EP 4130/PH6130

Second and Third Week

Frequentist Parameter Estimation

Numerical Recipes 1992 edition Chapter 15
Bevington
arXiv: 0712.3028 by L. Verde

`AstroMl` treatment of the above subject will be discussed later using more "formal" notations

- Estimate best fit parameters of a model (function) given data and associated errors.
- Error estimates on the parameters.
- Statistical Measure of Goodness of fit.

Maximum likelihood Estimation

- Maximize the probability of data given the parameters or the *likelihood* of the data given the parameters . This form of parameter estimation is called *maximum likelihood estimation*

Consider data points y_i with measurement errors σ_i which is independent and has a Gaussian distribution around the true model (or the function)

$$P \propto \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{y_i - y(x_i,\theta)}{\sigma_i}\right)^2\right]$$

To get maximum likelihood estimate $dP/d\theta = 0$

Least-Squares Minimization

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

Best-fit parameters are obtained by minimizing chi-square
For models that are linear in theta, probability distribution of $\chi 2$ is chi-square PDF with v= N-M degrees of freedom, where N is the number of data points and M is the number of free parameters.

A rule of thumb : $\chi 2 \sim$ v for a good fit (sometimes called “chi by eye”)

$< \chi 2 >$ = v and its standard deviation = sqrt(v)
Reduced chi-square = $\chi 2$ / v

Chi-Square G.O.F (Numerical Recipes Notation)

Probability that observed χ^2 for a correct model is less than a particular value $\hat{\chi}^2$
Is given by the cdf of the chi-square probability distribution

$$\mathcal{P}(\chi^2 < \hat{\chi}^2, \nu) = \mathcal{P}(\nu/2, \hat{\chi}^2/2) = \Gamma(\nu/2, \hat{\chi}^2/2)$$

where Γ is the incomplete gamma function. (Refer to Numerical Recipes Chap 6 & 15 for notation)

$$Q \equiv 1 - \mathcal{P}(\nu/2, \chi^2/2)$$

is the probability that the observed χ^2 should exceed a particular value $\hat{\chi}^2$ by chance

Q is a quantitative measure of goodness of fit of a model is sometimes called chi-square GOF.

This is called p-value

Chi-Square Probability and Goodness of Fit

Chi-square probability or Chi-square likelihood for chi-square=Q and DOF by k is given by

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{k/2-1} \exp(-Q/2)$$

p-value or chi-square goodness of fit $p = \int_{\chi^2}^{\infty} p(Q|k)dQ = 1 - \frac{\gamma(\nu/2, \chi^2/2)}{\Gamma(\nu/2)}$

Probability of finding $\chi 2_{\text{min}}$ as high as the one we got or higher if hypothesis is correct and is called p-value

Chi-square c.d.f given in terms of $P(\nu/2, \chi^2/2)$ (regularized Gamma function)

$$P(a,x) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} dt$$

Chi-Square Probability and GOF in Python

Chi-square likelihood in Python

```
stats.chi2(ν).pdf(χ2)
```

How to calculate p-value from chi-square (χ^2) and Degrees of freedom (ν) in python

```
p-value = 1-stats.chi2(ν).cdf(χ2)
```

OR

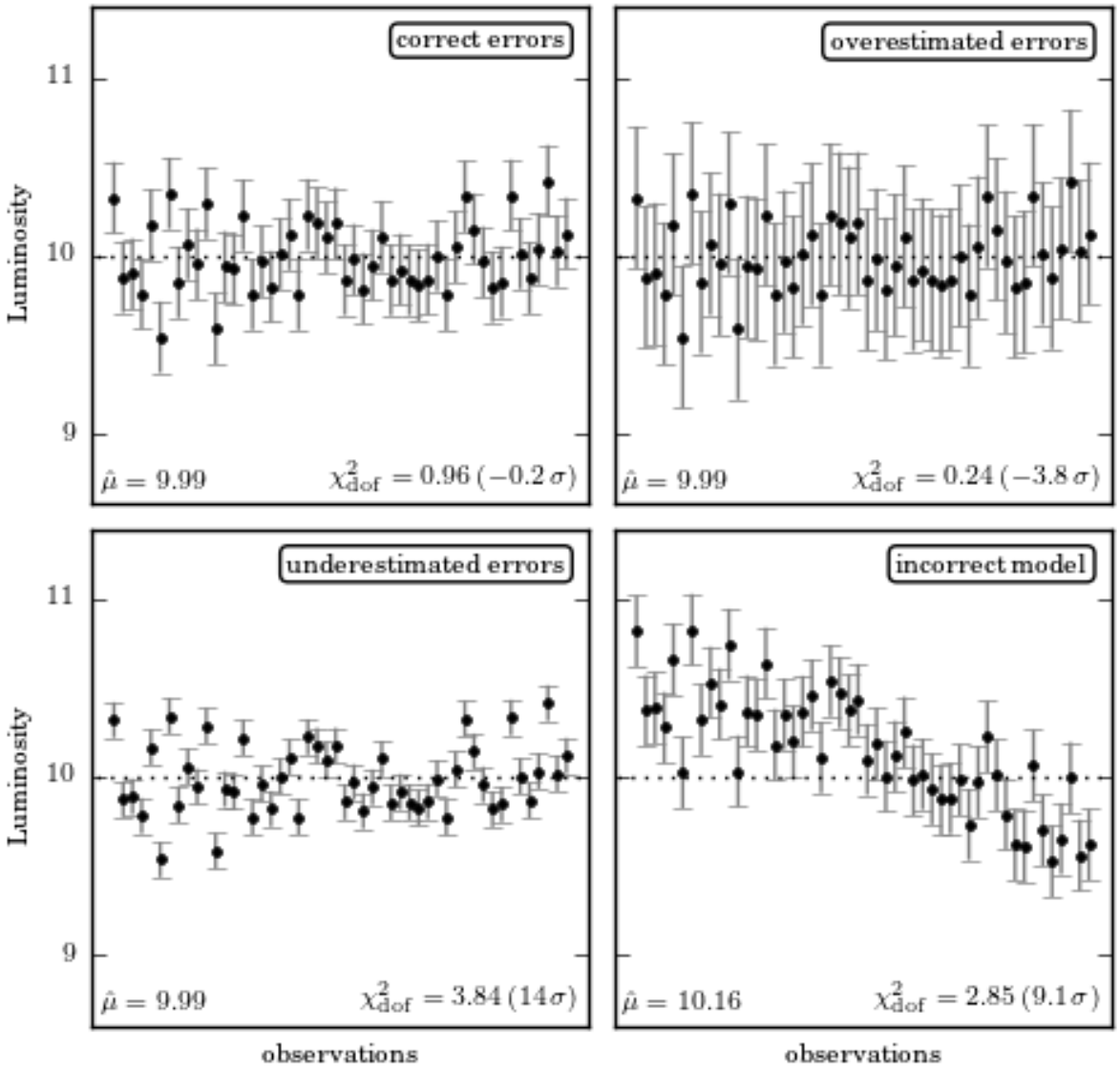
```
stats.chi2(ν).sf(χ2)
```

If p-value is a very small

- Model is wrong and can be rejected.
- Errors are really smaller than stated.
- Measurement errors are not Gaussianly distributed.

If p-value is a very large or close to 1

- Errors maybe overestimated
- Data are correlated and correlations are ignored in the fit
- Distribution is more compact than a Gaussian distributed. But this is almost never the case.



