Appendix A

Statistics in HEP data analysis[1]

This Appendix introduces an overview of two aspects of statistical methods used in High Energy Physics (HEP)- parameter estimation and hypothesis testing. For the detail, it is recommended to refer to the relevant textbooks [1, 2, 3] and literatures [4], and references quoted in this Appendix. In an experiment of HEP, an observable x usually is a random variable, its *pdf* is expressed as $f(x,\theta)$ with parameter θ, what the experiment obtained is an sample of x:$\overrightarrow{x} = (x_1, ..., x_n)^T$. The task of the statistical inference is based on the sample of data to determine the value and error, or a confidence interval at given confidence level for the parameter θ, or infer observable's *pdf* $f(x,\theta)$.

A.1 Parameter estimation

Parameter θ is estimated with a function of sample of observed data: $\hat{\vartheta}(x_1, ..., x_n)$, which is called an estimator of θ. The sample of observed data $\overrightarrow{x} = (x_1, ..., x_n)^T$ is also a random variable, the value of the estimator to a specific measurement of $(x_1, ..., x_n)^T$ is called an estimate. Throughout this Appendix, we will use same donation to denote estimate and estimator. An good estimator should have properties of consistency, unbiasedness and high efficiency.

The consistency means when the size of the data sample $(x_1, ..., x_n)$ goes to infinity, the estimator $\hat{\theta}$ converges to the true value of parameter θ.

The bias of an estimator is defined as the difference of the expectation of the estimator and the true value θ: $E(\hat{\theta}) = \theta + b(\theta)$. The unbiasedness is a property of an estimator in finite sample, namely, it is required $E(\hat{\theta}) = \theta$. If it has to be estimated with a biased estimator, then the bias b of the estimator should be known or can be obtained by some way.

The efficiency is a measure of the variance of an estimator. Under the regularity conditions, namely, if the range of $\vec{x}$ is independent of θ and the first and second derivatives of the sample's joint *pdf* - Likelihood function $L(\vec{x}|\theta) = \prod_{i=1}^{n} f(x_i,\theta)$ - with respect to θ exist, there exists a lower bound on the variance of the estimates derived from an estimator, which is called the minimum variance bound MVB, given by Cramer-Rao

[1] By Yong-Sheng Zhu

inequality:

$$MVB = \frac{(1 + \frac{\partial b}{\partial \theta})}{I(\theta)}, \tag{A.1}$$

where $I(\theta)$ is the Fisher information:

$$I(\theta) = E[(\frac{\partial lnL}{\partial \theta})^2] = \int (\frac{\partial lnL}{\partial \theta})^2 \cdot Ld\vec{x} = E[-\frac{\partial^2 lnL}{\partial \theta^2}] = \int (-\frac{\partial^2 lnL}{\partial \theta^2}) \cdot Ld\vec{x}. \tag{A.2}$$

The efficiency of an estimator $\hat{\theta}$ is defined as $e(\hat{\theta}) = MVB/V(\hat{\theta})$. Apparently, we hope the efficiency of the used estimator is close or equal to 1.

The mean-square error (MSE) of an estimator is a convenient quantity which combine the uncertainties in an estimator due to bias and variance:

$$MSE = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + b^2. \tag{A.3}$$

A.1.1 Estimators for mean and variance

Suppose we are interested in the expectation μ of an observable x (random variable) and its variance, σ^2. We have a set of n independent measurements x_i, which have same unknown expectation μ and common unknown variance σ^2. This corresponds to, for instance, a set of n measurements for an observable x in an experiment. Then their consistent and unbiased estimate are the sample mean $\bar{x}$ and sample variance S^2, respectively:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{A.4}$$

$$\hat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{A.5}$$

The variance of $\hat{\mu}$ is σ^2/n, while the variance of $\hat{\sigma^2}$ is

$$V(\hat{\sigma^2}) = \frac{1}{n}(m_4 - \frac{n-3}{n-1}\sigma^4), \tag{A.6}$$

where m_4 is the 4th central moment of x.

For the known μ, the consistent, unbiased estimator of variance is

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{A.7}$$

which gives a somewhat better estimator of σ^2 compared with Eq. A.5 for unknown μ case. For the binomial, Poisson and Gaussian variables x_i, which are often used in data analysis, the sample mean is an efficient estimator for μ; For the normal variables, Eq. A.7 is an efficient estimator of σ^2 in the case of known μ, and sample variance S^2 is an asymptotic efficient estimator for σ^2.

