

A General Theory of Goodness of Fit in Likelihood fits

Rajendran Raja
Fermilab

Feb 10, 2006, FNAL Wine and Cheese

Started working on problem after Durham '02 conference
SLAC PHYSTAT03 Gave a solution to the unbinned
goodness of fit problem

Oxford-Phystat2005-Generalized this to Binned and
Unbinned goodness of fit as a general theory

Calculation of Errors- Once the GoF problem is solved, one
can use Bayes' theorem to calculate posterior densities
without Bayesian Priors. Mathematical Proof.

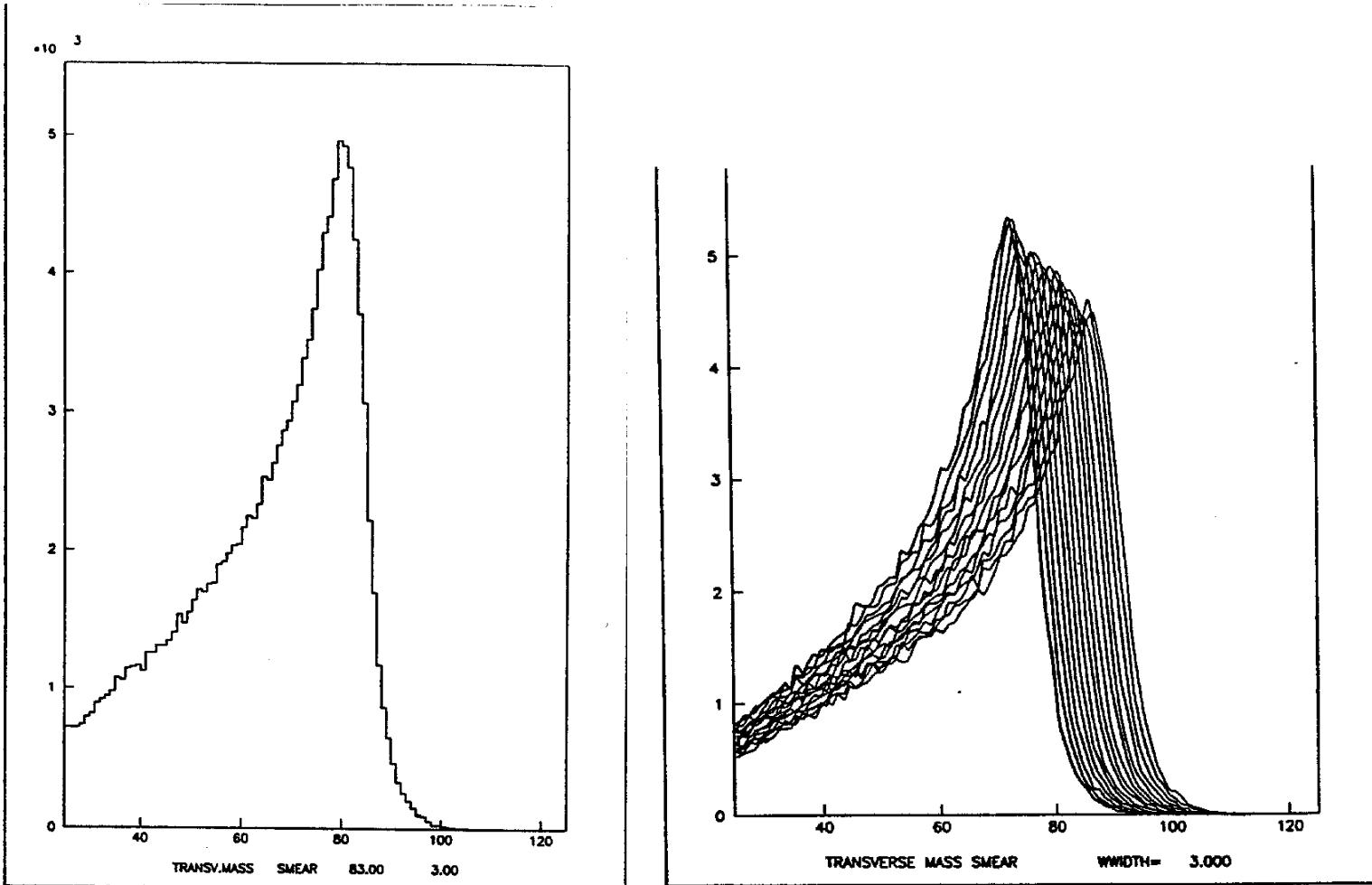
Can provide Fortran Subroutines do perform unbinned
goodness of fit, if interest.

See physics/0509008 for further details-60 page writeup

Overview of Goodness of fit techniques

- The χ^2 technique introduced by Karl Pearson is widely used. It is used for binned fitting. Analytic theory fully developed. Error on fitted parameter is obtained in the Gaussian approximation by $\Delta\chi^2=1$. Probably the most commonly used measure of goodness of fit. Drawback—Does not handle bins with low statistics well.
- G^2 method (binned)-Known to Statisticians-in HEP- Baker and Cousins(83).
 - » $G^2(\text{Multinomial}) = \sum_{k=1}^{n_b} n_k \left(\log_e \frac{n_k}{T_k} \right)$
 - where n_k is the number observed in the k^{th} bin
 - T_k is the number theoretically expected in the k^{th} bin
 - n_b is the number of bins in the histogram.
 - » $G^2(\text{Poisson}) = \sum_{k=1}^{n_b} T_k - n_k + n_k \left(\log_e \frac{n_k}{T_k} \right)$
 - The theory normalization and experimental normalization may be different
 - » No analytic theory of distribution of G^2 . However, can show that for large statistics, asymptotes to χ^2 .
- For maximum likelihood fitting, binned and unbinned methods exist. The theory we will derive will bring all of the above under one framework.
- Kolmogorov-Smirnov, Aslan-Zech (energy scheme) are not likelihood techniques and will not be covered here.

An example where χ^2 fitting is inadequate- W transverse mass



Format of talk

- State Maximum likelihood principle.
- Show why Likelihood cannot act as a goodness of fit variable- Not invariant under transformation of variables
- Define Likelihood ratio of Theoretical likelihood/Data Likelihood as derived from data. Show this is invariant under transformation of variables.
- Apply theory for unbinned goodness of fit using PDE's
- Illustrative Example for 1d unbinned fits.
- Illustrative example for unbinned fitting for an extreme case.
- Derive Goodness of fit formulae for binned fitting
- Show these asymptote to a χ^2 like variable for large statistics.
- Show theory applies for χ^2 fitting.
- Calculation of errors- Problem of inverse probabilities-Derive Bayes' theorem- Describe Bayesian Paradigm
- Show that Goodness of fit with "Data Likelihood from Data" is incompatible with Bayesian Prior.
- Introduce New Paradigm in Statistics that permits inversion of probabilities without Bayesian priors-Uses Bayes' theorem and Goodness of fit.
- Illustrate ideas using examples.

Notation

s denotes signal. Can be multi-dimensional. Denotes theoretical PARAMETERS. Can be multidimensional.

c denotes configurations and signifies DATA. Can be multi-dimensional

$P(c|s)$ signifies the conditional probability density in c , given s . I.e It defines the theoretical model-theory pdf which obeys normalization condition.

$P(s|c)$ signifies the conditional probability density in s , given c .

$$\int P(c | s) dc = 1$$

Let \vec{c}_n denote the dataset $c_i, i=1,n$

Then

$$L \equiv P(\vec{c}_n | s) = \prod_{i=1}^{i=n} P(c_i | s)$$

is the likelihood of observing the dataset \vec{c}_n

n is the number of elements in a dataset (experiment)

N denotes the number of datasets in an ensemble.

Maximum Likelihood method for fitting theory to data

- Due to R.A. Fisher(Philo. Trans.Roy.Soc Ser A 222 309-368 (1922))
- Maximize

$$L \equiv P(\vec{c}_n | s) = \prod_{i=1}^{i=n} P(c_i | s)$$

- The maximum likelihood point s^* yields the best fit to the data of the theoretical model specified by $P(c|s)$
- This yields the optimum estimate for the true value of s . However, no goodness of fit criterion exists. Likelihood at maximum likelihood point is NOT such a measure.
- Key Observation- χ^2 GoF relies on two distributions. Theory and data. Maximum likelihood only uses Theory distribution (evaluated at the data points).

To show that Likelihood does not furnish goodness of fit measure

- Goodness of fit must be invariant under change of variable $c' = c'(c)$

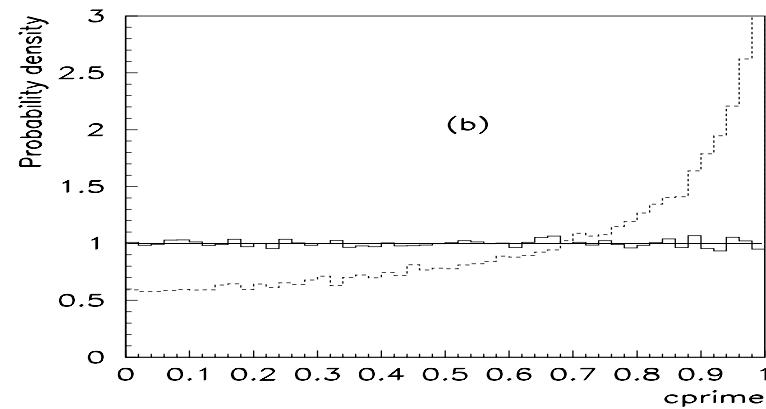
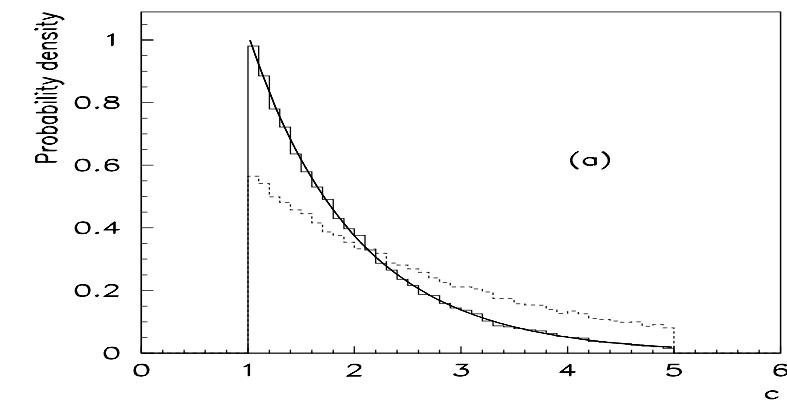
$$P(c | s) = \frac{\exp(-c/s)}{s(\exp(-(c_1/s)) - \exp(-(c_2/s)))}$$

$$c_1 = 1.0, c_2 = 5.0$$

$$c'(c) = \int_{c_1}^c P(t | s^*) dt$$

$$\left| \frac{\partial c'}{\partial c} \right| = P(c | s^*)$$

$$P(c' | s^*) = P(c | s^*) \left| \frac{\partial c}{\partial c'} \right| = 1$$



Likelihood Ratios

- Introduce the concept of "data likelihood" as derived from data. If $P(c)$ is the "frequentist" pdf of the data, the the "data likelihood" for dataset \vec{c}_n is defined as

$$P^{data}(\vec{c}_n) = \prod_{i=1}^{i=n} P(c_i)$$

- This object may be thought of as the pdf of the n-object \vec{c}_n .
- Then the likelihood ratio $L_R = \frac{P(\vec{c}_n | s)}{P^{data}(\vec{c}_n)}$ $L'_{R'}(c') = \frac{P(c' | s)}{P^{data}(c')} \begin{vmatrix} \frac{\partial c}{\partial c'} \\ \vdots \\ \frac{\partial c}{\partial c'} \end{vmatrix} = L_R(c')$
- Is invariant under change of variable since the Jacobians cancel in the numerator and denominator. We will show that it asymptotes to a χ^2 distribution for large statistics (binned case) and hence will satisfy as a GOF measure.
- Maximum Likelihood point s^* and Maximum likelihood ratio point are the same, since denominator does not depend on s .
- Likelihood ratios multiply. $L_R^{m+n} = L_R^m \times L_R^n = \frac{P(\vec{c}_m | s)}{P^{data}(\vec{c}_m)} \times \frac{P(\vec{c}_n | s)}{P^{data}(\vec{c}_n)}$

Historical use of Likelihood ratios

- Likelihood ratios are enshrined in statistics by the Neyman-Pearson lemma which states that the cut

$$L_R = \frac{P(\vec{c}_n | s_1)}{P(\vec{c}_n | s_2)} > \varepsilon$$

will have the optimum power in differentiating between the hypotheses s_1 and s_2 where ε is a cut chosen to obtain the requisite purity.

- Note that the Neyman-Pearson use of the likelihood ratio is *between two theoretical distributions* both in the numerator and the denominator.
- We believe, we are introducing the notion of data likelihood as evaluated using the data pdf alone. This quantity restores the goodness of fit in both the binned and unbinned likelihood fits, yielding a general theory of goodness of fit.
- It provides a method of inverting probabilities without a Bayesian prior
- What is it- It is merely a generalized pdf of the n-object \vec{c}_n

GoF in Unbinned Likelihood Fits

- We need to evaluate

$$L_R = \frac{P(\vec{c}_n | s)}{P^{data}(\vec{c}_n)}$$

- For the unbinned case. This is done if we have an algorithm to evaluate the denominator for each event c . We accomplish this by means of Probability Density Estimators (PDE's) also known as Kernel Density Estimators (KDE's).
- Essentially each event is smeared by a Kernel Density (Gaussians are popular kernels). All the Kernels are added up and the sum divided by n , the number of events. The resulting PDE approximates to the pdf of the data $P(c)$.
- We evaluate the likelihood ratio at the maximum likelihood point s^* . In order to quantify the goodness of fit, we evaluate the distribution of the negative log likelihood ratio (NLLR) for an ensemble of fits for Monte Carlo generated to mimic the theoretical curve. NO analytic theory as yet available.