Astroml figure 4.1

No of sigmas = (reduced χ -1)/(Error in Reduced χ)

Error in Reduced χ = 1/√2(N-1)

Confidence Intervals

A *confidence region* (or *confidence interval*) is a region of M-dimensional space that contains a certain percentage (fraction) of the total probability distribution.

For eg. 99% confidence region implies that there is 99 % chance that the true parameter value falls within this region around the measured value. (Note that Bayesians use a different definition)

Usually in Physics/Astrophysics literature what is plotted in papers is 68.3%, 95.4% and 99.7% Confidence level.

These are obtained by contours of constant $\Delta\chi 2$ around the minimum $\chi 2$
There is formal mathematical relation between $\Delta\chi 2$, confidence intervals and standard errors.

Confidence Intervals (single parameter)

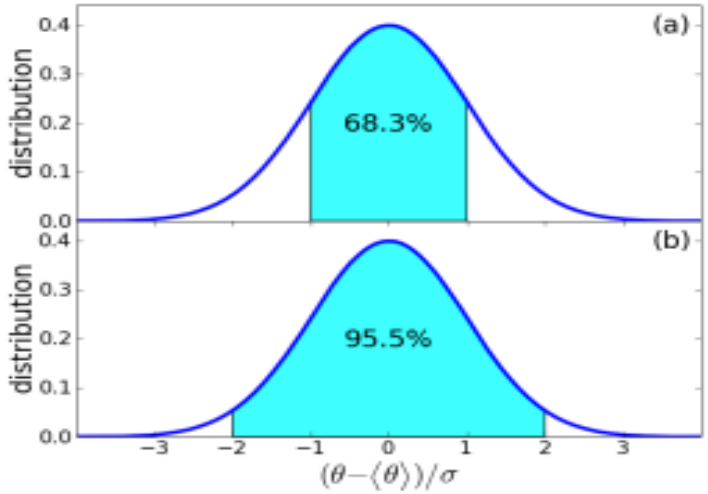


Figure 5: Confidence intervals for the Gaussian distribution of mean $\langle\theta\rangle$ and standard deviation σ. If we draw N values of θ from a Gaussian distribution, 68.3% of the values will be inside the interval $[\langle\theta\rangle - \sigma, \langle\theta\rangle + \sigma]$ as shown in panel (a), whereas 95.5% of the values will be inside the interval $[\langle\theta\rangle - 2\sigma, \langle\theta\rangle + 2\sigma]$ as shown in panel (b).

arXiv: 1009.2755

$$\mathrm{prob}(\theta_- \leq \hat{\theta} \leq \theta_+) = \int_{\theta_-}^{\theta_+} d\theta\, \mathrm{prob}(\theta) = C\,,$$

C.I. (single parameter, asymmetric distribution)

- Prev. equation does not uniquely define a C.I. in case of an asymmetric distribution.

1. Symmetric interval: θ_- and θ_+ are symmetric around the parameter estimate, i.e., $\hat{\theta}-\theta_- = \theta_+ - \hat{\theta}$.

2. Shortest interval: $\theta_+ - \theta_-$ is smallest for all intervals that satisfy Eq. (21).

3. Central interval: The probabilities above and below the interval are equal, i.e., $\int_{-\infty}^{\theta_-} d\theta\, \mathrm{prob}(\theta) = \int_{\theta_+}^{\infty} d\theta\, \mathrm{prob}(\theta) = (1-C)/2$.

Example of asymmetric likelihood

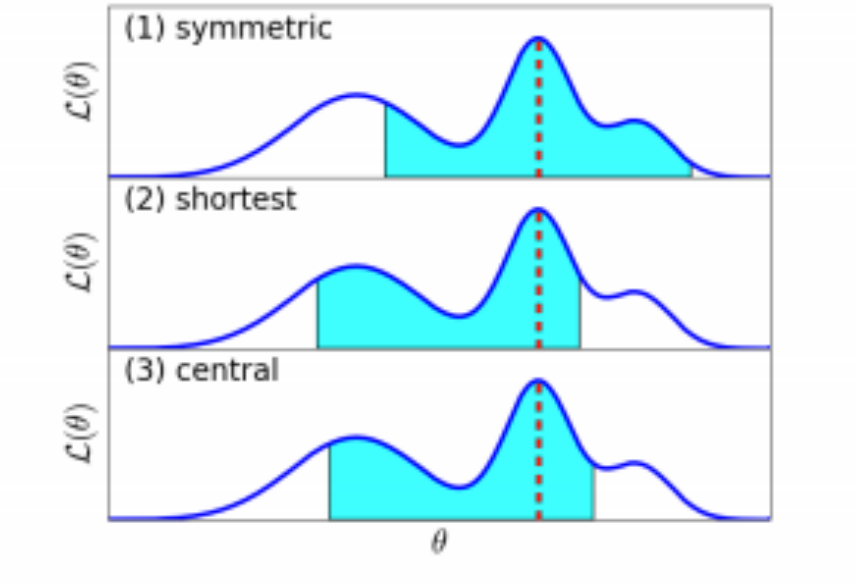


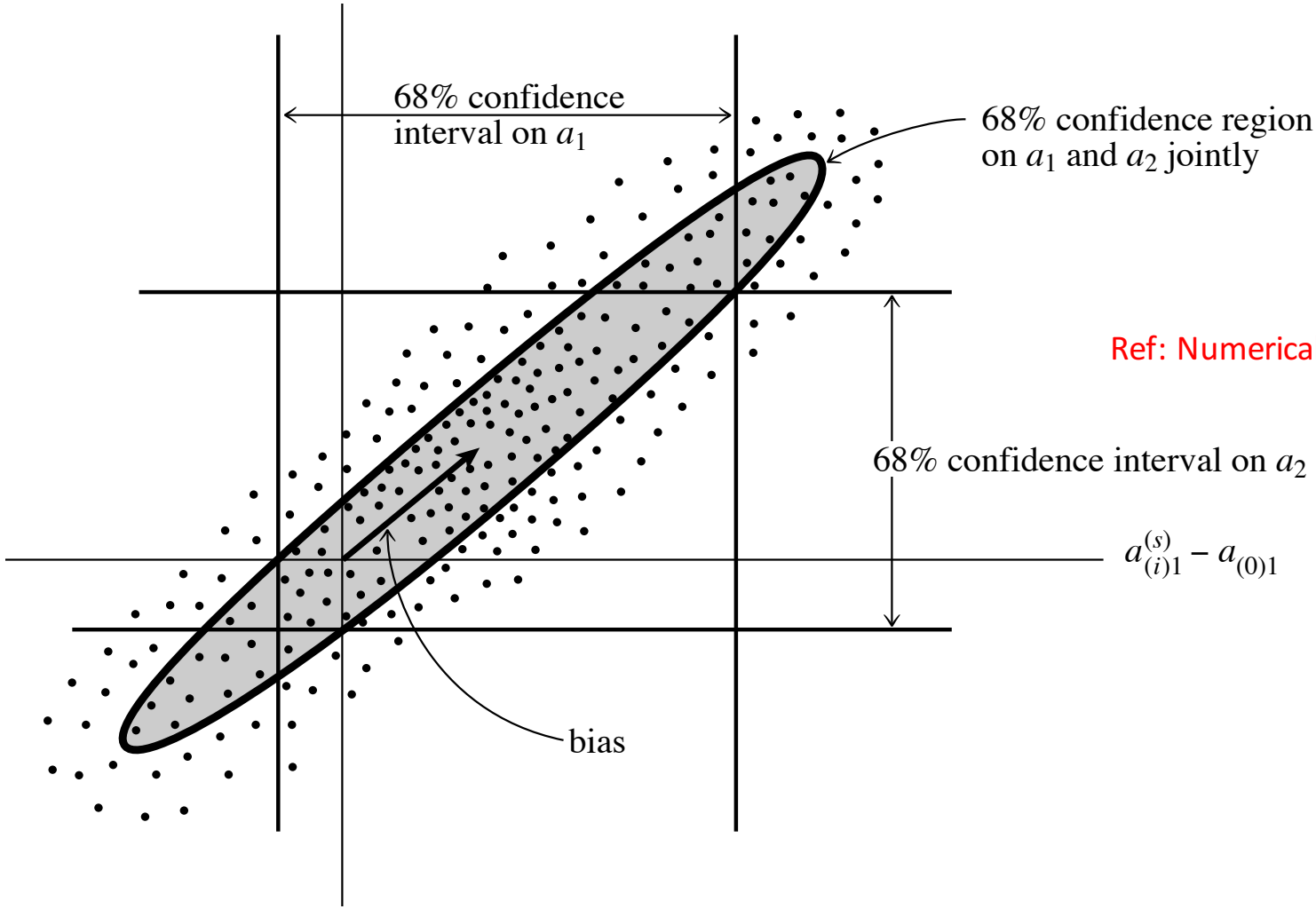
Figure 6: Different types of 68.3% confidence intervals for a multimodal likelihood function. The vertical dashed red line indicates the maximum-likelihood estimate $\hat{\theta}$. The panels are numbered according to the definitions in the main text.