For the Gaussian distributed x_i, Eq. A.6 becomes $V(\hat{\sigma^2}) = 2\sigma^4/(n-1)$ for any $n \geq 2$, and for large n the standard deviation of $\hat{\sigma}$ (the "error of the error") is $\sigma/\sqrt{2n}$.

If the x_i have different, known variance σ_i^2, which corresponds to the situation that different experiments measure the same quantity with different uncertainties. Assume x_i can be considered as a measurement of the Gaussian distributed variable $N(\mu, \sigma_i^2)$, then the unbiased estimator of the physics quantity μ is a weighted average

$$\hat{\mu} = \frac{1}{\omega} \sum_{i=1}^{n} \omega_i x_i, \tag{A.8}$$

where $\omega_i = 1/\sigma_i^2$, $\omega = \sum_i \omega_i$, and the standard deviation of $\hat{\mu}$ is $1/\sqrt{\omega}$.

A.1.2 The method of maximum likelihood(ML)

From the statistical point of view, the method of maximum likelihood (ML) is the most important general method of estimation, as the ML estimator of parameter has many good properties.

The ML estimators for parameter and its error

Suppose $x_i, i = 1, ..., n$ are the n independent measurements of a random variable x with the *pdf* $f(x, \vec{\theta})$, where $\vec{\theta} = (\theta_1, ..., \theta_k)^T$ are k parameters to be determined, then the ML estimators $\hat{\vec{\theta}}(x_1, ..., x_n)$ are the values of $\vec{\theta}$ that maximize the likelihood function

$$L(\vec{x}|\hat{\vec{\theta}}) = \prod_{i=1}^{n} f(x_i; \vec{\theta}). \tag{A.9}$$

Since both lnL and L are maximized for the same parameters values $\vec{\theta}$ and it is usually easier to work with lnL, therefore, the ML estimators can be found by solving the likelihood equations

$$\frac{\partial lnL}{\partial \theta_i} = 0, \qquad i = 1, ..., k. \tag{A.10}$$

The ML estimator is invariant under change of parameter, namely, under an one-to-one change of parameters from $\vec{\theta}$ to $\vec{\eta}$, the ML estimators $\hat{\vec{\theta}}$ transform to $\vec{\eta}(\vec{\theta})$: $\hat{\vec{\eta}}(\vec{\theta}) = \vec{\eta}(\hat{\vec{\theta}})$. Moreover, the ML estimators are asymptotic unbiased. When the likelihood function satisfies the regularity conditions, the ML estimators are consistent estimators. If there exist the efficient estimators for parameters or their functions, then the efficient estimators must be the ML estimators, and the likelihood equations give the unique solutions; while if the efficient estimators do not exist, the ML estimators give possibly minimum variance for $\vec{\theta}$. For large size n and the likelihood function satisfies the regularity conditions, $\hat{\vec{\theta}}$ asymptotically distributed as a normal variable with the mean being the true values $\vec{\theta}$ and the variances reach the MVB.

The ML estimators give only the values of the parameters. To know the errors of the parameters, one has to know the variances of parameters. The expression of the covariance between parameters $\hat{\theta}_i$ and $\hat{\theta}_j$ for any size of sample n is

$$V_{ij}(\hat{\vec{\theta}}) = \int (\hat{\theta}_i - \theta_i)(\hat{\theta}_j - \theta_j) L(\vec{x}|\vec{\theta}) d\vec{x}, \qquad i, j = 1, ..., k. \tag{A.11}$$

The calculation of this integral is sometimes troublesome, however, in general, it can be calculated with numerical method in any case.