Probability Density Estimators(PDE)

E.Parzen, Ann Math. Statis 32, 1065-1072

- If d is the dimension of the vector c , then

$$\langle c^\alpha \rangle = \frac{1}{n} \sum_{i=1}^{i=n} c_i^\alpha$$

$$E^{\alpha,\beta} = \langle c^\alpha c^\beta \rangle - \langle c^\alpha \rangle \langle c^\beta \rangle$$

$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}h)^d \sqrt{\det(H)}} \exp\left(\frac{-H^{\alpha\beta}c^\alpha c^\beta}{2h^2}\right)$$

- $H=E^{-1}$. h is a smoothing factor

$$P(c) \approx PDE(c) = \frac{1}{n} \sum_{i=1}^{i=n} \mathcal{G}(c - c_i)$$

$$P(c) = \int P(c) G_\infty(c - c_i) dc_i$$

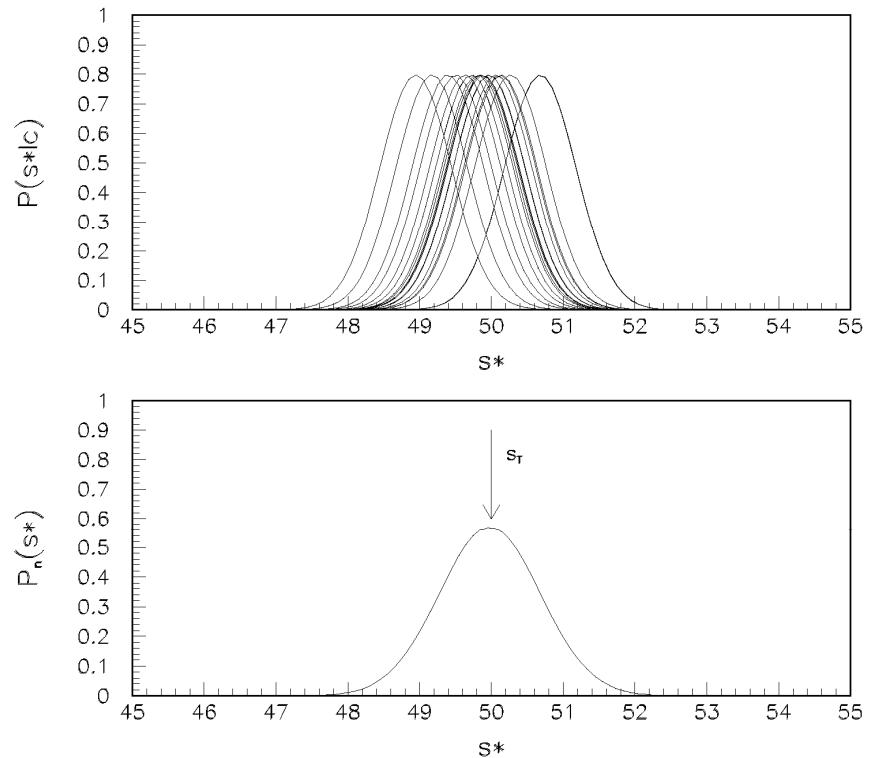
Probability Density Estimators

$$G_\infty(c - c_i) \equiv \lim_{n \rightarrow \infty} G(c - c_i) = \delta(c - c_i)$$

- This is assured by making the smoothing factor depend on the number of events.

$$h \approx n^{-1/(d+4)}$$

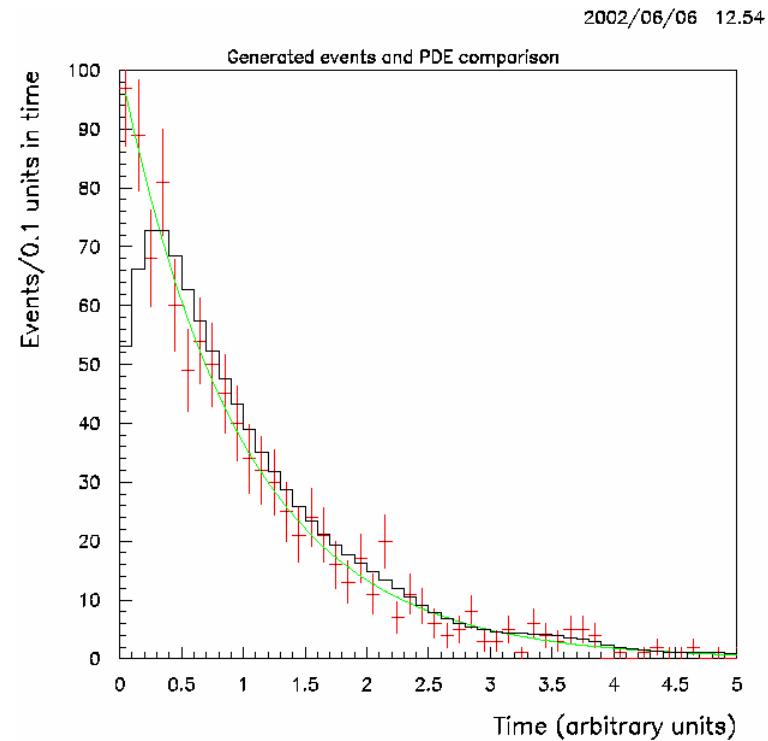
- PDE's are generalizable to arbitrary dimensions.



Illustrative Example

$$P(c|s) = \frac{1}{s} \exp\left(-\frac{c}{s}\right)$$

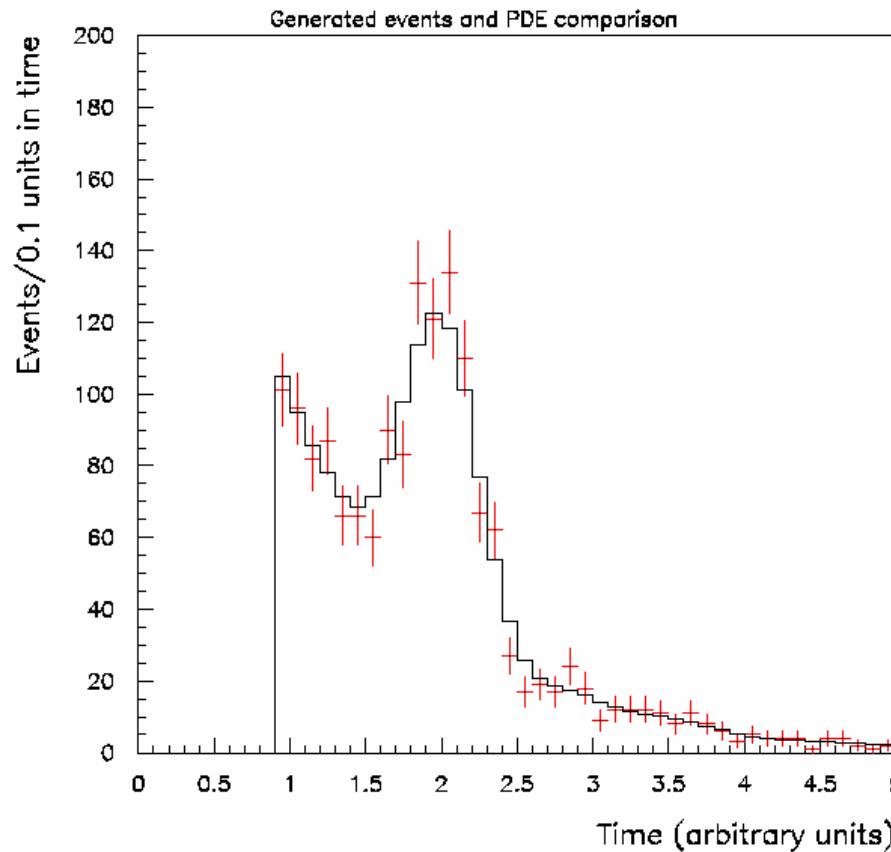
$$\mathcal{G}(c) = \frac{1}{(\sqrt{2\pi}sh)} \exp\left(-\frac{c^2}{2s^2h^2}\right)$$



Histogram (with errors) shows 1000 events generated as an exponential $P(c|s)$, for $s=1.0$. Superimposed is the PDE estimator (solid histogram). Boundary effect near $c=0$.

PDE tracks data

2002/06/06 12.53



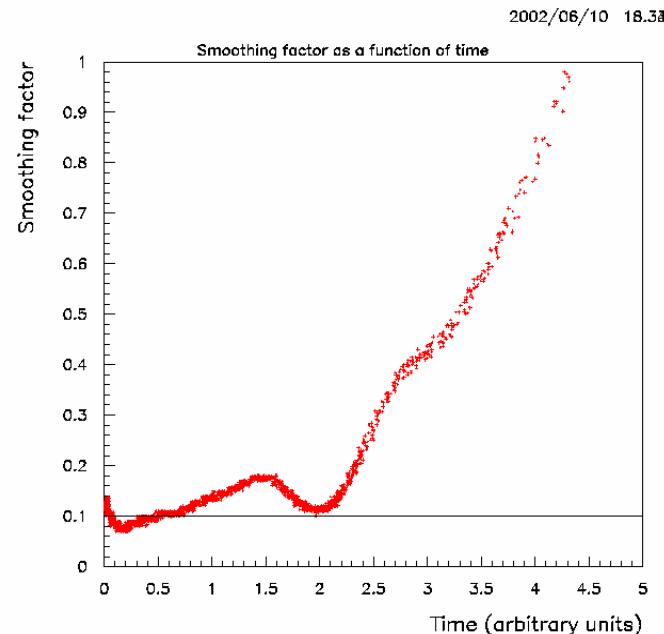
Histogram with errors shows exponential (1000 events) with superimposed Gaussian of 500 events ($\mu=2.0$, $\sigma=0.2$). The PDE estimator (Solid) histogram.

Improve the smoothing factor

- Smoothing factor should be allowed to vary as a function of event density. Estimate event density using constant smoothing factor and then apply the formula

$$h(c) = \left(\frac{n PDE(c)}{(t_2 - t_1)} \right)^{-0.6}$$

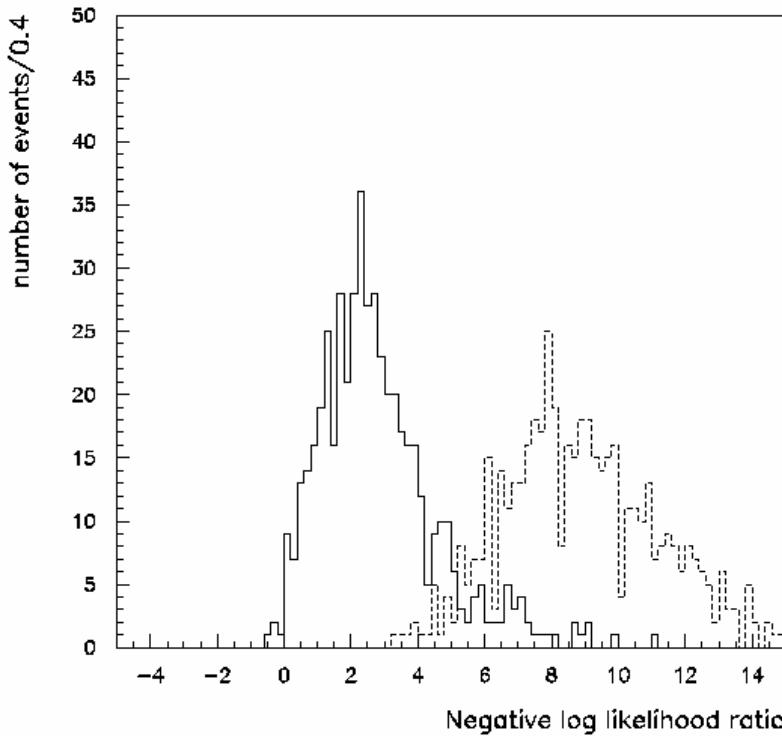
- t_1-t_2 are the lower and upper limits of c
- Iterate 3 times till things settle down.



Variation of h as a function of c for the above example determined iteratively.

Variable vs constant smoothing factor

2002/06/07 14.49



Solid Histogram shows NLLR for 500 distributions using iterative smoothing. The dashed curve corresponds to the case for constant smoothing factor h .

Results of unbinned and binned fitting

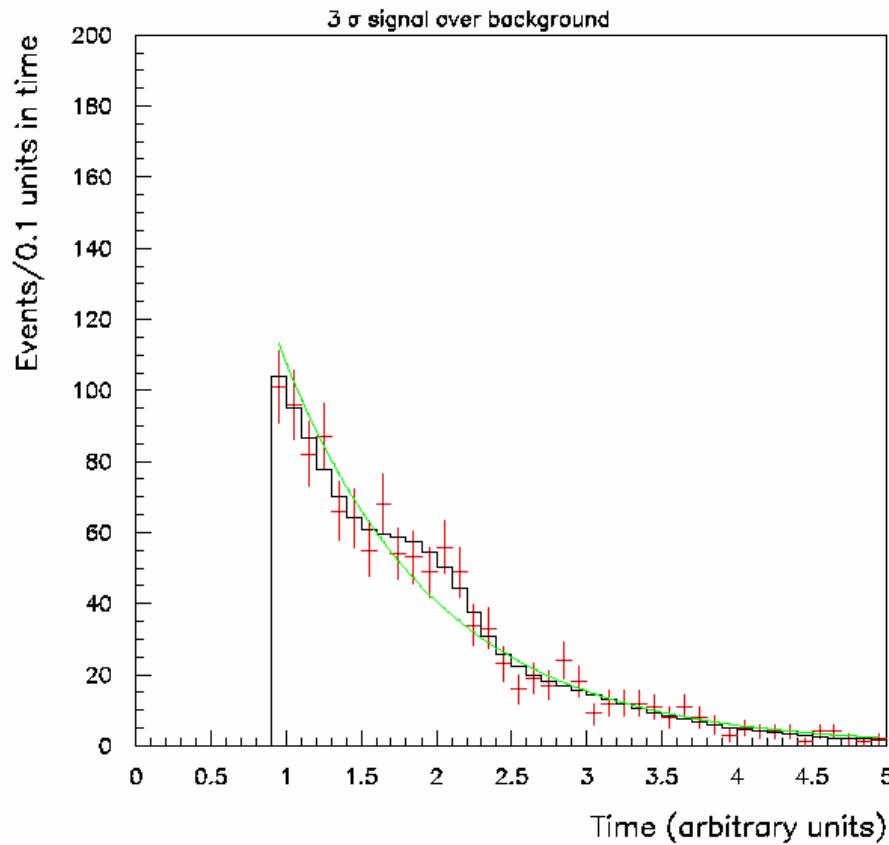
- 1000 background events

TABLE I:

Number of Gaussian events	Unbinned fit \mathcal{NLLR}	Unbinned fit $N\sigma$	Binned fit χ^2 39 d.o.f.
500	189.	103	304
250	58.6	31	125
100	11.6	4.9	48
85	8.2	3.0	42
75	6.3	1.9	38
50	2.55	-0.14	30
0	0.44	-1.33	24

An example of an unbinned fit.