Confidence Intervals in > 1 dimension

First, consider the case where the central-limit theorem indeed ensures that some (multidimensional) likelihood function is approximately Gaussian at its maximum position. Such (multivariate) Gaussians, with P-dimensional mean vector $\vec{\mu}$ and $P \times P$ covariance matrix $\mathbf{\Sigma}$,

$$\text{prob}(\vec{x}|\vec{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^P \det \mathbf{\Sigma}}} \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu})^T \cdot \mathbf{\Sigma}^{-1} \cdot (\vec{x} - \vec{\mu})\right] , \qquad (22)$$

One sigma contour of a 2-d gaussian marks a 39.4% confidence region.



Ref: Numerical Recipes 1992 ed . Chapter 15

Usually people plot the joint confidence region on both parameters of interest

Cookbook for Confidence Intervals

Numerical Recipes (1992) Sect 15.6

- Assume μ is the number of free parameters for which you want to plot the joint confidence interval and p% is the confidence limit desired .

- $\Delta\chi^2$ is distributed as a chi-square distribution with μ degrees of freedom. $\Delta\chi^2 = \chi^2 - \chi^2_{min}$

Calculate $\Delta\chi^2$ such that chi-square probability for μ free parameters is less than p.

$P(\chi^2 < \Delta\chi^2 , \mu) = p\%$

	ν					
p	1	2	3	4	5	6
68.3%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.4%	4.00	6.17	8.02	9.70	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.8

$\Delta\chi^2$ tables as a function of number of free parameters (p) and confidence level (%)

Numerical recipes Sect 15.6

Delta chi-square Intervals in python

```
>>> stats.chi2(2).cdf(2.3)
0.68336323062094673
```

Inverse of cdf in scipy is ppf

```
>>> stats.chi2(2).ppf(0.6833)
2.2996006508750853
```

Ref: Numerical Recipes 1992 ed . Chapter 15

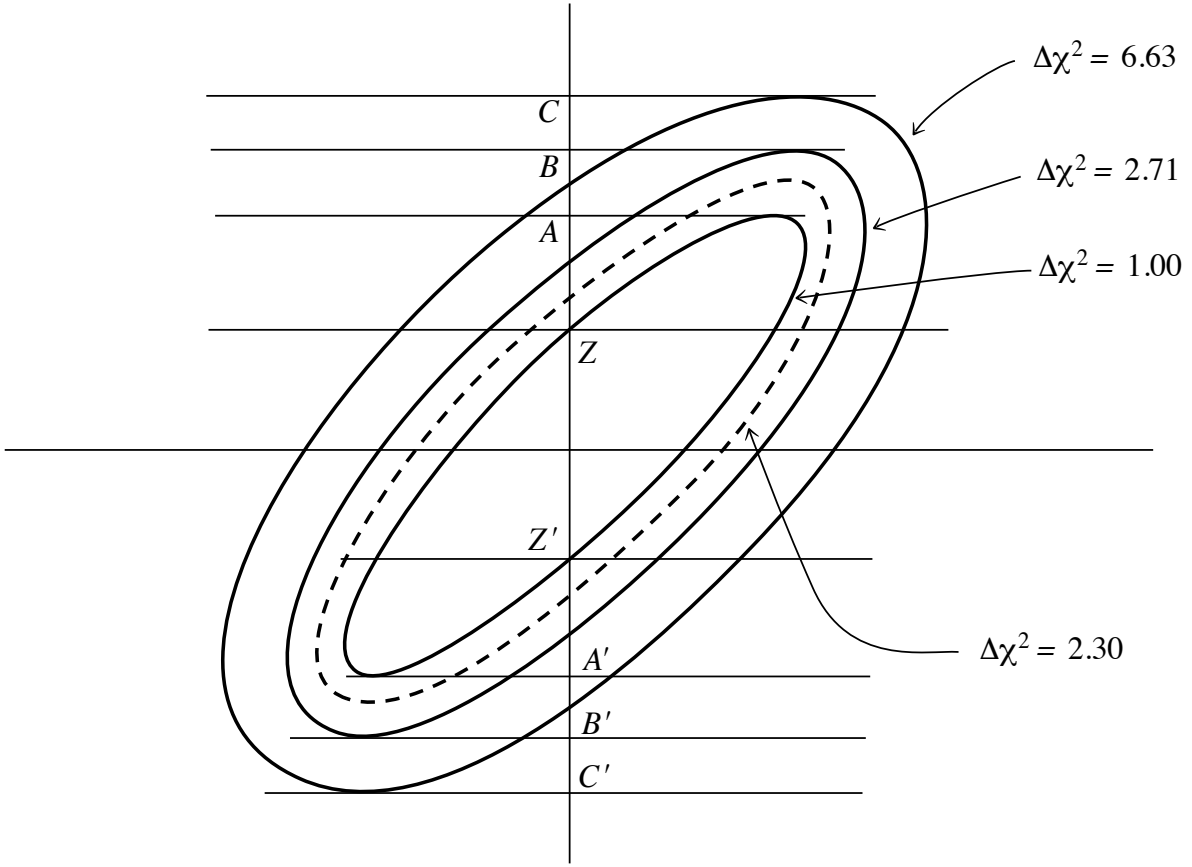


Figure 15.6.4. Confidence region ellipses corresponding to values of chi-square larger than the fitted minimum. The solid curves, with $\Delta\chi^2 = 1.00, 2.71, 6.63$ project onto one-dimensional intervals AA', BB', CC'. These intervals — not the ellipses themselves — contain 68.3%, 90%, and 99% of normally distributed data. The ellipse that contains 68.3% of normally distributed data is shown dashed, and has $\Delta\chi^2 = 2.30$. For additional numerical values, see accompanying table.