For the case that $\hat{\theta}$ is an efficient estimator of single parameter, following equation is applicable for any size of sample n:

$$V(\hat{\theta}) = \frac{(1+\frac{\partial b}{\partial \theta})^2}{(-\frac{\partial^2 lnL}{\partial \theta^2})_{\theta=\hat{\theta}}}; \tag{A.12}$$

In particular, if $\hat{\theta}$ is an unbiased efficient estimator

$$V(\hat{\theta}) = \frac{1}{(-\frac{\partial^2 lnL}{\partial \theta^2})_{\theta=\hat{\theta}}}. \tag{A.13}$$

For multi-parameters and large n, if there exists a set of k jointly sufficient statistics $t_1, ..., t_k$ for the k parameters $\theta_1, ..., \theta_k$, the inverse of the covariance matrix $V_{ij} = cov(\hat{\theta}_i, \hat{\theta}_j)$ for a set of ML estimators can be calculated by

$$V_{ij}^{-1}(\hat{\vec{\theta}}) = (-\frac{\partial^2 lnL}{\partial \theta_i \partial \theta_j})_{\vec{\theta}=\hat{\vec{\theta}}}\ , \qquad i,j = 1, ..., k. \tag{A.14}$$

Besides, for large n and the likelihood function satisfying the regularity conditions, the ML estimators $\hat{\vec{\theta}}$ asymptotically distributed as a multi-dimensional normal variable, then one has

$$V_{ij}^{-1}(\hat{\vec{\theta}}) = E(-\frac{\partial^2 lnL(\vec{x}|\vec{\theta})}{\partial \theta_i \partial \theta_j})_{\vec{\theta}=\hat{\vec{\theta}}} = \int (-\frac{\partial^2 lnL(\vec{x}|\vec{\theta})}{\partial \theta_i \partial \theta_j})_{\vec{\theta}=\hat{\vec{\theta}}} \cdot L d\vec{x}, \qquad i,j = 1, ..., k; \tag{A.15}$$

or, one can use the *pdf* of the random variable x to calculate the covariance matrix:

$$V_{ij}^{-1}(\hat{\vec{\theta}}) = n \int \frac{1}{f}(\frac{\partial f}{\partial \theta_i})(\frac{\partial f}{\partial \theta_j}) dx, \qquad i,j = 1, ..., k. \tag{A.16}$$

Wherein, the last equation uses only the *pdf* of the random variable x and does not need the measured data sample, which is particularly useful in the design stage of an experiment.

If the observable x is a normal random variable, or the size of sample n is sufficiently large, then the likelihood function is an asymptotically normal distribution and lnL is a parabolic function, a numerically equivalent way of determining s-standard-deviation errors is from the contour given by the $\vec{\theta}'$ such that

$$lnL(\vec{\theta}') = lnL_{max} - \frac{s^2}{2}, \tag{A.17}$$

where L_{max} is the value of lnL at the solution point. The extreme limits of this contour on the θ_i axis give an s-standard-deviation likelihood interval for θ_i. In the case lnL is not a parabolic function, the approximate 1-standard-deviation likelihood interval can also be estimated by this equation, and it will give an asymmetric positive and negative errors for each parameter, namely, $\sigma^+(\theta_i) \neq \sigma^-(\theta_i), i = 1, ..., k$.

The ML method for binned data

In the case that the size n of data sample $\vec{x} = (x_1, ..., x_n)^T$ is sufficiently large, the measurements often are expressed as a histogram binned data. For constant n, the likelihood function (joint *pdf*) of $n_i (i = 1, ..., m)$ measurement values appearing in i-th bin is expressed as a multinomial distribution

$$L(n_1, ..., n_m|\vec{\theta}) = n! \prod_{i=1}^{m} \frac{1}{n_i!} p_i^{n_i}. \tag{A.18}$$

The probability of one measurement value appearing in i-th bin is calculated with *pdf* $f(x|\vec{\theta})$

$$p_i = p_i(\vec{\theta}) = \int_{\triangle x_i} f(x|\vec{\theta}) dx. \tag{A.19}$$

Then the likelihood equation becomes

$$(\frac{\partial lnL}{\partial \theta_i})_{\vec{\theta}=\hat{\vec{\theta}}} = \frac{\partial}{\partial \theta_i} [\sum_{i=1}^{m} n_i ln p_i(\vec{\theta})]_{\vec{\theta}=\hat{\vec{\theta}}} = 0, \qquad i = 1, ..., m, \tag{A.20}$$

Solving this set of equations gives the ML estimators $\hat{\vec{\theta}}$.