2002/06/07 17.33

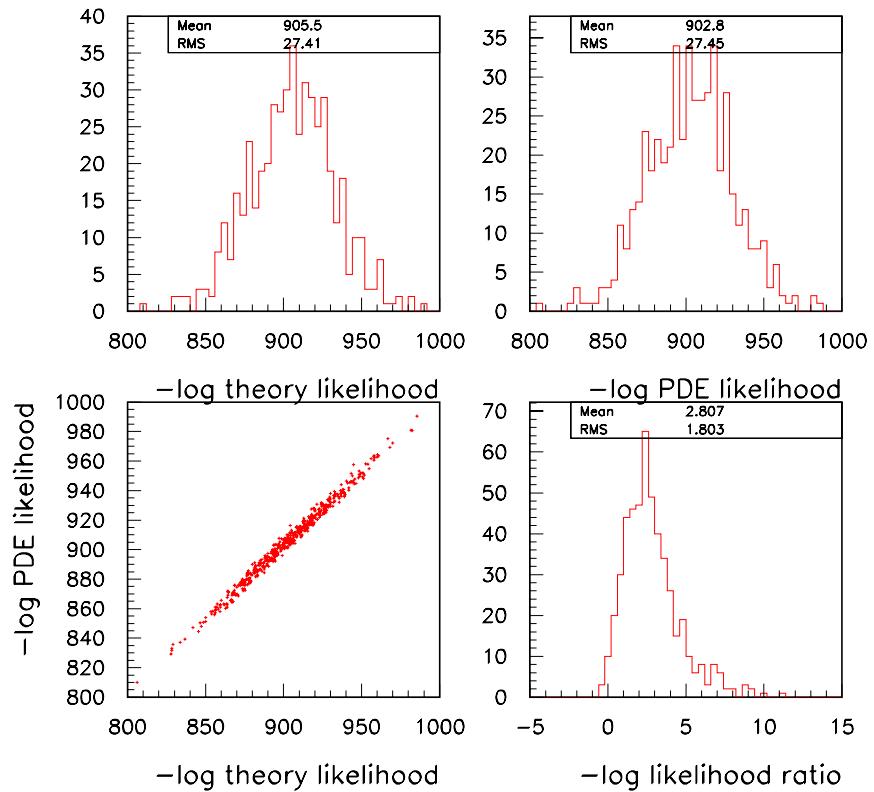


Histogram (with errors) is a Gaussian of 85 events ($\mu=2, \sigma=0.2$) superimposed on an exponential of 1000 events. This causes an NLLR that is 3σ away from the mean NLLR for the null hypothesis.

Likelihood vs Likelihood Ratio

- Negative log likelihoods, beside not being invariant are broad. (Top Left)
- Negative Log PDE likelihood are similar (Top right) and contain information from data!
- The two correlate (Bottom left)
- The difference is the narrow GoF, negative log likelihood ratio (Bottom right)

2002/08/23 12.31



Unbinned Goodness of Fit-Improving the PDE

- Apply same theory. We now need a way to estimate the "data density" for unbinned data.
- Steps- First, perform an unbinned likelihood fit.
- Second-Evaluate the "data likelihood" $P(\vec{c}_n)$ by the method of probability density estimators (PDE)- also known as Kernel Density Estimators (KDE).
- Reported on this in PHYSTAT03.
- To summarize, the best PDE method is to transform to hypercube co-ordinates, where $P(c'|s)$ is flat.
- The likelihood ratio L_R is invariant and will not depend on the co-ordinate system.
- For each event c_i , $i=1,n$, define a boxcar function centered on c_i of width h so that the kernel

$$G(c') = \frac{1}{h}; |c'| < \frac{h}{2}$$

$$G(c') = 0; |c'| > \frac{h}{2}$$

$$\int G(c') dc' = 1$$

- Use periodical boundary conditions on the kernels so that

$$G(c' - c'_i) = G(c' - c'_i - 1); c' > 1$$

$$G(c' - c'_i) = G(c' - c'_i + 1); c' < 0$$

- Treats every point in hypercube like every other.

Unbinned Likelihood gof

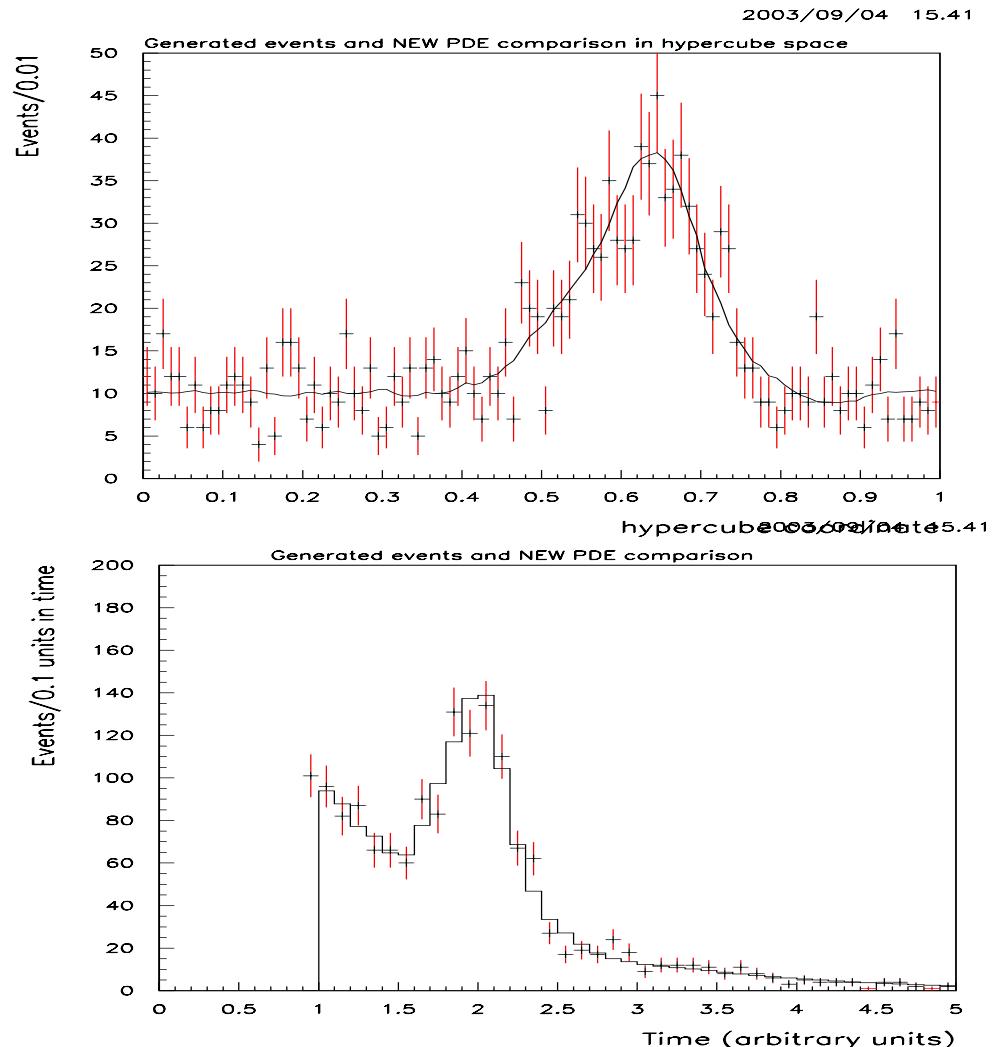
- Then the PDE for the data likelihood is given by

$$P^{data}(c) \approx PDE(c) = \frac{1}{n} \sum_{k=1}^{k=n} G(c - c_k)$$

$$\int P^{data}(c) dc = 1$$

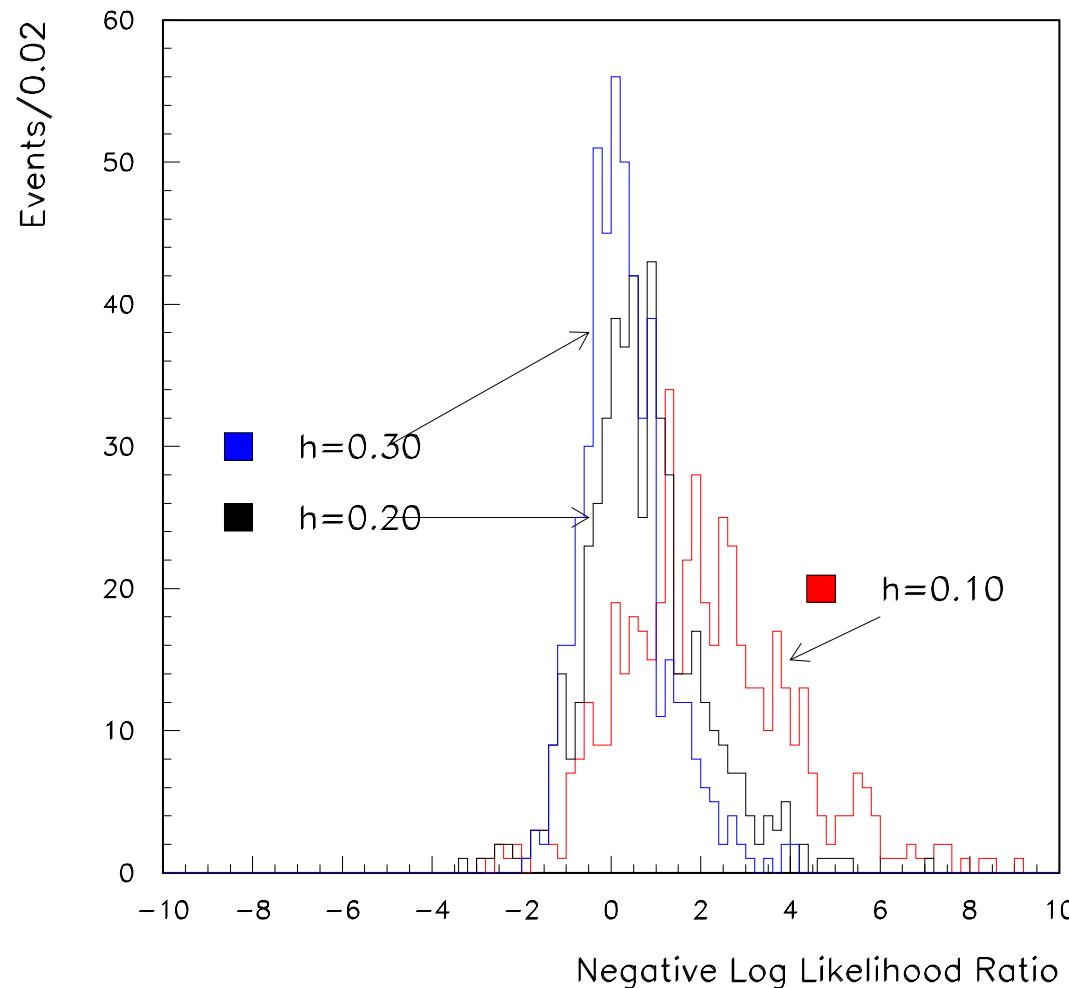
$$P(\vec{c}_n) = \prod_{i=1}^{i=n} P^{data}(c_i)$$

