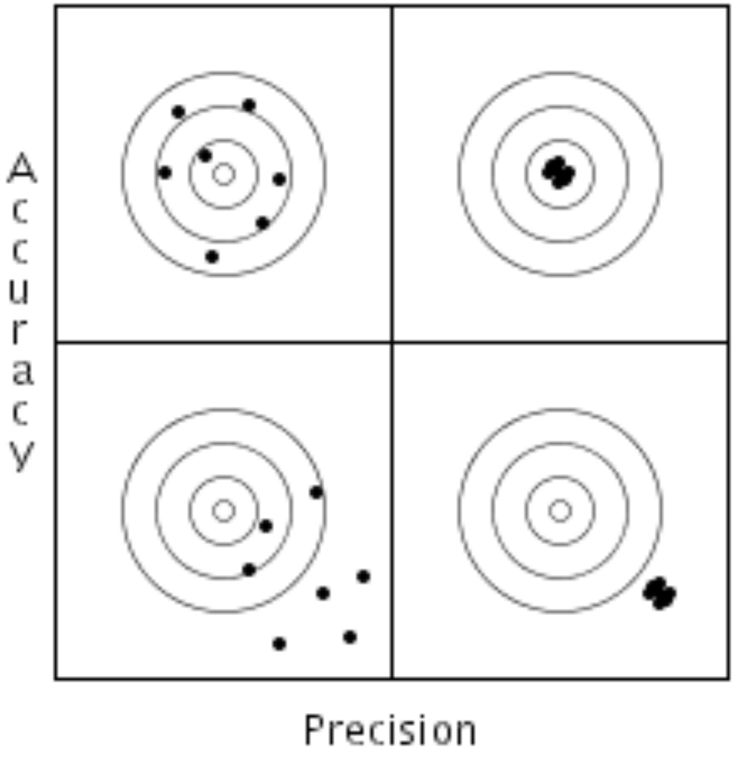


Example: Search for Peak with Unknown Mean



Systematics: “Another Class of Errors”

- Statistical errors:
  - Spread in values one would see if the experiment were repeated multiple times
    - ☞ RMS of the estimator for an ensemble of experiments done under the same conditions (e.g. numbers of events)
- But there is another source of uncertainty in results: systematics