How to calculate projected 1-D errors

Numerical Recipes (1992) Sect 15.6

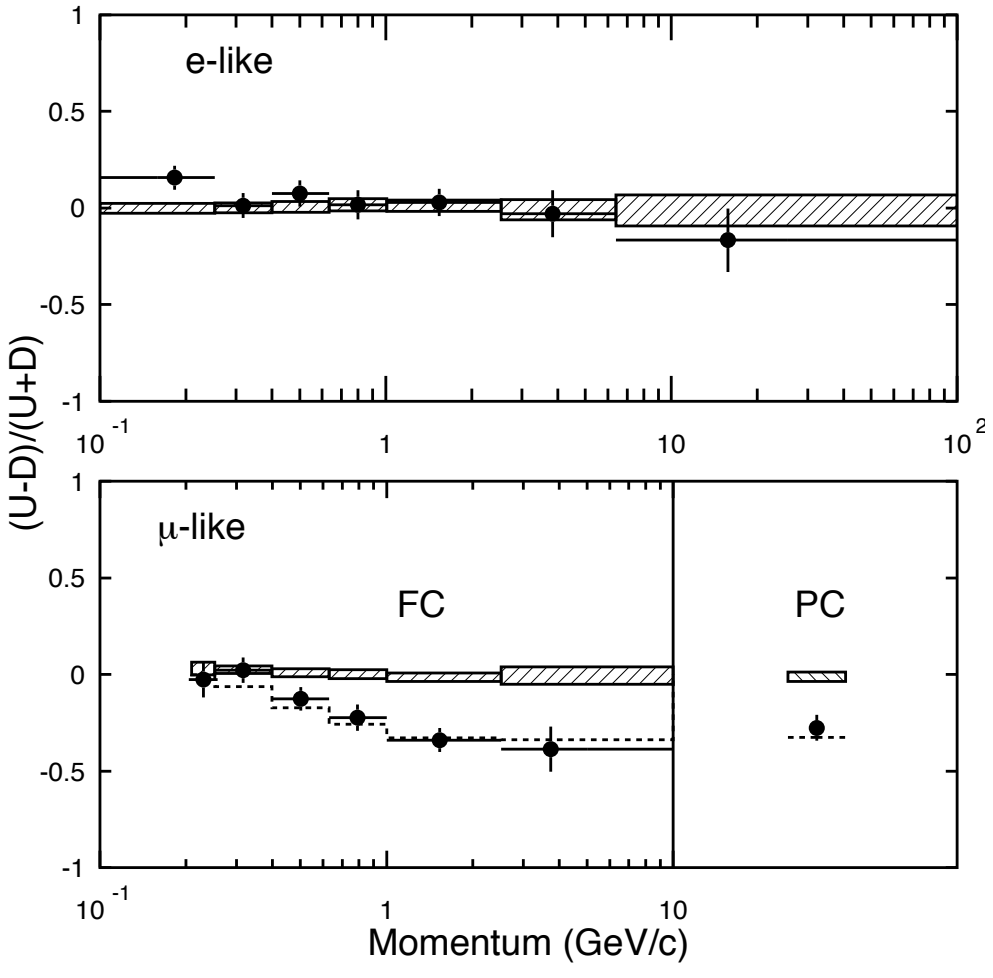
For one variable $$\delta a_1 = \pm\sqrt{\Delta\chi^2}\sqrt{C_{11}}$$

This gives relation between formal standard error $\boldsymbol{\sigma}_{11}$ = sqrt(C_{11}) and confidence interval (δa_1)

[C] = covariance matrix = Inverse of Fisher Information matrix or Inverse of curvature matrix
(to be defined later)

68% confidence interval ⟷ 1 **σ**
95% confidence interval ⟷ 2 **σ**
99% confidence interval ⟷ 3 **σ**

Note that this is valid only for 1 dimensions



hep-ex/9807003

$$\chi^2 = \sum_{cos\Theta,p} (N_{DATA} - N_{MC})^2/\sigma^2 + \sum_j \epsilon_j^2/\sigma_j^2,$$

Best-fit χ2 = 65.2 for 67 DOF

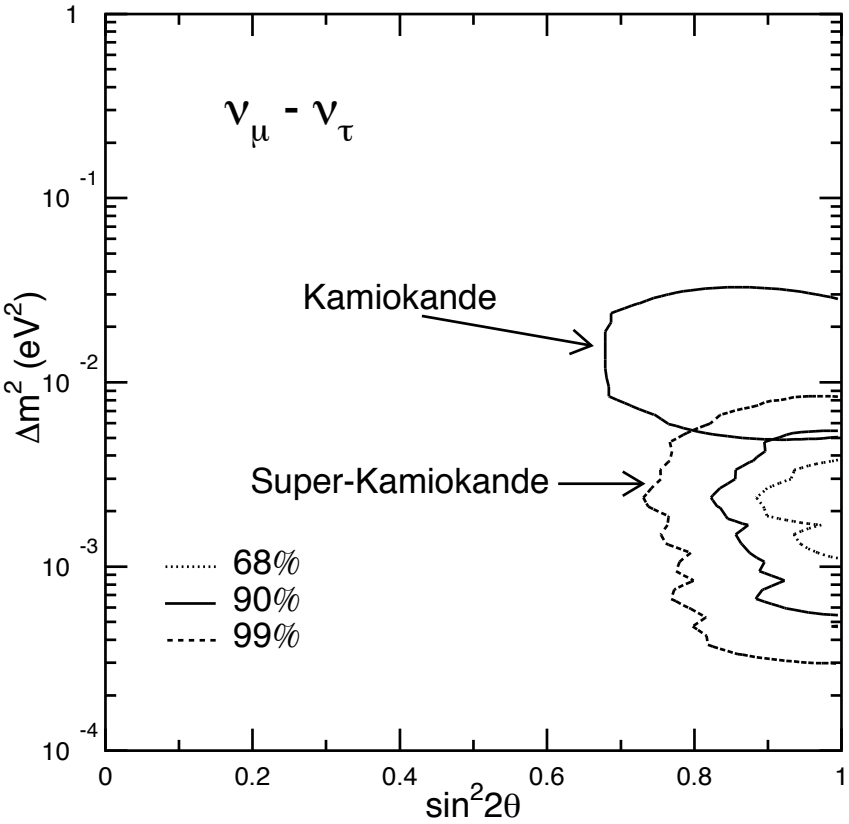
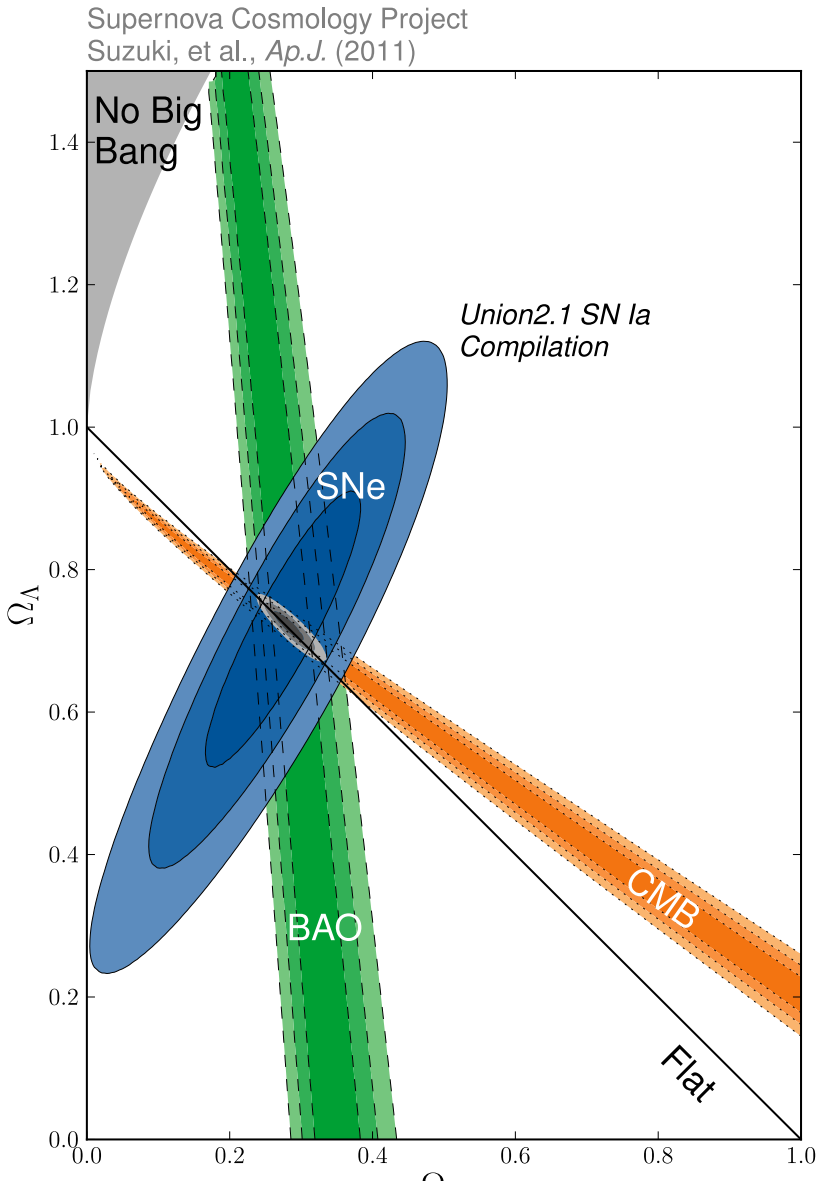