- Exponential + Gaussian- Data (crosses) and PDE histogram



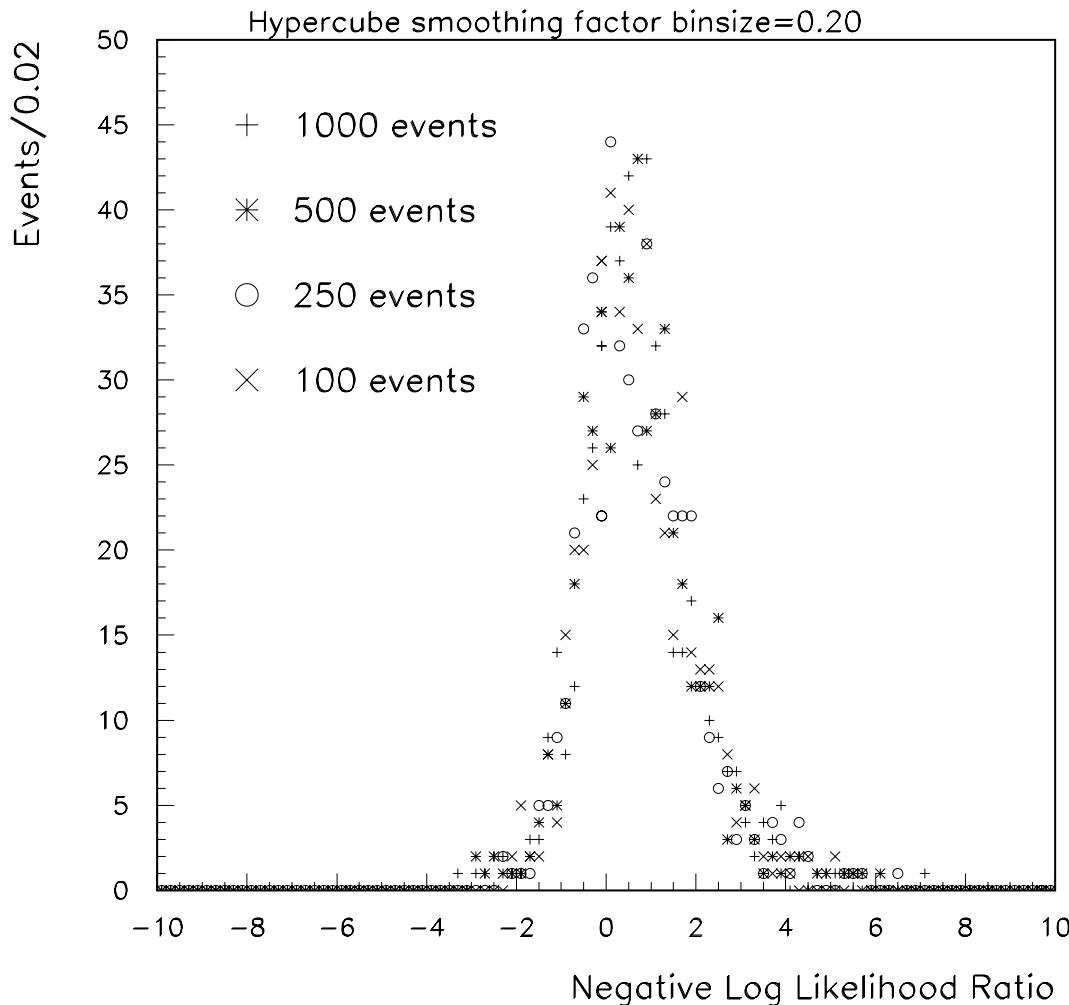
Dependence of NLLR on h

2004/01/22 11.24



Dependence of NLLR on n

2003/11/28 19.51



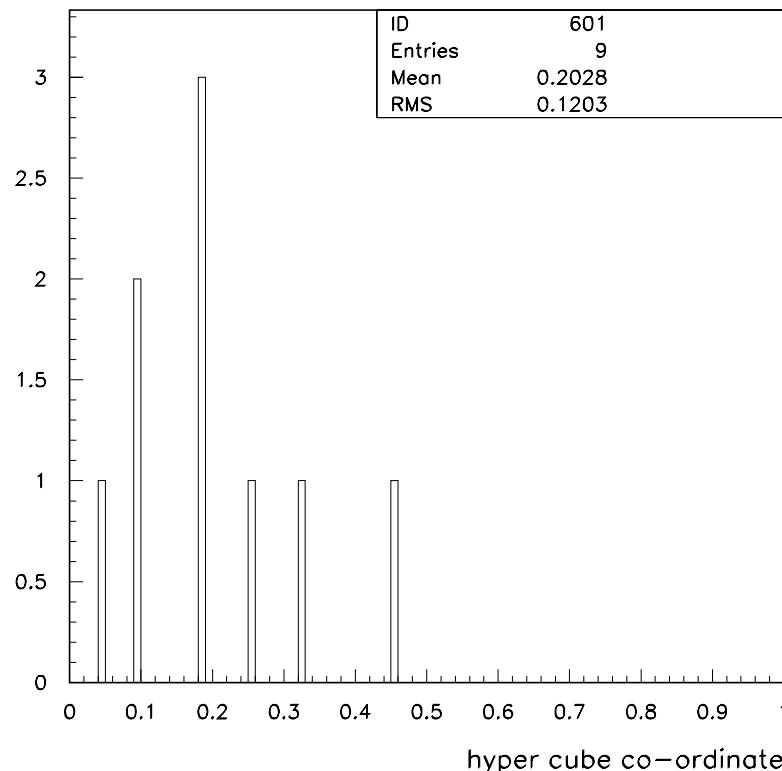
Extreme Example

- Problem-(set by Bruce Knuteson)-3 data points in 3D space have coordinates $(x,y,z)=(0.1,0.2,0.3)$, $(0.2,0.4,0.1)$, and $(0.05,0.6,0.21)$. What is the goodness of fit to the hypothesis that they are distributed according to $P(x,y,z)=\exp(-(x+y+z))$?
- Likelihood function is given by

$$L = \prod_{i=1}^{i=3} \frac{1}{s} \exp(-(x_i + y_i + z_i)/s); s = 1.0$$

$$L = \prod_{i=1}^{i=9} \frac{1}{s} \exp(-c_i/s); \vec{c}_9 = (0.1, 0.2, 0.3, 0.2, 0.4, 0.1, 0.05, 0.6, 0.21)$$

2005/06/24 15.24



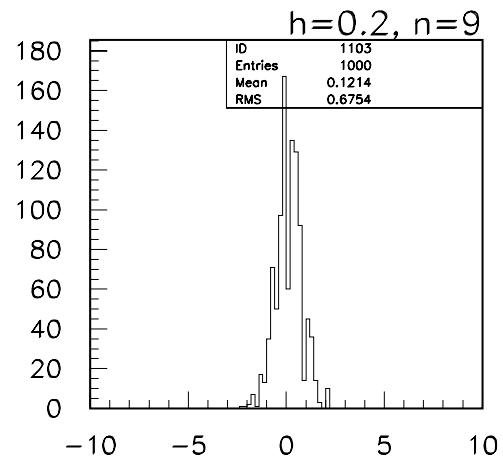
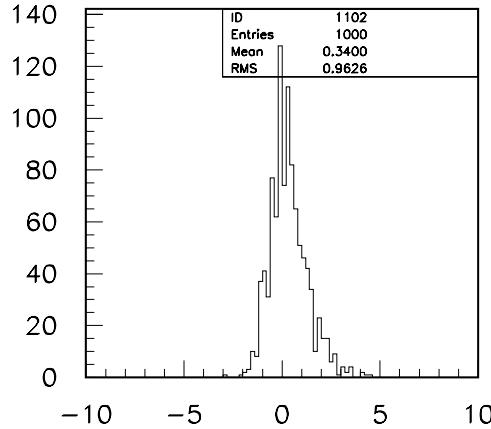
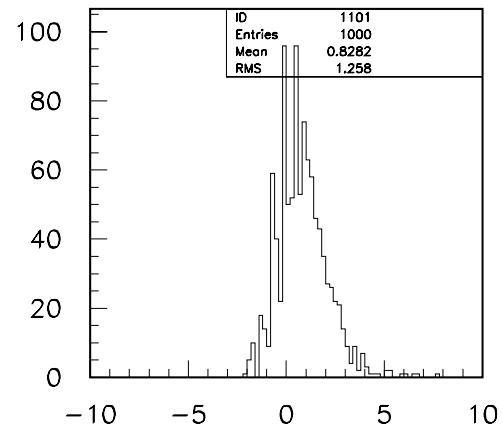
Results

- The data are a poor fit to the hypothesis

Smoothing Parameter h	NLLR	Probablilty to exceed NLLR
0.2	5.36	0.5%
0.3	5.84	<0.1%
0.4	1.77	1%

Extreme Example-NLLR vs h n=9

2005/06/24 15.24



$h=0.2, n=9$

$h=0.3, n=9$

$h=0.4, n=9$

Binned Likelihood goodness of fit

- Let there be n_b bins and let the k^{th} bin whose center abscissa is given by c_k contain n_k entries.
- Then the probability of obtaining the histogram containing a total of n events is given by the multinomial

$$P(\text{histogram}) = \frac{n!}{\prod_{k=1}^{n_b} n_k!} \prod_{k=1}^{n_b} P(c_k | s)^{n_k}$$

$$\sum_{k=1}^{n_b} n_k = n$$

- Histogram is degenerate with degeneracy factor

$$D = \frac{n!}{\prod_{k=1}^{n_b} n_k!}$$

- Each such histogram has the same goodness of fit
- So we can just solve the problem for one of them.

Binned Likelihood goodness of fit

- Then the binned likelihood ratio is

$$L_R = \prod_{k=1}^{k=n_b} \left[\frac{P(c_k | s)}{P^{data}(c_k)} \right]^{n_k}$$

- We can estimate the denominator by

$$P^{data}(c_k) \approx \frac{n_k}{n \Delta c_k}$$

where Δc_k is the bin width for the k^{th} bin. This yields

$$L_R = \prod_{k=1}^{k=n_b} \left[\frac{T_k}{n_k} \right]^{n_k}; G^2(\text{Multinomial}) = -\log_e L_R$$

where the T_k is the theoretically expected number of events in the k^{th} bin given by

$$T_k = n \Delta c_k P^{\text{bin average}}(c_k | s)$$

This is the same result as given by Baker and Cousins(NIM A221, 1984) for the multinomial but derived using the new theory.

Comparison of derivation with Baker and Cousins

- Baker and Cousins justify the use of a likelihood ratio by the likelihood ratio theorem and employ a L_R of theory to the true value.

$$L_R^{BC} = \prod_{k=1}^{k=n_b} \left[\frac{T_k}{\text{True Value}_k} \right]^{n_k}$$

- Then they say the maximum likelihood for the true value is when $(\text{True value})_k = n_k$. This yields the same result, but our rationale is very different. We form a ratio between the theoretical likelihood and the new concept the data likelihood as derived from data. Their theory cannot do unbinned likelihood GoF's.
- We now show that this quantity asymptotes to a χ^2 distribution for large n .

To show that NLLR asymptotes to a χ^2 distribution

$$\lambda_k = n_k - T_k; \sum_{k=1}^{k=n_b} \lambda_k = 0$$

$$L_R = \prod_{k=1}^{k=n_p} \left[1 - \frac{\lambda_k}{n_k} \right]^{n_k}$$

$$NLLR = -\log_e(L_R) = -\sum_{k=1}^{k=n_b} n_k \log_e \left[1 - \frac{\lambda_k}{n_k} \right]$$

$$= \sum_{k=1}^{k=n_b} n_k \left[\frac{\lambda_k}{n_k} + \frac{1}{2} \left(\frac{\lambda_k}{n_k} \right)^2 + \frac{1}{3} \left(\frac{\lambda_k}{n_k} \right)^3 + \dots \right]$$

$$\text{As } n \rightarrow \infty, NLLR \rightarrow \sum_{k=1}^{k=n_b} \frac{1}{2} \left(\frac{\lambda_k}{n_k} \right)^2 = \frac{1}{2} \sum_{k=1}^{k=n_b} \chi_k^2$$

$$E(L_R) = \exp(-n_b / 2)$$

Maximum likelihood fits only shapes not normalizations

- No information on the theoretical expected number of events. We evaluate the likelihood ratio on the observed data events. Both the theory likelihood and data likelihood integrate to unity. So no information on the normalization.
- Get around this by using the binomial distribution and demanding the observed number of events be in the first bin. Let N be the number of identical experiments in an ensemble. Then for the binomial, the likelihood ratio, using our theory, can be written

$$L_R = \left(\frac{n_t}{n} \right)^n \left(\frac{N-n_t}{N-n} \right)^{N-n} = \left(\frac{n_t}{n} \right)^n \left(\frac{1-n_t/N}{1-n/N} \right)^{N-n}$$

where n_t is the theoretical number of events expected

n is the experimental number of events observed

N is the number of experiments in the ensemble

- We let N go to infinity in two well known limits.
- Poisson— n_t and n finite.

$$L_R = e^{-(n_t-n)} \left(\frac{n_t}{n} \right)^n$$

Binned Likelihood gof- Poisson Normalization gof

- Multiplying this L_R with binned likelihood L_R yields the formulae

$$NLLR = \sum_{k=1}^{k=n_b} n_k \log_e \left(\frac{n_k}{T_k} \right); \text{shapes only}$$

$$T'_k = \frac{n_t T_k}{n}; \sum_{k=1}^{n_b} T'_k = n_t$$

$$G^2(Poisson) = \sum_{k=1}^{k=n_b} T'_k - n_k + n_k \log_e \left(\frac{n_k}{T'_k} \right); \text{shapes and normalization}$$

- Same as the "multinomial" and "poissonian" results of Baker and Cousins.
- Gaussian Limit of the Binomial, n and n_t also go to infinity as N goes to infinity. Yields

$$L_R = e^{-\frac{(n-n_t)^2}{2\sigma^2}}$$

$$\sigma^2 = Npq; p = n/N; q = 1-p$$

To show that Pearson χ^2 distribution is an NLLR

- Let the contents of the k^{th} bin be denoted by c_k .
- Let the theoretical expectation for the k^{th} bin be s_k .
- Let the standard deviation of the k^{th} bin be σ_k .
- Then,

$$P(c_k | s_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(c_k - s_k)^2}{2\sigma_k^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\chi_k^2}{2}\right)$$

$$-\log_e(P(c_k | s_k)) = \frac{\chi_k^2}{2} + \log_e(\sqrt{2\pi}\sigma_k)$$

- People mistakenly conclude that χ^2 is the negative log of a likelihood. Correct expression is to note that

$$P(c_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(c_k - c_k)^2}{2\sigma_k^2}\right) = \frac{1}{\sqrt{2\pi}\sigma_k}$$

$$L_R = \frac{P(c_k | s_k)}{P(c_k)} = \exp\left(-\frac{\chi_k^2}{2}\right)$$

Why is this theory important to HEP?

- Very useful in multivariate analyses. Even though we have shown examples in 1 dimension, theory is easily generalizable to multi-dimensions.
Theoretical model can also be Monte Carlo events.
Theory and data are interpolated by use of PDE's.
- Theory can be signal + background Monte Carlo
- NLLR distribution obtained by comparing one set of theory (numerator) vs many such generations of "data from theory" for denominator.
- GoF is useful in evaluating one set of theoretical curves wrt another.
- Likelihood can be done in correlated variables or uncorrelated variables.
- Excellent tool for LHC analyses.