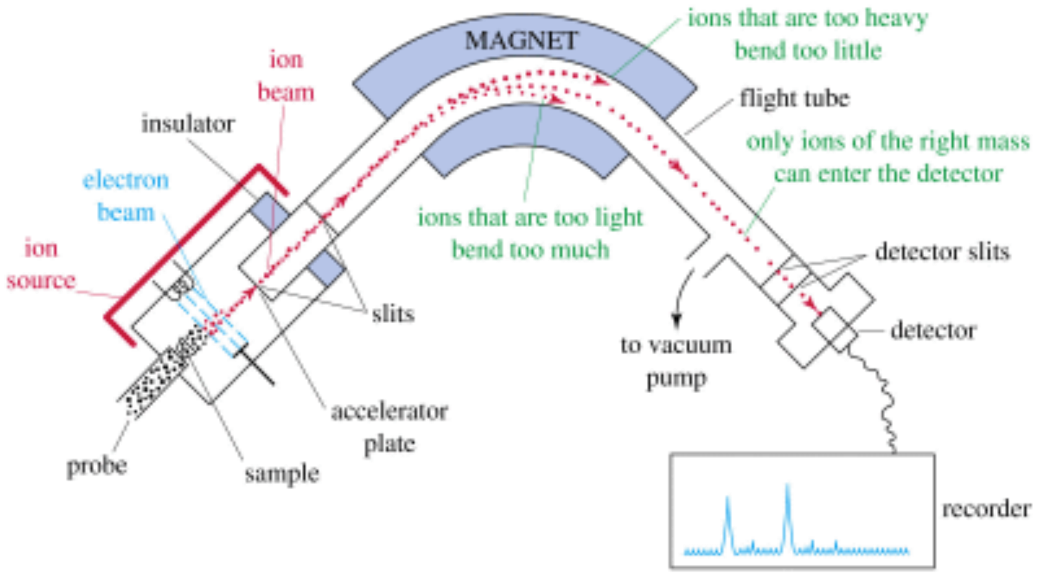


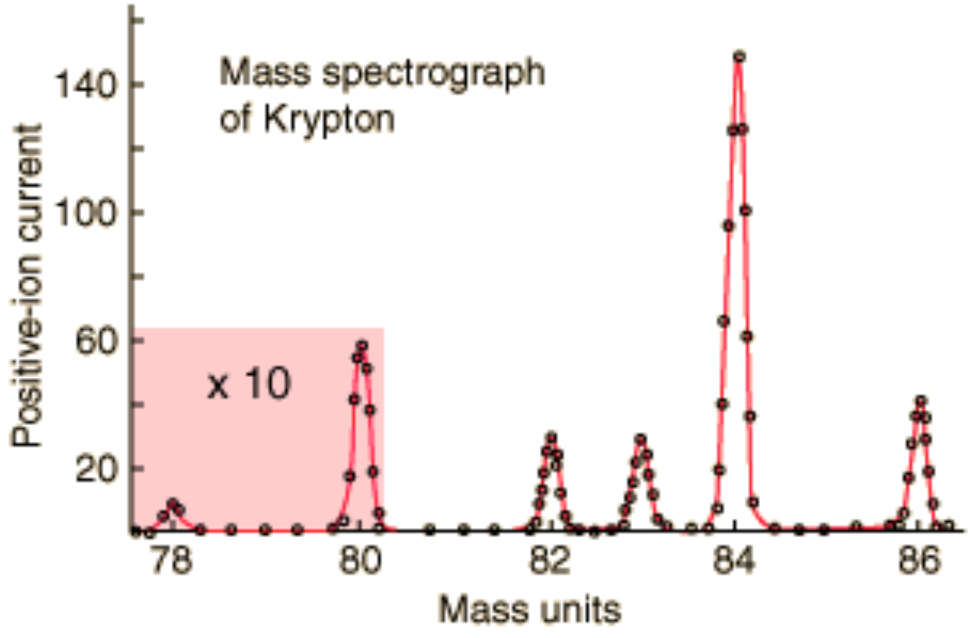
Simple Example

- Mass spectrometer

$$m = \frac{qrB^2}{2V}$$

- Stat error: resolution/sqrt(N)
  - ☞ Measure V,B for each run
  - ☞ Average fluctuations
- Common errors do not average out
  - ☞ Scale of B,V
  - ☞ Radius r
  - ☞ Velocity selection
  - ☞ Energy loss (residual pressure)
  - ☞ Etc, etc.





Combination of Errors

- Normally, independent errors are added in quadrature
  - For instance, if measurements of *r,V,B* are uncorrelated, then (to first order)