FIG. 2. The 68%, 90% and 99% confidence intervals are shown for $\sin^2 2\theta$ and Δm^2 for $\nu_\mu \leftrightarrow \nu_\tau$ two-neutrino oscillations based on 33.0 kiloton-years of Super–Kamiokande data. The 90% confidence interval obtained by the Kamiokande experiment is also shown.

hep-ex/9807003

$\Delta\chi 2$ intervals of 2.6,5.0, and 9.6 Correspond to 68%,90% and 99% confidence intervals

2015 Physics Nobel Prize



arXiv:1105.3470

68.3%, 95.4% and 99.7% confidence levels

$$\chi^2_{\rm stat} = \sum_{\rm SNe} \frac{\left[\mu_B(\alpha,\beta,\delta,M_{\rm B}) - \mu(z;\Omega_m,\Omega_w,w)\right]^2}{\sigma^2_{\rm lc} + \sigma^2_{\rm ext} + \sigma^2_{\rm sample}}.$$

2011 Physics Nobel Prize

Generalization of chi-square for Poisson data

- In case of bins with very low counts (or Poisson data) for comparing models with data we use :

$$C = 2\sum_{i=1}^{N}[f(x_i,\theta) - y_i + y_i \ln(y_i/f(x_i,\theta))]$$

Cash ApJ 229, 939 (1979)
Baker and Cousins, NIM 221, 437 (1984)

In astrophysical literature ONLY, C is called Cash statistics or C-stat. Based on C , one can calculate goodness of fit of a model in the same way as for a Gaussian distribution.

Also this can be used for model comparison as ΔC between two models of parameters also has a $\chi 2$ distribution with DOF = no of free parameters

Derivation of Cash statistics

- Obtained from likelihood ratio test. Ref. Baker and Cousins, NIM 221, 437 (1984)

Define $$\lambda = \frac{L_p(y,n)}{L_p(n,n)}$$ where n = observed data and y= expected model

where $$L_p(y,n) = \prod_i \exp(-y_i) y_i^n / n!$$

$$L_p(n,n) = \prod_i \exp(-n_i) n_i^{n_i} / n_i!$$

$\chi^2 = -2 \ln \lambda$ asymptotically obeys chi-square distribution (Wilk's theorem)

and evaluates to the definition of Cash statistics on previous slide

THE ASTROPHYSICAL JOURNAL, 228:939–947, 1979 March 15

PARAMETER ESTIMATION IN ASTRONOMY THROUGH APPLICATION OF THE LIKELIHOOD RATIO

WEBSTER CASH
Space Sciences Laboratory, Department of Physics, University of California, Berkeley
Received 1977 August 22; accepted 1978 August 24

ABSTRACT

Many problems in the experimental estimation of parameters for models can be solved through use of the likelihood ratio test. Applications of the likelihood ratio, with particular attention to photon counting experiments, are discussed. The procedures presented solve a greater range of problems than those currently in use, yet are no more difficult to apply. The procedures are proved analytically, and examples from current problems in astronomy are discussed.

Subject heading: functions: numerical methods

Recent exposition on Cash statistic

2.1. *The method of maximum likelihood and the C statistic*

The N Poisson data points D_i are assumed to be measurements from a parent model that describes the properties of the source. Models can be either fully specified with no free parameters, or more commonly featuring a number of free parameters. The likelihood of the data with the model is

$$\mathcal{L} = \prod_{i=1}^{N} \frac{e^{-S_i} S_i^{D_i}}{D_i!} \tag{1}$$

where D_i is an integer number of counts (the i–th data point) and S_i the mean value of the model for that data point. It is convenient to calculate the logarithm of the likelihood,

$$\ln \mathcal{L} = \sum_{i=1}^{N} (-S_i + D_i \ln S_i - \ln D_i!)$$

and then define the *Cash* or C statistic as

$$C = -2 \ln \mathcal{L} - B = 2 \sum_{i=1}^{N} (S_i - D_i + D_i \ln(D_i/S_i)) = \sum_{i=1}^{N} C_i \tag{2}$$

where

$$B = 2 \sum_{i=1}^{N} (D_i - D_i \ln D_i + \ln D_i!)$$

and

$$C_i = 2 \left(S_i - D_i + D_i \ln(D_i/S_i)\right).$$

arXiv:1912.05444

Relation between Cash statistic and χ^2

it can be shown that the C statistic is approximately equal to the χ^2 statistic:

arXiv:1912.05444

$$C \simeq 2\sum_{i=1}^{N}\left(S_i - D_i - D_i\left(-\frac{d}{D_i} - \frac{1}{2}\left(\frac{d}{D_i}\right)^2\right)\right) =$$

$$2\sum_{i=1}^{N}\left(S_i - D_i + d + \frac{d^2}{2D_i}\right) = \sum_{i=1}^{N}\frac{(D_i - S_i)^2}{D_i} = \sum_{i=1}^{N}\frac{(D_i - S_i)^2}{S_i}\frac{S_i}{D_i}, \quad (3)$$

where, by definition, $S_i - D_i + d = 0$. Since $D_i \sim \text{Poiss}(S_i)$, an estimate of the deviation d is given by the standard deviation of the Poisson distribution, i. e., $|d| \simeq \sqrt{S_i}$, further approximated as $|d| \simeq \sqrt{D_i}$, as also suggested by [9]. Using this approximation, each term in Equation 3 differs from a χ^2 distribution by a factor

$$\frac{S_i}{D_i} = \left(1 - \frac{d}{D_i}\right) \simeq \left(1 \pm \frac{1}{\sqrt{D_i}}\right)$$