Errors in Fitted Parameters

- Problem of “inverse probability”.
- Theoretical likelihood is given by

$$L \equiv P(\vec{c}_n | s) = \prod_{i=1}^{i=n} P(c_i | s)$$

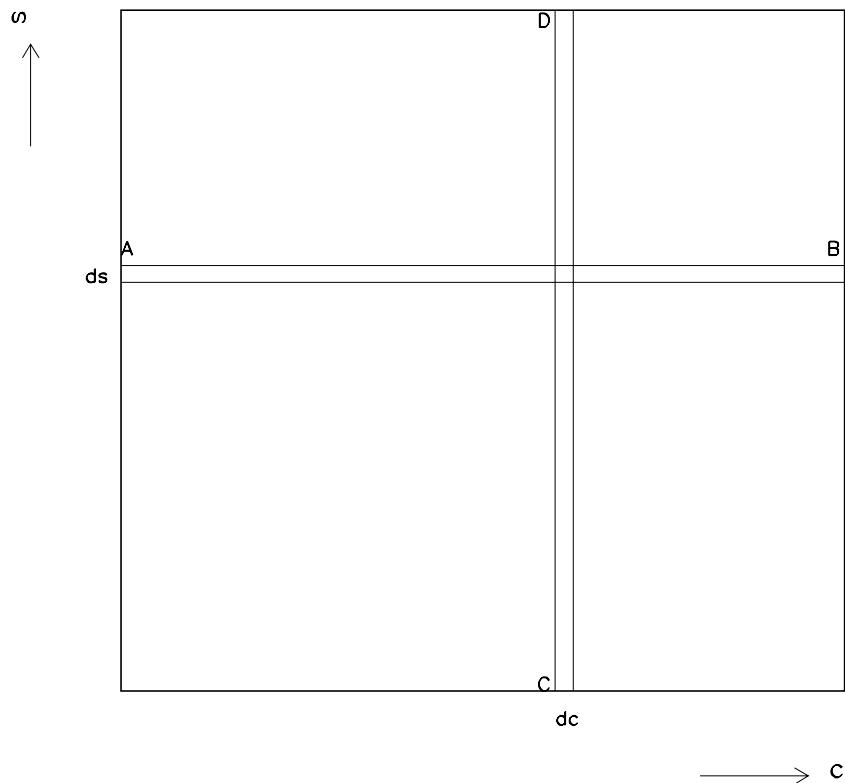
- In order to estimate the errors in s , we need to determine the “inverse probability”

$$P(s | \vec{c}_n)$$

One needs Bayes’ theorem. In Bayesian language this is also known as “Posterior density”.

Bayes' Theorem-Simple derivation

2003/09/20 17.09



Define a joint probability density $P(s,c)$ such that

$$\iint P(s,c) ds dc = 1$$

Then define projections $P(c)$, $P(s)$ such that

$$\int P(s,c) dc = P(s); \int P(s,c) ds = P(c);$$

$$\int P(s) ds = 1; \int P(c) dc = 1$$

Define conditional probability $P(c|s)$ along line AB

$$P(c|s) = \frac{P(s,c)}{\int P(s,c) dc} = \frac{P(s,c)}{P(s)}$$

Define conditional probability $P(s|c)$ along line CD

$$P(s|c) = \frac{P(s,c)}{\int P(s,c) ds} = \frac{P(s,c)}{P(c)}$$

Bayes' Theorem-Simple derivation

- Then

$$P(s, c) = P(s | c)P(c) = P(c | s)P(s)$$

$$\text{or } P(s | c) = \frac{P(c | s)P(s)}{P(c)} \text{ Bayes' Theorem}$$

$$\text{and } P(c) = \int P(c | s)P(s)ds$$

$$P(s) = \int P(s | c)P(c)dc$$

- Generalizing to dataset \vec{c}_n

$$P(s, \vec{c}_n) = P(s | \vec{c}_n)P(\vec{c}_n) = P(\vec{c}_n | s)P(s)$$

$$\text{or } P(s | \vec{c}_n) = \frac{P(\vec{c}_n | s)P(s)}{P(\vec{c}_n)}$$

$$\text{and } P(\vec{c}_n) = \int P(\vec{c}_n | s)P(s)ds$$

$$P(s) = \int P(s | \vec{c}_n)P(\vec{c}_n)d\vec{c}_n$$

The Bayesian Paradigm

- Note that Bayes' theorem is a theorem in mathematics and not statistics, since it can be derived for any function of 2 variables with finite integral. The "conditional probabilities" are appropriately normalized "slices" of such a function and the functions $P(s)$ and $P(c)$ are projections of such a function.
- It is up to us how to identify the various functions and map them to our theory.
- In order to invert the probability, one needs to specify one projection and one slice function. This can be used to determine the joint probability, from which the inverse probability may be computed.
- The Bayesian paradigm is to assume that $P(s)$, the projection on the parameter axis of the joint probability is a function that is obtained by prior knowledge of the parameter. This function is supplied during inversion and is known as the Bayesian prior.

The Bayesian Paradigm

- Several kinds of Bayesian methods- for providing the prior-
 - » Objective Bayesianism (E.T.Jaynes et al) Assume the prior is flat in the variable of interest within some reasonable bounds- Immediate objection- Flat in which variable, s , s^2 , $\log_e s$, $\log_e(\log_e s)$ etc? Results will vary.
 - » Subjective Bayesianism (de Finetti et al). Prior knowledge depends on the prior "history" of the experimenter. So different priors should be used. A collider experiment can and will have more priors than experimenters and as many different results!
 - » Hierarchical Bayesianism- Parameterize the prior by some new parameters. These parameters will have their own priors and so on!
 - » Empirical Bayesianism- Attempts to terminate the above infinite regression by determining the parameters from data.

The Bayesian Paradigm

- All Bayesians use Bayes' theorem. So they term the standard method of obtaining errors from χ^2 fitting ($\Delta\chi^2=1$) and from likelihood fitting ($\Delta L=1/2$) as illegal, since these do not use priors.
- The Bayesians then compute the posterior density as

$$P(s | \vec{c}_n) = \frac{P(\vec{c}_n | s)P(s)}{P^{Bayes}(\vec{c}_n)} = \frac{P(\vec{c}_n | s)P(s)}{\int P(\vec{c}_n | s)P(s)ds}$$

where $P^{Bayes}(\vec{c}_n) = \int P(\vec{c}_n | s)P(s)ds$ an uninteresting theoretical constant!

Bayesians then compute

$$\int P(s | \vec{c}_n)P^{Bayes}(\vec{c}_n)d\vec{c}_n = P(s) \text{ an n independent Bayesian prior}$$

- Bayesians do not have GoF.

All Bayesians make this substitution

$$P^{\text{Bayes}}(\vec{c}_n) = \int P(\vec{c}_n | s)P(s)ds$$

142 Conditional Prevision and Probability

Theory of Probability

A critical introductory treatment

Volume 1

BRUNO DE FINETTI

(1905–1985)

Translated by

ANTONIO MACHI

Assistant Professor of Mathematics
at the University of Rome,
Italy

and

ADRIAN SMITH

Professor of Mathematics
at the University of Nottingham,
UK

Wiley Classics Library Edition Published 1990



INTERSCIENCE PUBLISHERS

JOHN WILEY & SONS

Chichester · New York · Brisbane · Toronto · Singapore

4.6 LIKELIHOOD

4.6.1. *Bayes's theorem*—in the case of events E , but not random quantities X —permits us to write $\mathbf{P}(\cdot | H)$ in the form we met above, a form which is often more expressive and practical:

$$(5) \quad \mathbf{P}(E|H) = \mathbf{P}(E)\mathbf{P}(H|E)/\mathbf{P}(H) = K \cdot \mathbf{P}(E)\mathbf{P}(H|E),$$

where the normalizing factor, $1/\mathbf{P}(H)$, can be simply denoted by K , and, more often than not, can be obtained more or less automatically without calculating $\mathbf{P}(H)$. For this reason, it is often convenient to talk simply in terms of *proportionality* (i.e. by considering $\mathbf{P}(\cdot | H)$ only up to an arbitrary, non-zero, multiplicative constant, which can be determined, if necessary, by normalizing).

One could say that $\mathbf{P}(\cdot | H)$ is proportional to $\mathbf{P}(\cdot)$ and to $\mathbf{P}(H|\cdot)$, where the dot stands for E , thought of as varying over the set of all the events of interest. More concisely, this is usually expressed by saying that

'final probability' = K 'initial probability' × 'likelihood', where $= K$ denotes proportionality, and we agree to call:

the *initial* and *final* probabilities those not conditional or conditional on H , respectively (i.e. evaluated before and after having acquired the additional knowledge in question, H), and

the *likelihood* of H given E , the $\mathbf{P}(H|E)$ thought of as a function of E (and possibly multiplied by any factor independent of E , e.g. $1/\mathbf{P}(H)$), the use of which would allow the substitution of '=' for ' $= K$ ', or anything resulting from the omission of common factors, more or less cumbersome, or constant, or dependent on H). The term 'likelihood' is to be understood in the sense that a larger or smaller value of $\mathbf{P}(H|E)$ corresponds to the fact that the knowledge of the occurrence of E would make H either more or less probable (our meaning would be better conveyed if we spoke of the 'likelihoodization' of H by E).

4.6.2. This discussion leads to an understanding of how it should be possible to pass from the initial probabilities to the final ones through intermediate stages, under the assumption that we obtain, successively, additional pieces of information H_1, H_2, \dots, H_n (giving, altogether, $H = H_1 H_2 \dots H_n$). In fact, one can also verify analytically that

$$\begin{aligned} \mathbf{P}(E|H_1 H_2) &= \mathbf{P}(EH_1 H_2)/\mathbf{P}(H_1 H_2) \\ &= [\mathbf{P}(E)\mathbf{P}(H_1|E)\mathbf{P}(H_2|EH_1)]/[\mathbf{P}(H_1)\mathbf{P}(H_2|H_1)] \\ &= K \cdot \mathbf{P}(E) \cdot \mathbf{P}(H_1|E) \cdot \mathbf{P}(H_2|EH_1) \\ &= (\text{the probability of } E) \times (\text{the likelihood of } H_1 \text{ given } E) \\ &\quad \times (\text{the likelihood of } H_2 \text{ given } EH_1). \end{aligned}$$

The New Paradigm

- We now try and solve for the joint probability using Bayes' theorem by retaining the denominator in the likelihood ratio as being obtained from data. This results in an iterative inversion of probability that is reminiscent of "Fisher's fiducial probability" that preserves the GoF. R.A.Fisher Annals of Eugenics,6:391-398(1935). Fisher did not solve the GoF problem, so he could not counter Baysians, though he tried.
- This leads to $P_n(s) = \int P(s | \vec{c}_n) P^{data}(\vec{c}_n) d\vec{c}_n$
but $P^{data}(\vec{c}_n) d\vec{c}_n = \frac{dN}{N}$ the ensemble density
Leads to $P_n(s) = \lim N \rightarrow \infty \left(\frac{1}{N} \sum_{k=1}^{k=N} P(s | \vec{c}_n) \right)$
- This leads to a projection $P_n(s)$ on the parameter axis that is n dependent since
 $P(s | \vec{c}_n) \rightarrow \delta(s - s_T)$ as $n \rightarrow \infty$
 $P_n(s)$ also $\rightarrow \infty$ as $n \rightarrow \infty$
- So the projection on the parameter axis cannot be an n independent Bayesian prior. Need new definitions and paradigm.

The New Paradigm- Definitions

We redefine $P_n(s)$ as the distribution of the maximum likelihood value s^* when the experiment with data set with n members is repeated N times (ensemble) and $N \rightarrow \infty$. We re-label it as $P_n(s^*)$. s_T , the true value of s , is the maximum likelihood point of $P_n(s^*)$. This is an assumption of unbiasedness in the experiment. The true value is a number and does not have a distribution. The true value is unknown and unknowable with infinite precision. The function $P_n(s^*)$ is also unknowable.

To calculate errors we assume that given a single dataset \vec{c}_n , not only is the maximum likelihood value s^* knowable, but there is information present on the distribution of s^* as well-i.e errors on s^* are computable. We call such a function $P(s^* | \vec{c}_n)$

New Paradigm

- Each member dataset k of the ensemble provides a maximum likelihood value s_k^* and also has a function $P_k(s^* | \vec{c}_n)$ associated with it.
- Then the joint probability $P(s^*, \vec{c}_n)$ is given by

$$P(s^*, \vec{c}_n) = P(\vec{c}_n | s^*) P_n(s^*) = P(s^* | \vec{c}_n) P(\vec{c}_n)$$

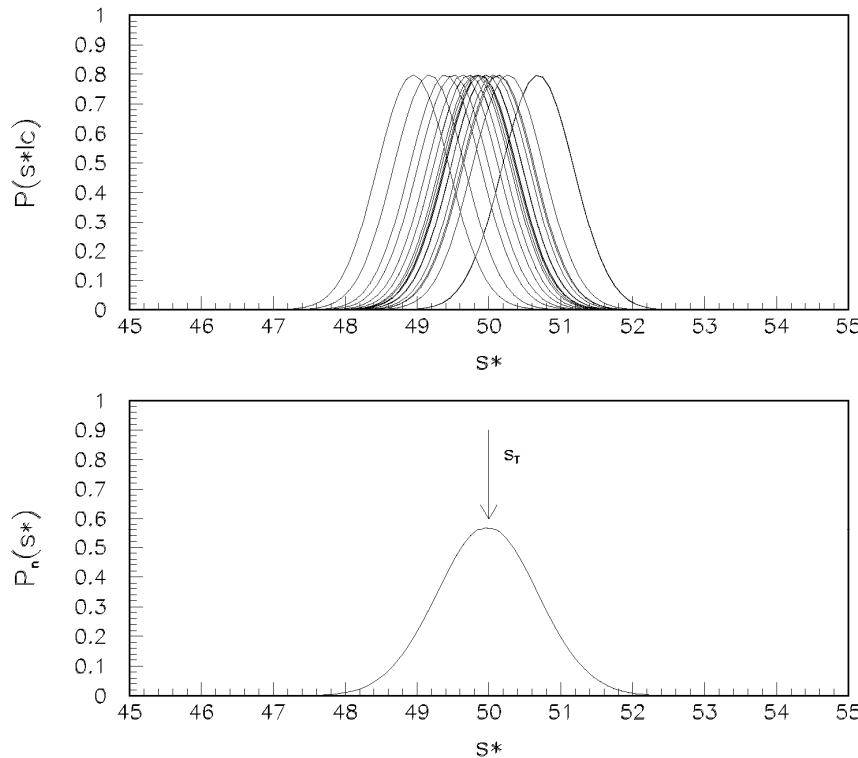
leading to

$$P(s^* | \vec{c}_n) = \frac{P(\vec{c}_n | s^*) P_n(s^*)}{\int P(\vec{c}_n | s^{*'}) P_n(s^{*'}) ds^{*'}}$$

$$P_n(s^*) = \int P(s^* | \vec{c}_n) P(\vec{c}_n) d\vec{c}_n = \sum_{k=1}^{k=N} P_k(s^* | \vec{c}_n)$$

- Not much use to us, since $P_n(s^*)$ is only available after a large number of experiments on the ensemble. After one experiment, what can we learn?