The extended ML method

If the size n of data sample is not a constant but a Poisson random variable with the expectation ν, then the likelihood function is the product of usual likelihood function and the Poisson probability of observing n events

$$L(\nu, \vec{\theta}) = \frac{\nu^n}{n!} e^{-\nu} \prod_{i=1}^{n} f(x_i, \vec{\theta}), \tag{A.21}$$

which is called the extended likelihood function [5]. Then the solutions of the likelihood equations

$$\frac{\partial lnL(\nu, \vec{\theta})}{\partial \theta_j} = 0, \qquad j = 1, ..., k, \tag{A.22}$$

$$\frac{\partial lnL(\nu, \vec{\theta})}{\partial \nu} = 0 \tag{A.23}$$

give the ML estimators $\hat{\vec{\theta}}$.

In the case that ν is irrelevant to $\vec{\theta}$, $\frac{\partial lnL(\nu,\vec{\theta})}{\partial \nu} = 0$ gives $\hat{\nu} = n$, the solutions of Eq. A.22 give the same $\hat{\vec{\theta}}$ as those from Eqs. A.9, A.10. If ν is a function of $\vec{\theta}$, the likelihood function becomes (dropping terms irrelevant to $\vec{\theta}$)

$$lnL(\vec{\theta}) = -\nu(\vec{\theta}) + \sum_{i=1}^{n} ln[\nu(\vec{\theta}) \cdot f(x_i, \vec{\theta})]. \tag{A.24}$$

The variances of the ML estimators $\hat{\vec{\theta}}$ derived from the extended likelihood function are usually smaller than those from usual likelihood function because the former uses the information from both n and $\vec{x}$.

For the binned data, the extended likelihood function is

$$L(n_1, ..., n_m|\vec{\theta}) = \prod_{i=1}^{m} \frac{1}{n_i!} \nu_i^{n_i} e^{-\nu_i}, \tag{A.25}$$

where the expectation of n_i, ν_i, is

$$\nu_i = \nu \int_{\triangle x_i} f(x|\vec{\theta}) dx, \qquad \nu = \sum_{i=1}^{m} \nu_i. \tag{A.26}$$

In the case that ν is irrelevant to $\vec{\theta}$, the likelihood equations become

$$\frac{\partial ln L}{\partial \theta_j}|_{\vec{\theta}=\hat{\vec{\theta}}} = \frac{\partial}{\partial \theta_j}[\sum_{i=1}^{m} n_i ln \nu_i]_{\vec{\theta}=\hat{\vec{\theta}}} = 0, \qquad j = 1, ..., m, \tag{A.27}$$

which has the same form of Eq. A.20 with $p_i(\vec{\theta})$ replaced by $\nu_i(\vec{\theta})$, and $\hat{\nu} = n$. If ν is a function of $\vec{\theta}$, then the likelihood equations are

$$\frac{\partial ln L}{\partial \theta_j}|_{\vec{\theta}=\hat{\vec{\theta}}} = \frac{\partial}{\partial \theta_j}[\sum_{i=1}^{m} n_i ln \nu_i - \nu]_{\vec{\theta}=\hat{\vec{\theta}}} = 0, \qquad j = 1, ..., m. \tag{A.28}$$

The variances of the ML estimators $\hat{\vec{\theta}}$ derived from these equations are usually smaller than those from usual likelihood function because the random property of n has been taken into account here.

Combining measurements with ML method

Suppose the observations in two independent experiments are $\vec{x} = (x_1, ..., x_l)^T$ and $\vec{y} = (y_1, ..., y_m)^T$, and their *pdf* $f_x(x, \vec{\theta})$ and $f_y(y, \vec{\theta})$, depend on same parameters $\vec{\theta}$, which are the quantities to be measured in the experiments. The joint likelihood function of these two experiments are

$$L(\vec{x}, \vec{y}; \vec{\theta}) = L(\vec{x}; \vec{\theta}) \cdot L(\vec{y}; \vec{\theta}) = \prod_{i=1}^{l} f_x(x_i, \vec{\theta}) \prod_{j=1}^{m} f_y(y_i, \vec{\theta}). \tag{A.29}$$

Solving the likelihood equations of this likelihood function with respect to parameters $\vec{\theta}$ and obtaining the ML estimator $\hat{\vec{\theta}}$ gives the combined measurement of these two experiments for parameters $\vec{\theta}$.