Example of usage of Cash statistics

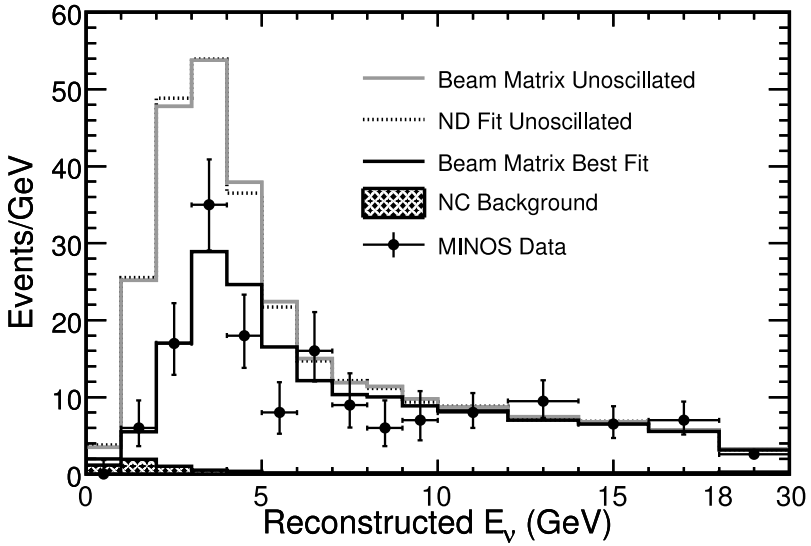


FIG. 3: Comparison of the Far Detector spectrum with predictions for no oscillations for both analysis methods and for oscillations with the best-fit parameters from the Beam Matrix extrapolation method. The estimated NC background is also shown. The last energy bin contains events between 18-30 GeV.

$$\chi^2 = \sum_{nbins} (2(e_i - o_i) + 2o_i \ln(o_i/e_i)) + \sum_{nsys} \frac{\Delta s_j^2}{\sigma_{s_j}^2}$$

hep-ex/0607088

Disconnect in usage of nomenclature

The jargon "Cash statistics" is used only in Astrophysics and NOT in particle physics

energy and the observed number of events in each bin is compared to the expected number of events for this oscillation hypothesis. The best fit parameters are those which minimize $\chi^2 = -2 \ln \lambda$ where λ is the likelihood ratio:

$$\chi^2 = \sum_{nbins} (2(e_i - o_i) + 2o_i \ln(o_i/e_i)) + \sum_{nsys} \frac{\Delta s_j^2}{\sigma_{s_j}^2} \quad (2)$$

where o_i and e_i are the observed and expected numbers of events in bin i, and the $\Delta s_j^2/\sigma_{s_j}^2$ are the penalty terms for nuisance parameters associated with the systematic uncertainties. The expected number of events depends

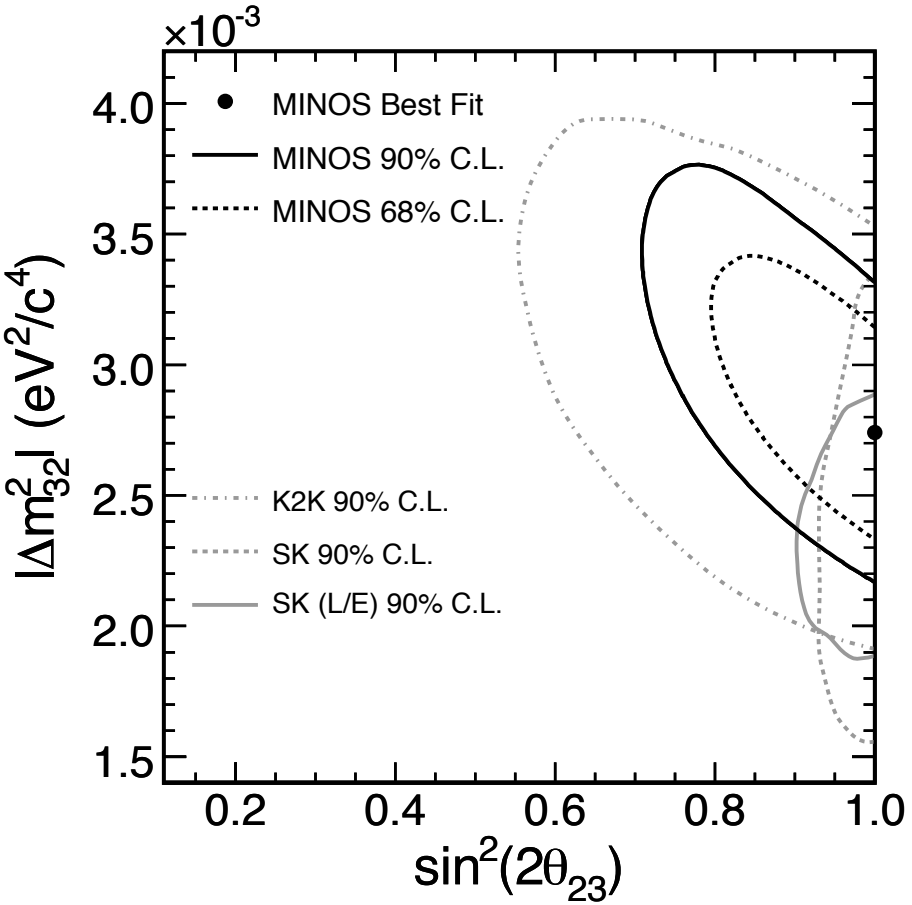
hep-ex/0607088 (Particle Physics)

certainties of $B(m)$ are small due to its being drawn from an area that is 25 times the area of the cluster. That is, the uncertainties of the cluster LF are dominated by the cluster field. We use the Cash (1979) statistic with a maximum likelihood estimator $C_{\rm stat}$ to estimate the parameters in the fitting.

$$C_{\rm stat} = 2\sum_j \left(M(m_j) - N(m_j) + N(m_j) \ln \left(\frac{N(m_j)}{M(m_j)} \right) \right), \quad (5)$$

where j runs over all the magnitude bins in the fit and N is the observed magnitude distribution for the cluster. Using the estimator $C_{\rm stat}$ allows us to estimate the goodness of fit (GOF) for the data following the Poisson distribution in the same way as a χ^2-distribution. The GOF of the LF fitting is defined by the ratio of the best-fit $C_{\rm stat}$ to the degrees of freedom (d.o.f) (i.e., GOF $= C_{\rm stat}$/d.o.f) and has a corresponding probability to exceed that provides information about tension between the best fit model and the data.

arXiv:1512.01244 (Astrophysics)



hep-ex/0607088 (based on minimization of Cash statistics)

χ^2 including covariances

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix} , \qquad \boldsymbol{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \\ 1 & x_N \end{bmatrix} ,$$

$$\boldsymbol{C} = \begin{bmatrix} \sigma_{y1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{y2}^2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_{yN}^2 \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X}$$

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - f(x_i)]^2}{\sigma_{yi}^2} \equiv [\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}]^{\top}\, \boldsymbol{C}^{-1}\, [\boldsymbol{Y} - \boldsymbol{A}\,\boldsymbol{X}]$$

(Definition of C above includes no-correlations in errors and hence off-diagonal elements are 0. Also A above is only true for a straight line fit of the form y=Ax+c)

More details in arXiv:1008.4686

MLE in case of Truncated/Censored Data

- Truncation : when S(x) = 0 for $x > x_{max}$ or $x < x_{min}$

- Censoring : Measurement of an existing source only resulted in *upper* limits

MLE in case of truncation becomes :

$$p(x_i|\mu,\sigma,x_{min},x_{max}) = C(\mu,x_{min},x_{max})\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where the normalization constant C is given by:

$$C(\mu,\sigma,x_{min},x_{max}) = (P(x_{max}|\mu,\sigma) - P(x_{min}|\mu,\sigma))^{-1}$$

Chi-square for errors in X-variable

- There is no uniform method to deal with this. Multiple methods used (BCES, ODR, etc) Some references:

  physics/051182 by Guido D'agostini

  astro-ph/9605002 by Akritas and Bershady

  arXiv:0705.2774 Also other techniques such as Orthogonal Distance Regression

Simplest recipe as follows: If σ_x and σ_y denote the error in X (independent) and Y (dependent) variables, then total error σ_t is given by

$$\sigma_t = \sqrt{\sigma_y^2 + \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2}$$

σ_t can be plugged in the expression for $\chi 2$ in the denominator

Classical Statistical Inference (astroML)

- **Point Estimation :** Best estimate for a model parameter based on available data

- **Confidence Estimation :** How confident should we be in our point estimate

- **Hypothesis Testing :** Are data consistent with a given hypothesis or model?