Error Bootstrap



The value of the distribution $P_n(s^*)$ at the true value is denoted by

$$P_n(s_T) = \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sigma} \text{ for the Gaussian case}$$

It depends on the individual error σ and the number n in the dataset.

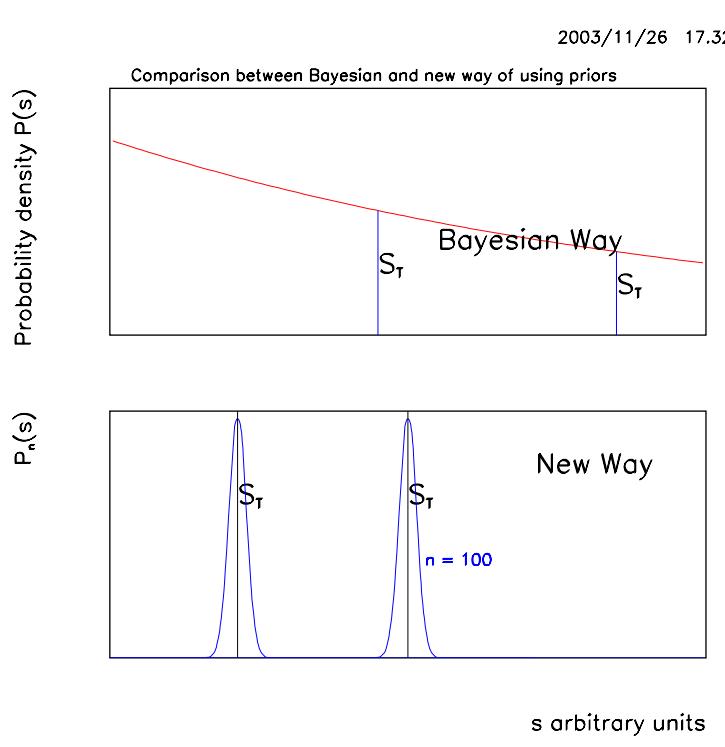
Error Bootstrap

- The quantity $P_n(s^*)$ is the ensemble average of all the posterior densities $P(s^* | \vec{c}_n)$. Its maximum likelihood value is the true value s_T .
 - We only have measurements from one member of the ensemble namely \vec{c}_n
- . We want to describe to the system our lack of knowledge of the true value. I.e. We want to say that it is at s_1 OR it is at s_2 OR it is at s_3 . At each value of s^* , we hypothesize that that is the true value.
- The likelihood ratio $L_R(s^*)$ gives the goodness of fit at that value.
 - At the true value $s^* = s_T$, the joint probability $P(s^*, c)$ is given by

$$P(s_T, \vec{c}_n) = P(s_T | \vec{c}_n)P(\vec{c}_n) = P(\vec{c}_n | s_T)P_n(s_T)$$

- As you change the value of s^* , the whole distribution $P_n(s^*)$ has to move so that the true value is at the new value of s^* . I.e. $P_n(s_T)$ in Bayes' equation is a constant.

Error Bootstrap



for true value s_T at s_1^*

$$P(s_1^*, \vec{c}_n) = P(s_1^* | \vec{c}_n) P(\vec{c}_n) = P(\vec{c}_n | s_1^*) P_n(s_T)$$

for true value s_T at s_2^*

$$P(s_2^*, \vec{c}_n) = P(s_2^* | \vec{c}_n) P(\vec{c}_n) = P(\vec{c}_n | s_2^*) P_n(s_T)$$

for true value s_T at an arbitrary s^*

$$P(s^*, \vec{c}_n) = P(s^* | \vec{c}_n) P(\vec{c}_n) = P(\vec{c}_n | s^*) P_n(s_T)$$

$$\int P(s^* | \vec{c}_n) ds^* = 1 = \frac{P_n(s_T)}{P(\vec{c}_n)} \int P(\vec{c}_n | s^{*'}) ds^{*'}$$

$$\frac{P_n(s_T)}{P(\vec{c}_n)} = \frac{1}{\int P(\vec{c}_n | s^{*'}) ds^{*'}}$$

$$P(s^* | \vec{c}_n) = \frac{P(\vec{c}_n | s^*)}{\int P(\vec{c}_n | s^{*'}) ds^{*'}}$$

This is the same formula as frequentists use! No Bayesian prior. But it is the first iteration. As we get more members of the ensemble, we will re-use Bayes' theorem iteratively.

Illustrative example

- Measure a mass whose true value s_T is unknown with an apparatus whose standard error σ is known to be 5 gms. A single data set consists of $n=100$ measurements. Then the likelihood for an arbitrary mass s is given by

$$P(c | s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(c-s)^2}{2\sigma^2}}$$

$$P(\vec{c}_n | s) = \prod_{i=1}^{i=n} P(c_i | s)$$

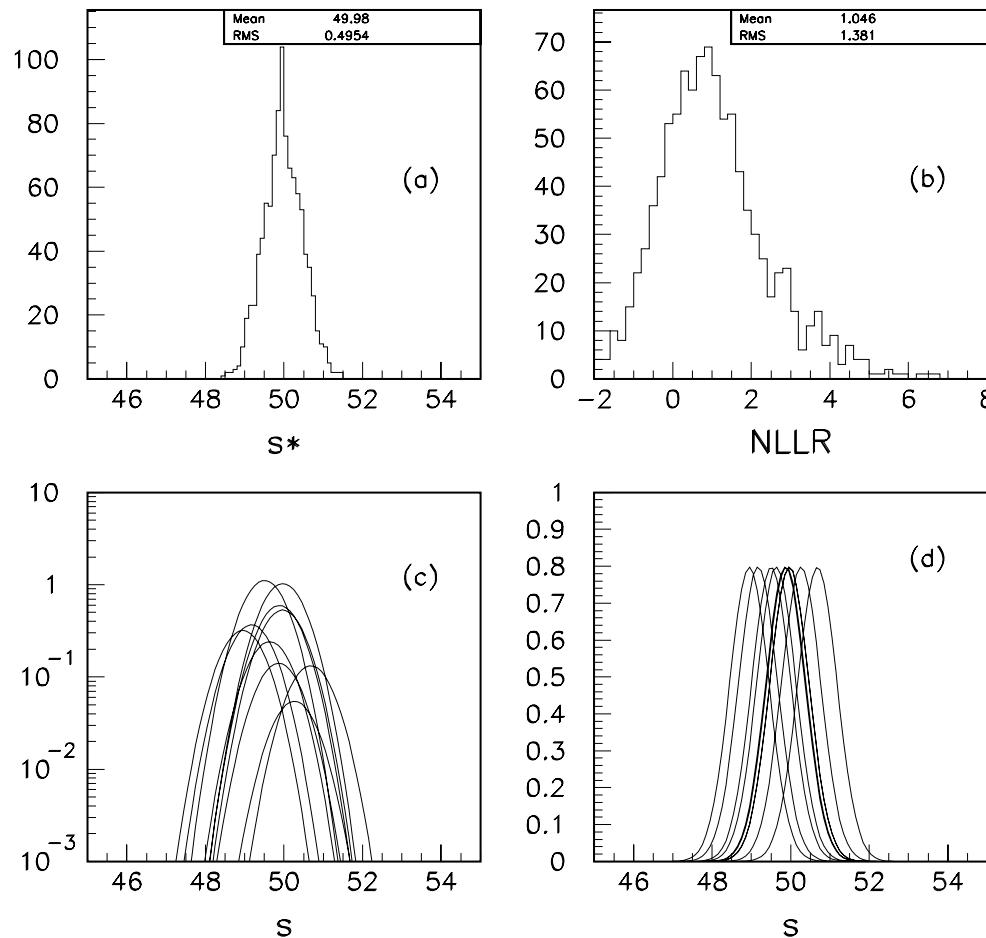
- Do goodness of fit using the method of unbinned likelihood fits. Obtain NLLR, likelihoods for each individual fit.
- Determine $P(s^* | \vec{c}_n)$ for each fit. Average over ensemble $N=1000$ fits to obtain a better value of $P_n(s^*)$. We reuse Bayes' theorem to re-evaluate posteriors, since we know $P_n(s^*)$ from the ensemble better than from individual measurements.
- One more iteration. For the k^{th} element of the ensemble

$$P_k^{\text{iter}}(s^* | \vec{c}_n) = \frac{P_k(\vec{c}_n | s^*) P_n(s^*)}{\int P(\vec{c}_n | s^*) P_n(s^*) ds^*}$$

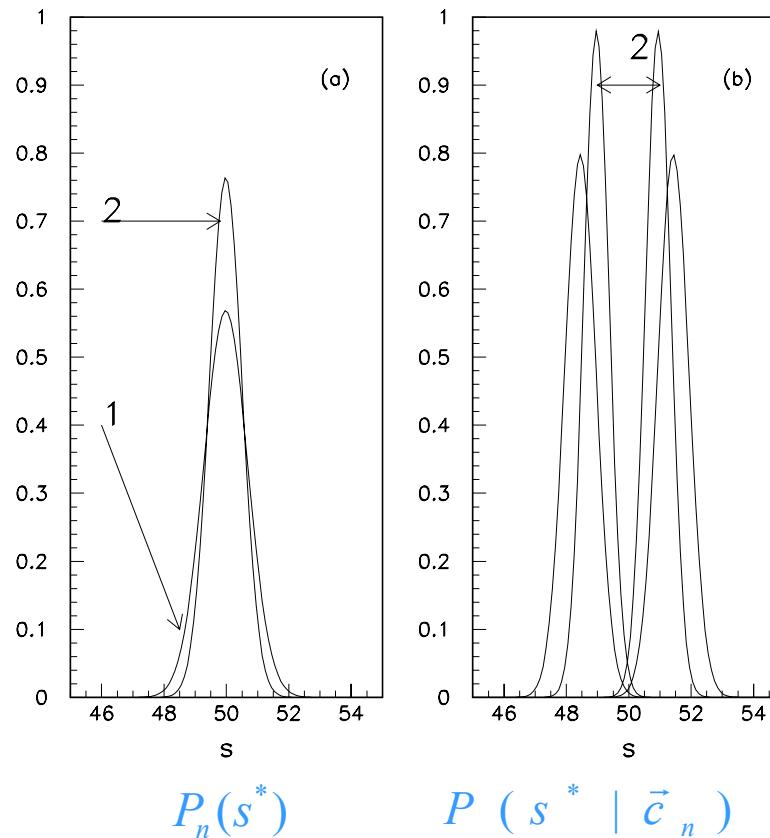
$$P_n^{\text{iter}}(s^*) = \frac{1}{N} \sum_{k=1}^{k=N} P_k^{\text{iter}}(s^* | \vec{c}_n)$$

Illustrative example

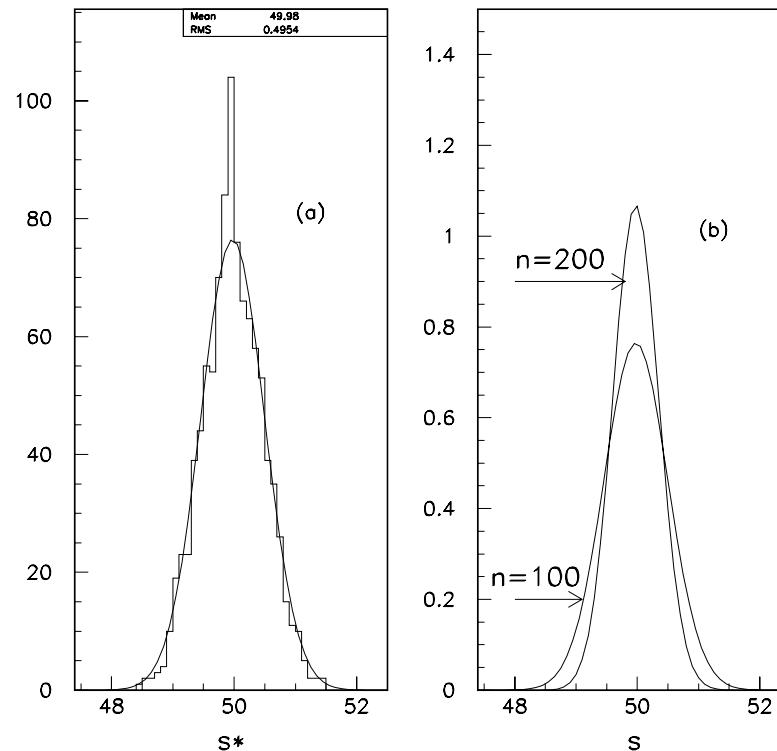
•



Illustrative example. Iterated functions



Illustrative example

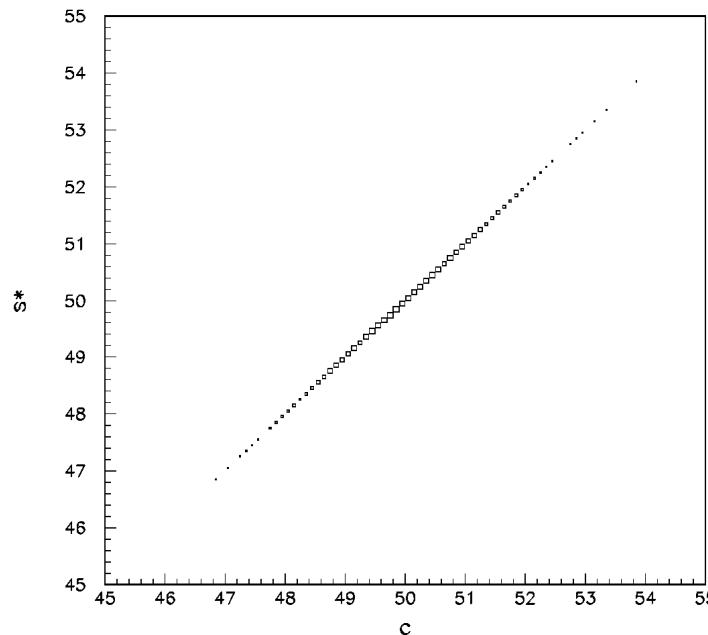


Fit to s^*
histogram

$P_n(s^*)$ for
 $n=100, 200$

Iterative Example

- Same as above, except $n=1$. I.e. One measurement. Measurement error is $\sigma=1$ gm. We can project to both s^* and c axes. First iteration- Do a maximum likelihood fit-trivial $s^*=c$.