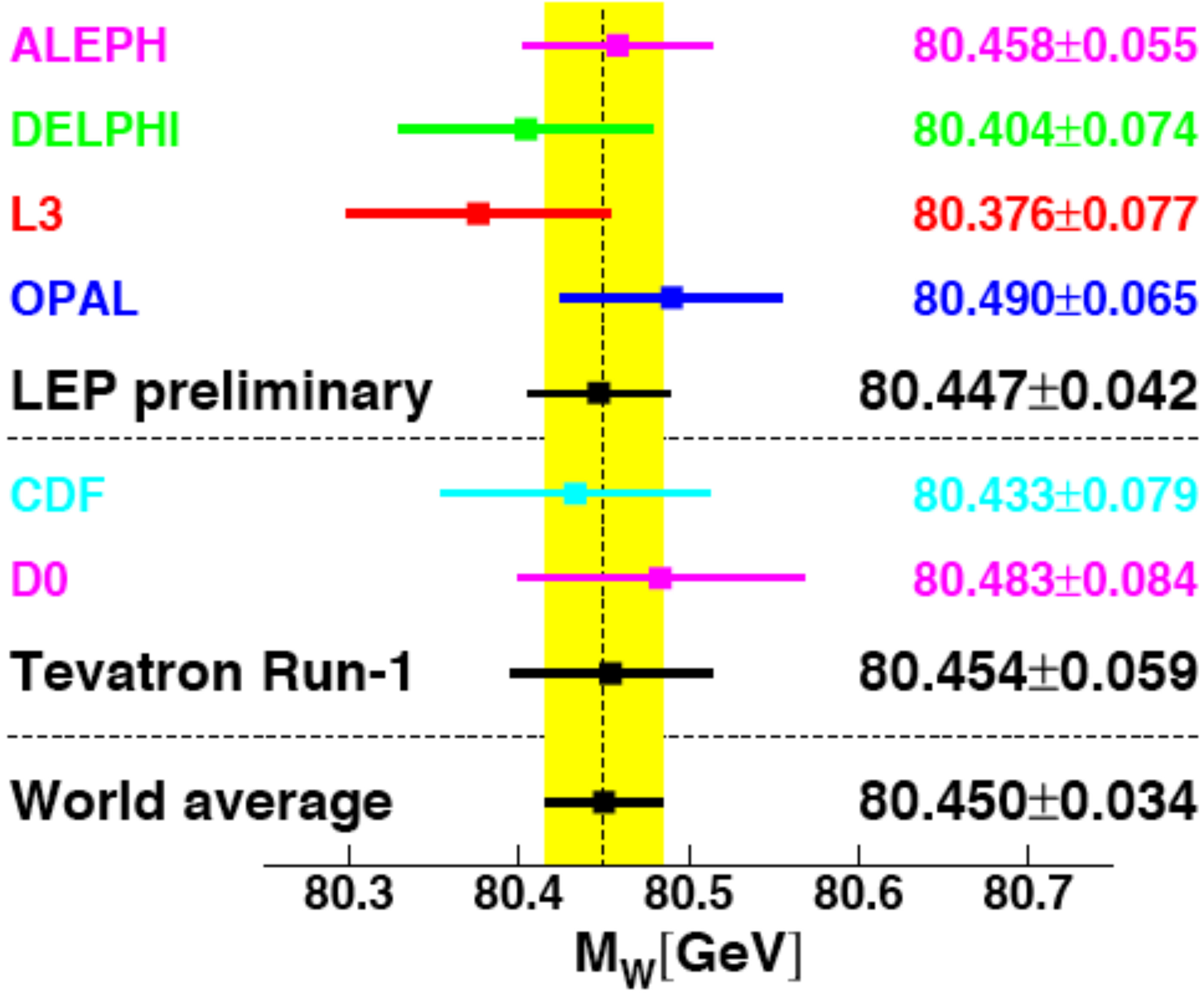
$$\frac{\sigma(m)}{m} = \sqrt{\left(\frac{\sigma(r)}{r}\right)^2 + \left(\frac{\sigma(V)}{V}\right)^2 + \left(2\frac{\sigma(B)}{B}\right)^2}$$

  - This is fine for a single ion
    - ☞ But when we average (take more data), have to take into account the fact that errors on *r,V,B* correlate measurements of mass for each ion

Quadrature Sum

- Stat and syst errors are typically quoted separately in experimental papers
  - E.g. c=[0.9 ± 0.2 (stat.) ± 0.1 (syst.)] ft/nsec
  - It is understood that the first number scales with the number of events while the second may not
    - ☞ Splitting like this gives a feeling of how much a measurement could be improved with more data
    - ☞ It is also understood that stat and syst errors are uncorrelated (if this is not the case, have to say so explicitly !)
    - ☞ It is also understood that stat errors are uncorrelated between different experiments, while syst errors could be correlated (modeling, bias)

Classic Example (one of many)



Combining Errors

- For one measurement with stat and syst errors, this is easy
  - Suppose we measure $x_1 = \langle x_1 \rangle \pm \sigma_1 \pm S$
    - ☞ Split into “random” and “systematic” parts
    - ☞ $x_1 = \langle x_1 \rangle + x^R + x^S$
    - ☞ $\langle x^R \rangle = \langle x^S \rangle = 0$, $\langle (x^R)^2 \rangle = \sigma_1$, $\langle (x^S)^2 \rangle = S$
    - ☞ Total variance $V[x_1] = \langle x_1^2 \rangle - \langle x_1 \rangle^2 = \langle (x^R + x^S)^2 \rangle = \sigma_1^2 + S^2$
    - ☞ Syst and stat errors are combined in quadrature

Systematic Errors and Fitting

- Use covariance matrix in χ^2:

$$\chi^2 = \sum_i \sum_j d_i V_{ij}^{-1} d_j$$

  - ☞ $d_i=(y_i-y_i^{\text{fit}})$
  - ☞ Can apply the same recipe for ML fit (e.g. $L \sim \exp(-\chi^2/2)$)

Practical Implications

- In the full formalism, can still use χ^2/df test to determine the goodness of fit
  - ☞ But this will not work unless correlations are taken into account
  - ☞ For simplicity, if all stat errors are roughly equal and all systematic errors are common, can do the fit with stat errors only (this will determine stat errors on parameters), then propagate syst errors
- Limitations
  - ❑ More points do not improve the systematic error
  - ❑ Goodness of fit would not reveal unsuspected sources of systematics
    - ☞ All points move together -- same goodness of fit