Two types of statistical paradigms to address the statistical inference questions : classical or *Frequentist* paradigm and the *Bayesian* paradigm.

Likelihood Function

- Starting point of both MLE and Bayesian estimation is *likelihood* of the data. The data likelihood represents a quantitative description of our measuring process also called P(D|M)

$$\mathcal{L} \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^{N} p(x_i|M(\theta))$$

where *M* stands for the model and θ is model parameter vector

- Although the interpretation of L is the probability of the data given the model, L is not a true properly normalized PDF. Likelihood of a single data point is a pdf. But the product of such functions is no longer normalized to 1.

- Likelihood is a function of both the data and the model.

Best-fit Model parameters which maximize P(D|M) are called point estimates.

Frequentist	Bayesian
Probabilities refer to relative frequencies of events	Refer to degree of subjective belief, not limiting frequency
Parameters are fixed unknown constants. Probability statements about parameters are meaningless	Probability statements can be made about things other than data including the model parameters and models themselves.
Frequentists consider model parameters to be fixed and data to be random	Bayesians consider data to be fixed and model parameters to be random
Confidence intervals should have well-defined long run frequency properties. 95% c.i. should bracket the true value of the parameter with a limiting frequency of at least 95%.	Inferences about a parameter are made by producing its probability. *Distribution quantifies amount of uncertainty of our knowledge about the parameter.* These are called credible intervals.

Bayesian data analysis often computationally intensive compared to frequentist analysis

Maximum Likelihood Estimators

- They are *consistent* estimators; they converge to the true parameter values as the number of data points increases.

- They are *asymptotically normal* estimators. The distribution of the parameter estimate, as the number of data points increase to infinity approaches a normal distribution centered on MLE with a certain spread.

- They asymptotically achieve the theoretical minimum possible variance called Cramer-Rao bound

(they achieve best possible error given the data at hand)

MLE Confidence Intervals

- We first compute the error matrix or Fisher information matrix.

$$F_{jk} = \frac{-\partial^2 \ln L}{\partial\theta_j \partial\theta_k}$$

- Inverse of the Fisher information matrix gives a lower bound on the variance of any unbiased estimator of θ . This is called Cramer-Rao bound.

$$\sigma_{jk} = \sqrt{[F^{-1}]_{jk}}$$

- Diagonal elements of σ correspond to marginal parameters for θ . IF non-diagonal elements are 0, inferred values of parameters are uncorrelated.

Jackknife, bootstrap, etc.

To estimate a parameter we have various tools such as maximum likelihood, least squares, etc.

Usually one also needs to know the variance (or the full sampling distribution) of the estimator – this can be more difficult.

Often use asymptotic properties, e.g., sampling distribution of ML estimators becomes Gaussian in large sample limit; std. dev. from curvature of log-likelihood at maximum.

The jackknife and bootstrap are examples of "resampling" methods used to estimate the sampling distribution of statistics.

In HEP we often do this implicitly by using Toy MC to determine sampling properties of statistics (e.g., Brazil plot for 1σ, 2σ bands of limits).

Glenn Cowan
Lecture notes

The Bootstrap (Efron, 1979)

Glenn Cowan lecture notes

Idea is to produce a set of "bootstrapped" data samples of same size as the original (real) one by sampling from some distribution that approximates the true (unknown) one.

By evaluating a statistic (such as an estimator for a parameter θ) with the bootstrapped-samples, properties of its sampling distribution (often its variance) can be estimated.

If the data consist of n events, one way to produce the bootstrapped samples is to randomly select from the original sample n events *with replacement* (the non-parametric bootstrap).

> That is, some events might get used multiple times, others might not get used at all.

In other cases could generate the bootstrapped samples from a parametric MC model, using parameter values estimated from real data in the MC (parametric bootstrap).

The Bootstrap (cont.)

Glenn Cowan lecture notes

Call the data sample $\boldsymbol{x} = (x_1, \ldots x_n)$, observed data are $\boldsymbol{x}_{\text{obs}}$, and the bootstrapped samples are $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$

Idea is to use the distribution of

$$\hat{\theta}(\mathbf{x}^*) - \hat{\theta}(\mathbf{x}_{\text{obs}})$$

as an approximation for the distribution of

$$\hat{\theta}(\mathbf{x}) - \theta$$

In the first quantity everything is known from the observed data plus bootstrapped samples, so we can use its distribution to estimate bias, variance, etc. of the estimator $\hat{\theta}$.

Bootstrap Resampling

- Bootstrap and Jackknife methods rely on resampling of the dataset $\{x_i\}$ (instead of generating them from probability distributions).
- We do not know h(x) [parent distribution of data] and the best we can do is computations relying on various estimates of h(x), based upon the data (called f(x)).

  Bootstrapping is based upon the approximation

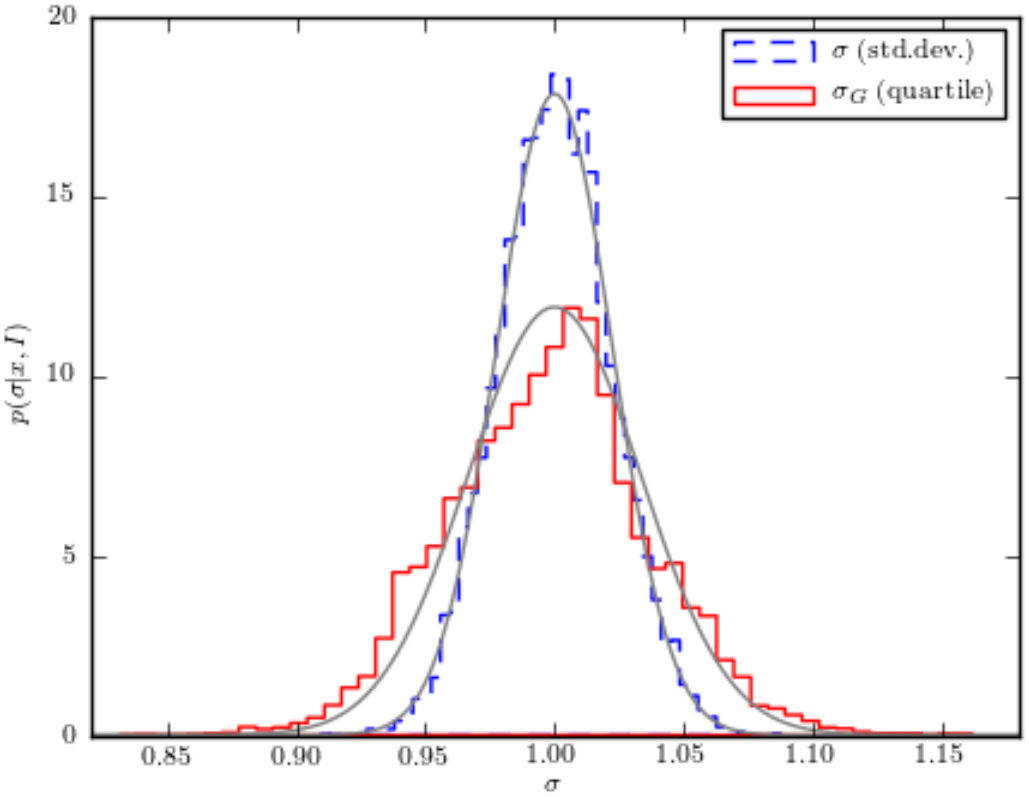
$$f(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i)$$

f(x) maximizes the probability of obtaining h(x) and is the MLE of h(x). Therefore, f(x) can be used as a proxy for h(x).