Iterative solution

Iterate the posterior densities first

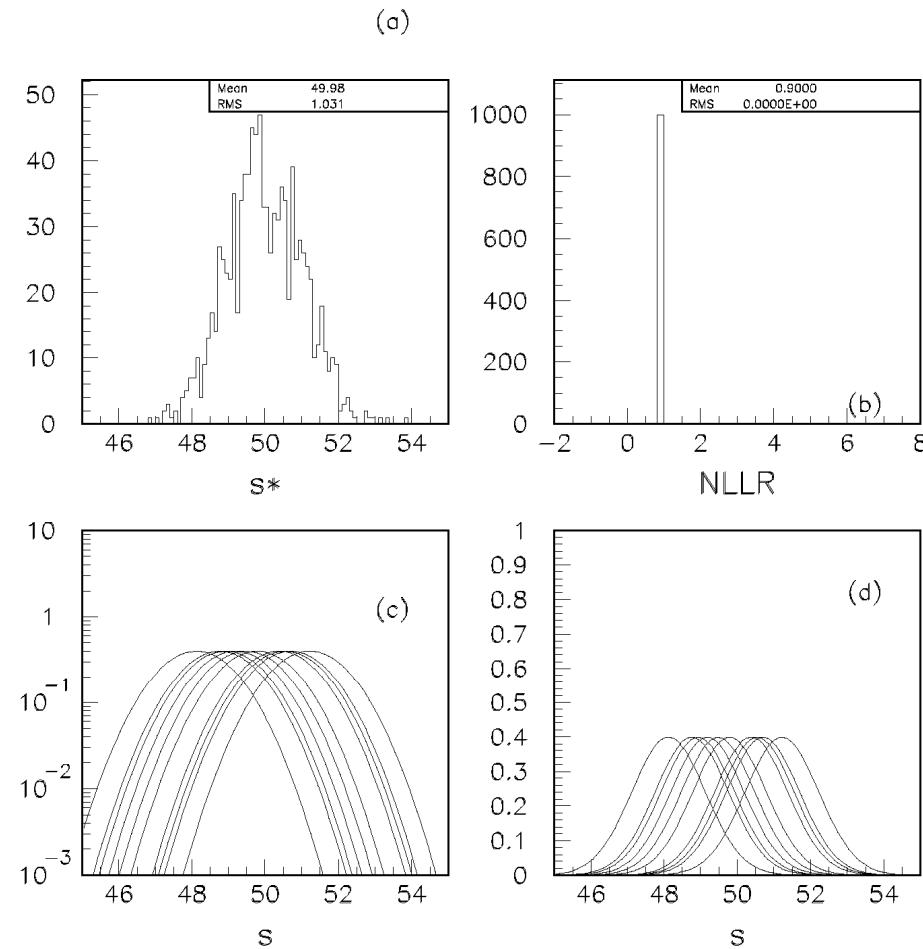
$$P_k^{iter}(s^* | \vec{c}_n) = \frac{P_k(\vec{c}_n | s^*) P_n(s^*)}{\int P(\vec{c}_n | s^*) P_n(s^*) ds^*}$$

$$P_n^{iter}(s^*) = \frac{1}{N} \sum_{k=1}^{k=N} P_k^{iter}(s^* | \vec{c}_n)$$

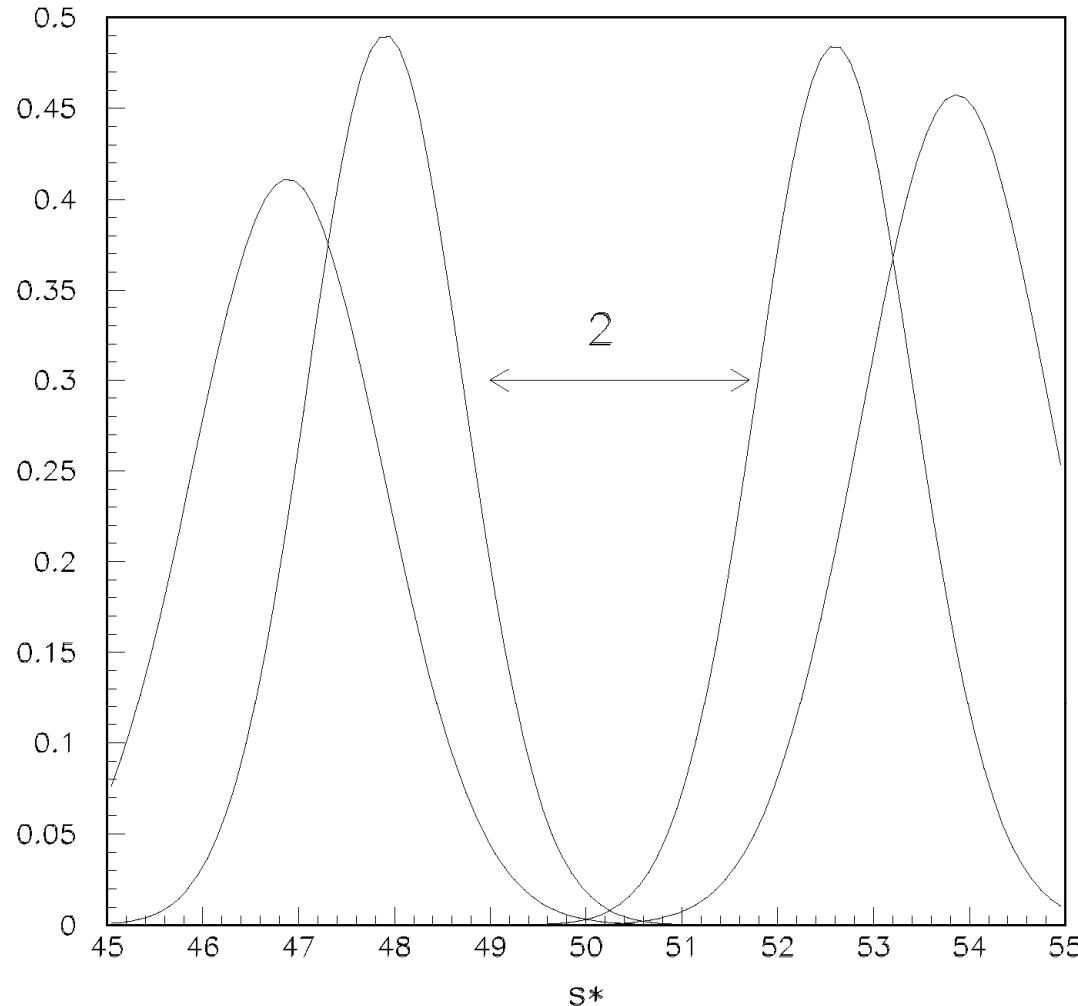
Similarly for the theoretical likelihood,

$$P_k^{iter}(\vec{c}_n | s^*) = \frac{P_k(s^* | \vec{c}_n) P(\vec{c}_n)}{\int P_k(s^* | \vec{c}_n) P(\vec{c}_n) d\vec{c}_n}$$

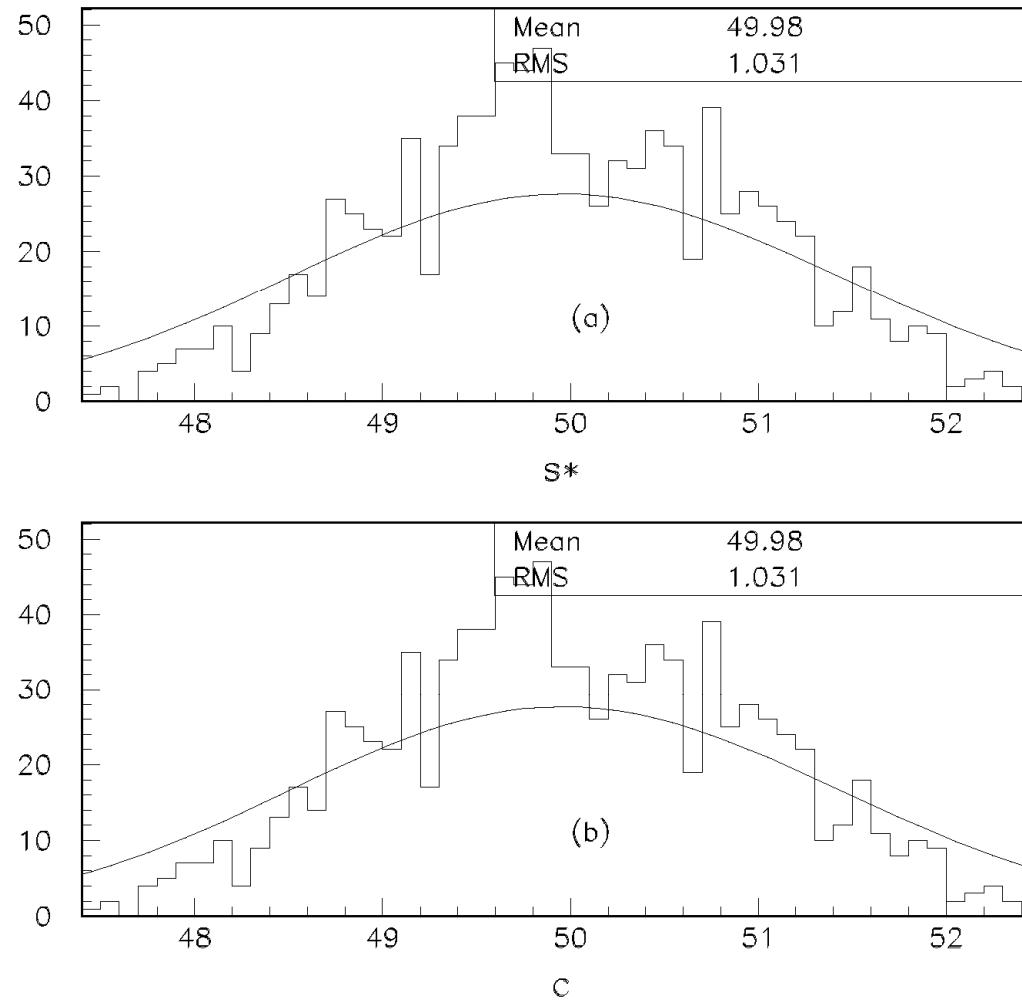
$$P^{iter}(\vec{c}_n) = \frac{1}{N} \sum_{k=1}^{k=N} P_k^{iter}(\vec{c}_n | s^*)$$



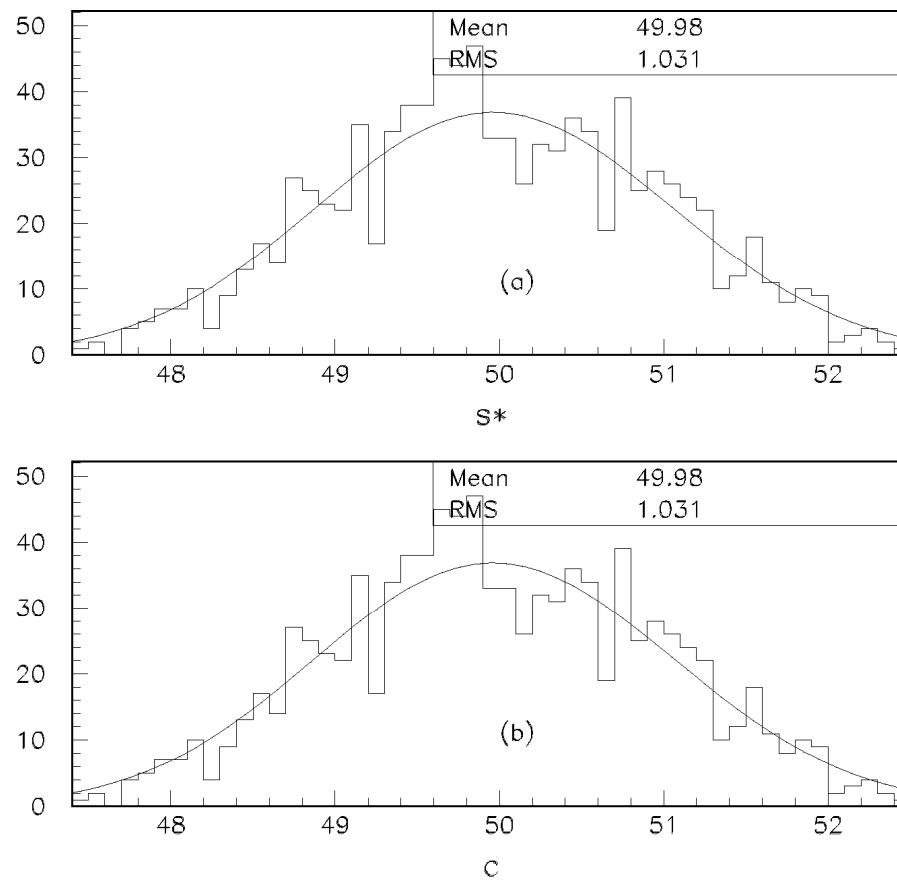
How is $P(s^* | \vec{c}_n)$ iteratively modified by Bayes' theorem



First iteration

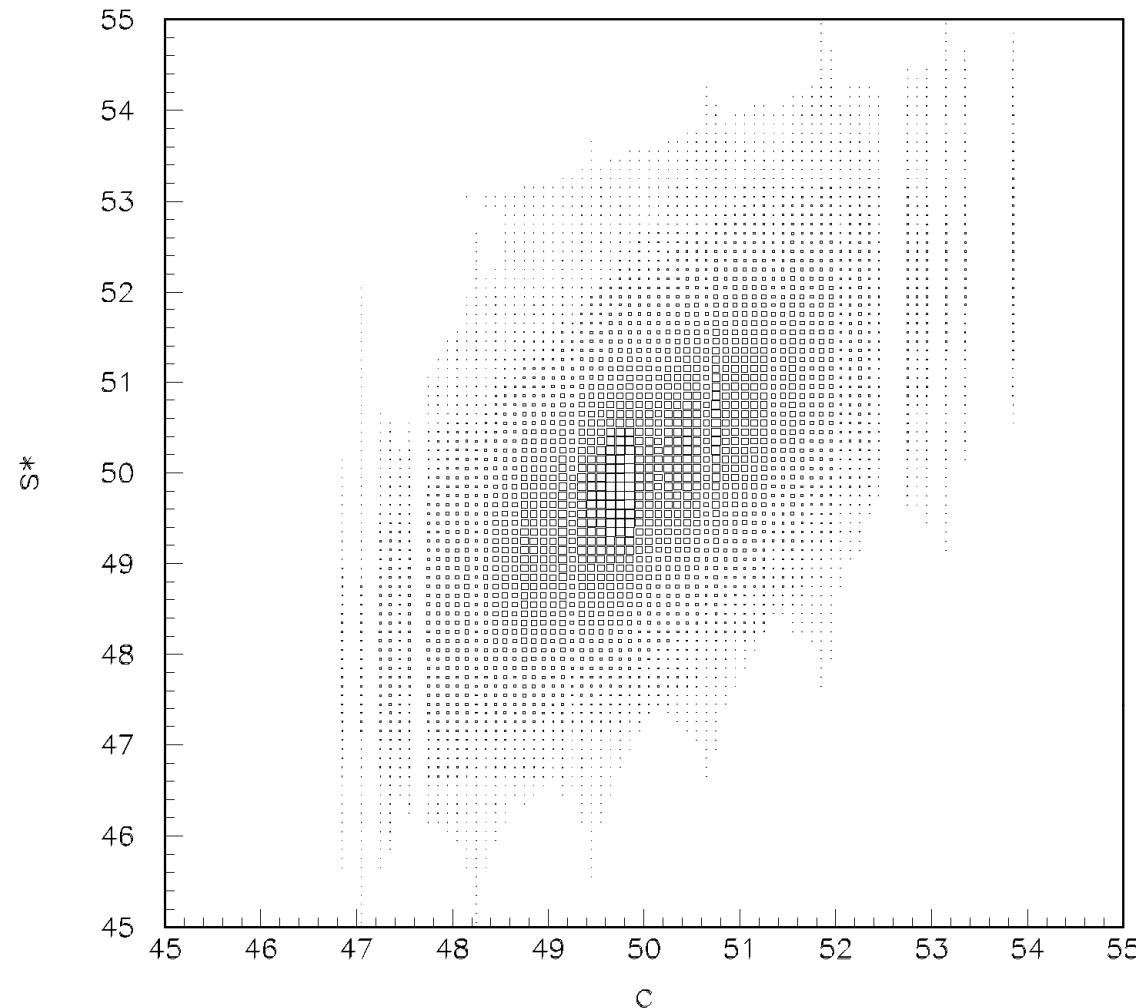


Second iteration



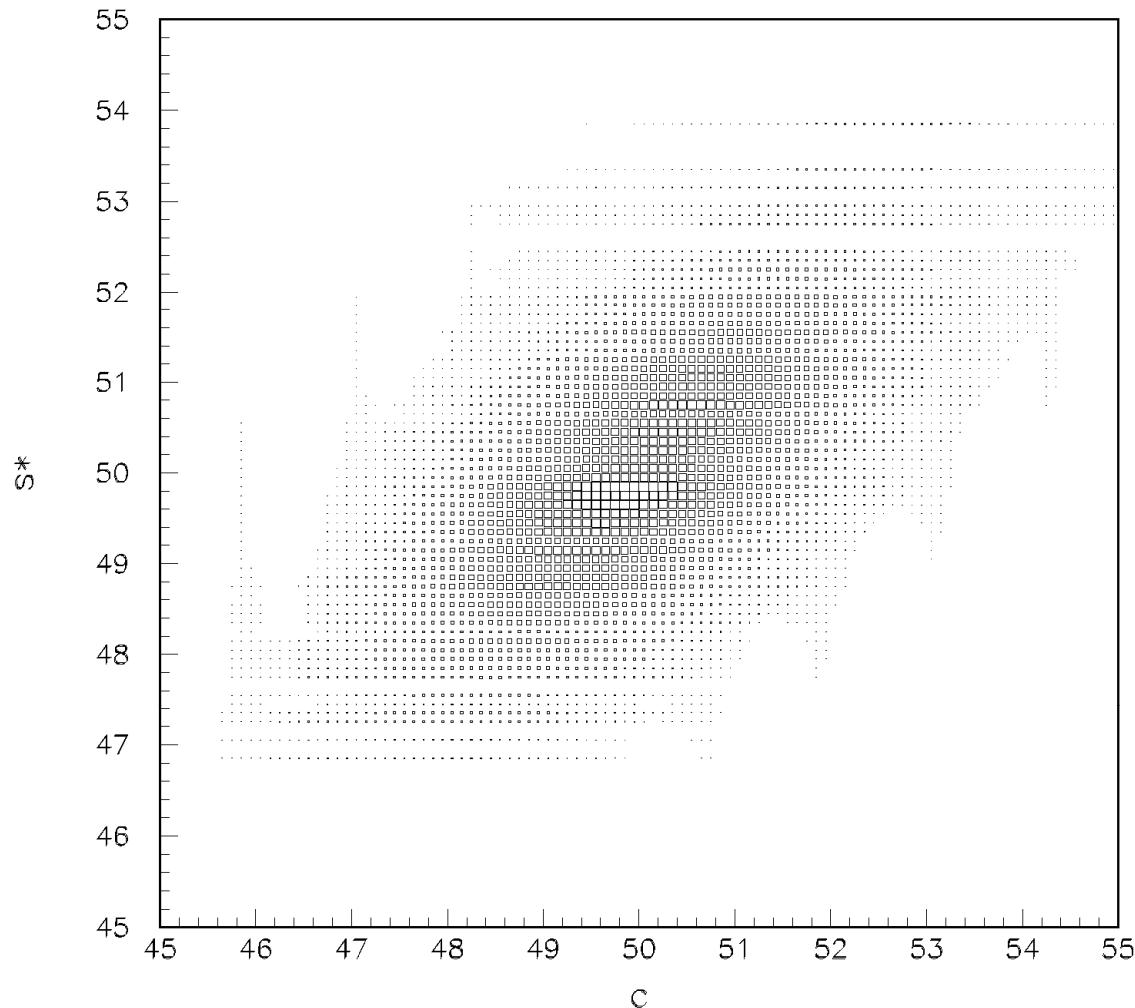
Joint probability

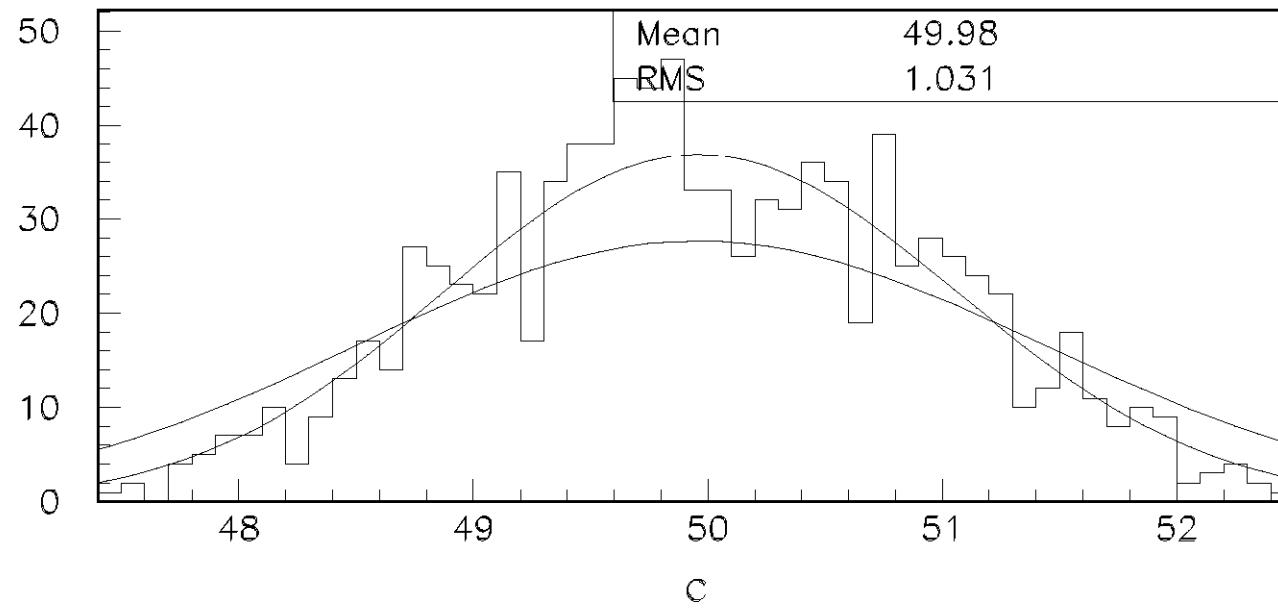
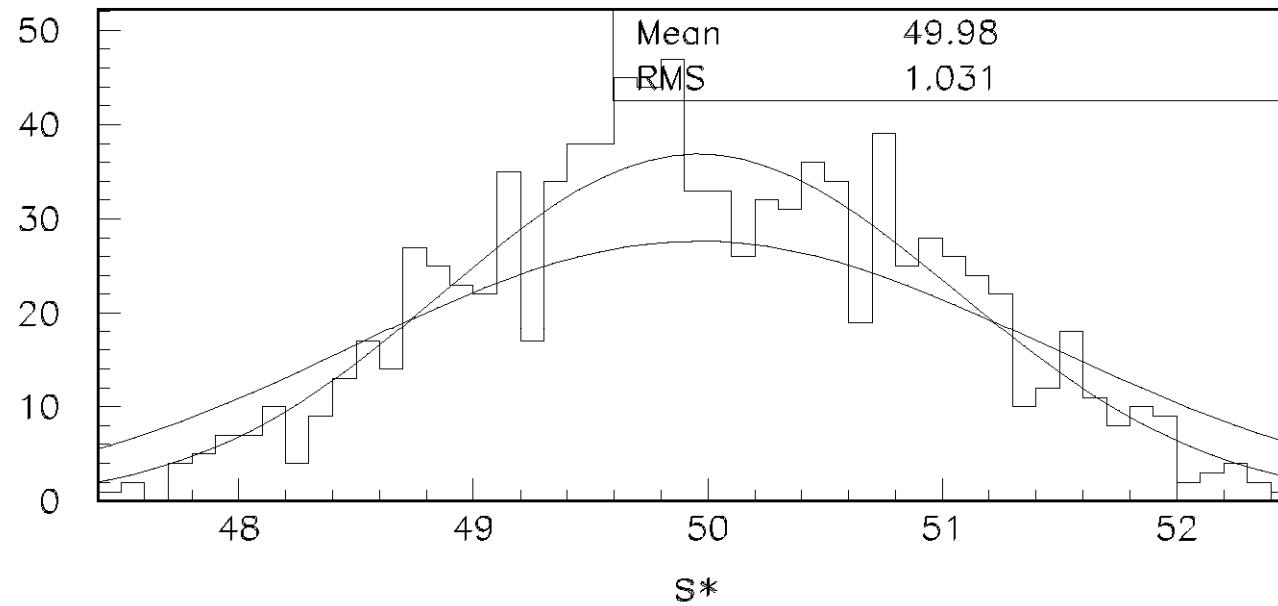
$$P(s^* | \vec{c}_n) P(\vec{c}_n)$$



Joint Probability

$$P(\vec{c}_n \mid s^*) P_n(s^*)$$





Bayes' theorem Equations Alternate forms valid for the new paradigm.

$$\frac{P(\vec{c}_n | s^*)}{P(\vec{c}_n)} = \frac{P(s^* | \vec{c}_n)}{P_n(s^*)}$$

$$\frac{P(\vec{c}_n | s^*)}{\int P(\vec{c}_n | s^{*'}) P_n(s^{*'}) ds^{*'}} = \frac{P(s^* | \vec{c}_n)}{\int P(s^* | \vec{c}_n) P(\vec{c}_n) d\vec{c}_n}$$

$$L_R^k(s^*) = \frac{P_k(\vec{c}_n | s^*)}{\langle P(\vec{c}_n | s^{*'}) \rangle} = \frac{P_k(s^* | \vec{c}_n)}{\langle P(s^* | \vec{c}_n) \rangle}$$

where $\langle \rangle$ denotes ensemble average

k is the ensemble member number

Also, the equation for $P_n(s^*)$ becomes an eigenfunction of PDE with a kernel. See long write-up

Conclusions

- We have given a general theory of goodness of fit that applies equally well to binned and unbinned likelihood fits. We have shown that the usual χ^2 is also a special case of this general theory.
- We have shown an iterative way to obtain posterior densities that does not require the Bayesian prior but uses Bayes' theorem.
- Method has applications to Tevatron and LHC analyses using multivariate techniques.
- Information from other experiments combined by multiplying likelihood ratios.