Basic idea of bootstrap is to draw new samples from the measured dataset itself. Drawing these ``bootstrap'' resamples can be done with replacement , i.e. same data point can occur multiple times in our data sample. (Non-parametric bootstrap).

For more details see Efron (1979); Lupton (1993); G.J. Babu lecture notes in PSU school

Example of Bootstrap Resampling



The bootstrap uncertainty estimates for the sample standard deviation σ (dashed line; see eq. 3.32) and σ_G (solid line; see eq. 3.36). The sample consists of N = 1000 values drawn from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. The bootstrap estimates are based on 10,000 samples. The thin lines show Gaussians with the widths determined as $s/\sqrt{2(N-1)}$ (eq. 3.35) for σ and $1.06s/\sqrt{N}$ (eq. 3.37) for σ_G.

$\sigma_G = 0.7413(q_{75} - q_{25})$

Formulae for
Standard error for σ and σ_G provided
In astroML book

The Jackknife

Glenn Cowan lecture notes

Invented by Quenouille (1949) and Tukey (1958).

Suppose data sample consists of n events: $\boldsymbol{x} = (x_1, \ldots x_n)$.

We have an estimator $\hat{\theta}(\boldsymbol{x})$ for a parameter θ.

Idea is to produce pseudo data samples $\boldsymbol{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots x_n)$
by leaving out the ith event.

See, e.g., Notes on Jackknife and Bootstrap by G. J. Babu:

www.iiap.res.in/astrostat/School10/LecFiles/
JBabu_JackknifeBootstrap_notes.pdf

Assume α is the jackknife estimate of the sample (with one data point removed) and α_N is the value estimated from the full dataset.

Bias-corrected Jackknife estimate of α is given by:

$$\alpha_J = \alpha_N + \Delta\alpha \tag{1}$$

$$\Delta\alpha = (N-1)(\alpha_N - \frac{1}{N}\sum_{i=1}^{N}\alpha_i^*)$$

Standard error for a jackknife estimator is given by:

$$\sigma_\alpha = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}[N\alpha_N - (N-1)\alpha_i^*]^2} \tag{2}$$

- Confidence Limits for α can be computed using Student's t-distribution with t defined as : $t = (\alpha - \alpha_J)/\sigma_\alpha$ with N-1 degrees of freedom.

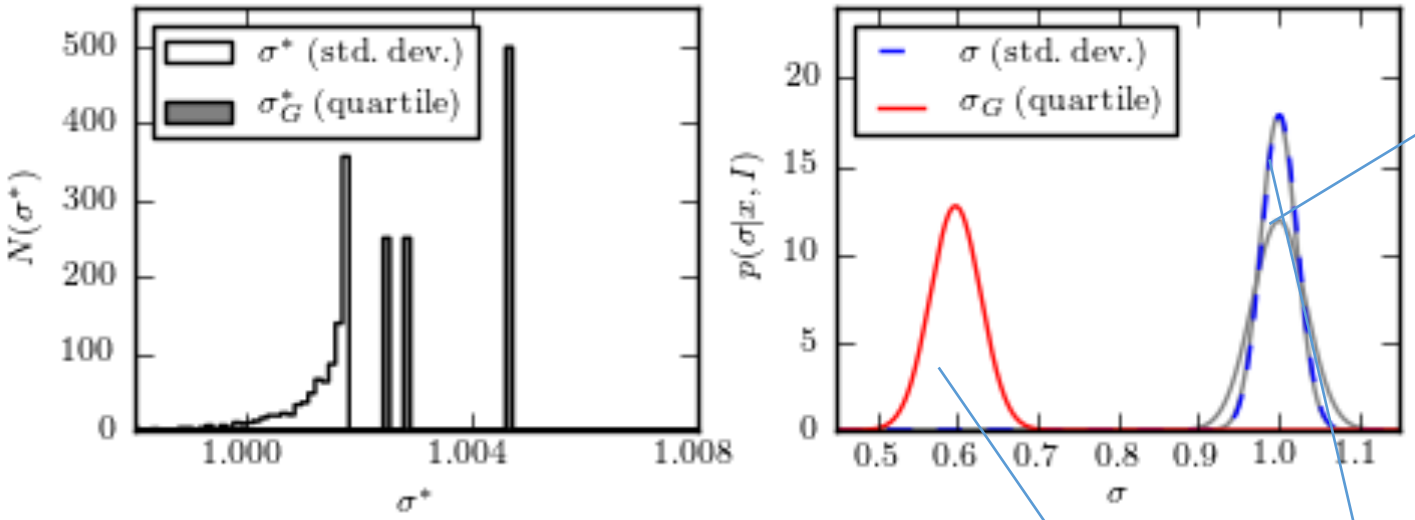
- Jackknife standard error is more reliable than jackknife bias correction.

- Jackknife bias correction completely fails for any robust statistics (median, quantiles, rank-based statistics.

Qt: Should one use bootstrap or jacknife?

For smooth statistics they produce similar results. Bootstrap better for calculating confidence intervals

In Machine learning , cross-validation and aggregating are closely related to jackknife and bootstrap



The jackknife uncertainty estimates for the width of a Gaussian distribution. This example uses the same data as figure 4.3. The upper panel shows a histogram of the widths determined using the sample standard deviation, and using the interquartile range. The lower panel shows the corrected jackknife estimates (eqs. 4.33 and 4.35) for the two methods. The gray lines show the theoretical results, given by eq. 3.35 for σ and eq. 3.37 for σ_G. The result for σ matches the theoretical result almost exactly, but note the failure of the jackknife to correctly estimate σ_G (see the text for a discussion of this result).

Expected distribution for $\boldsymbol{\sigma}_G$

Computed using (1) and (2)

Pdf from jacknife mean and sigma for sigma

Pdf from jacknife mean and sigma for sigma)G

Jackknife & Bootstrap Functions in AstroML

```
from astroML.resample import jackknife

[mean,stddev,raw_distribution] =
jackknife(data,user_statistic,kwargs,return_raw_distribution,pass_indices)

Example: mu1,sigma_mu1,mu1_raw =
jackknife(data,np.std,kwargs=dict(axis=1,ddof=1),return_raw_distrbution=True
)

astroML.resample.bootstrap(data,n_bootstraps,user_statistic,kwargs,pass_indi
ces,random_state)

Example: mu2_bootstrap = bootstrap(data,n,sigmaG,kwargs=dict(axis=1))
```

Lookup astroML.resample.jacknife and astroML.resample.bootstrap for full documentation