In the case $f_x(x, \theta)$ and $f_y(y, \theta)$ are Gaussians and the parameter θ is the mean of Gaussians, the combined estimator of the parameter and its variance have simple forms:

$$\hat{\theta} = (\frac{\theta_x}{\sigma_x^2} + \frac{\theta_y}{\sigma_y^2})/(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}), \tag{A.30}$$

$$V(\hat{\theta}) = 1/(\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}), \tag{A.31}$$

where θ_x and σ_x are the measured value of parameter θ and its error from experiment x, respectively. Above expressions can be directly extended to the situation of multi-experiments.

If the likelihood function is unknown, and only the results of parameter θ and its errors, $\theta_i, \sigma_i^+, \sigma_i^-$, are reported in each experiment, the combined results for parameter θ and its errors of multi-experiments can be deduced with the method suggested by R.Barlow [6]. The essence of the method is using the measured values $\theta_i, \sigma_i^+, \sigma_i^-$ to construct an approximate parametric likelihood function for each experiment. The *variable width Gaussians* are concluded as the best approximation for our purpose. The likelihood function can be approximated as

$$lnL(\theta_i|\theta) = -\frac{(\theta-\theta_i)^2}{2V_i(\theta)}, \tag{A.32}$$

where the true value of parameter is θ, and the measured value is θ_i in i-th experiment. For the *linear σ parametrization*, we have

$$V_i(\theta) = [\sigma_i(\theta)]^2, \qquad \sigma_i(\theta) = \sigma_i + \sigma_i'(\theta-\theta_i), \tag{A.33}$$

$$\sigma_i = \frac{2\sigma_i^+\sigma_i^-}{\sigma_i^+ + \sigma_i^-}, \qquad \sigma_i' = \frac{\sigma_i^+ - \sigma_i^-}{\sigma_i^+ + \sigma_i^-}, \tag{A.34}$$

where σ_i^+, σ_i^- are the measured positive and negative errors of θ_i in i-th experiment. For the *linear V parametrization*,

$$V_i(\theta) = V_i + V_i'(\theta-\theta_i), \tag{A.35}$$

$$V_i = \sigma_i^+\sigma_i^-, \qquad V_i' = \sigma_i^+ - \sigma_i^-. \tag{A.36}$$

Thus, the joint likelihood function of multi-experiments for parameter θ is approximately

$$lnL(\theta) = -\frac{1}{2}\sum_i \frac{(\theta-\theta_i)^2}{V_i(\theta)}. \tag{A.37}$$

The best estimate of θ, $\hat{\theta}$, is determined by the maximum of above likelihood function. For the linear σ form, the solution is

$$\hat{\theta} = \Sigma_i\omega_i\theta_i/\Sigma_i\omega_i, \tag{A.38}$$

$$\omega_i = \frac{\sigma_i}{[\sigma_i + \sigma_i'(\hat{\theta}-\theta_i)]^3}. \tag{A.39}$$

For the linear V form, the solution is

$$\hat{\theta} = \Sigma_i\omega_i[\theta_i - \frac{V_i'}{2V_i}(\hat{\theta}-\theta_i)^2]/\Sigma_i\omega_i, \tag{A.40}$$

$$\omega_i = \frac{V_i}{[V_i + V_i'(\hat{\theta}-\theta_i)]^2}. \tag{A.41}$$

Two sets of equations shown above are non-linear for $\hat{\theta}$ and the solution must be found by iteration. The $\triangle lnL$=0.5 points of the likelihood function in Eq. A.37 are used to determine the positive and negative errors for $\hat{\theta}$, which also need to be determined numerically. The program of combining results from multi-experiments using parametrization likelihood function has been coded, and obtainable under http://www.slac.stanford.edu/ barlow/statistics.html.

A.1.3 The method of least squares(LS)

The LS estimator for parameter and its error

Suppose n observations $\vec{y} = (y_1, ..., y_n)^T$ are measured at n points $\vec{x} = (x_1, ..., x_n)^T$, the covariance matrix of observations $\vec{y}$ is expressed as $V_{ij} = cov(y_i, y_j)$, and the true values of $\vec{y}$, $\vec{\eta}$, are described by model $\eta_i = f(x_i, \vec{\theta})$,$i = 1, ..., n$, where $\vec{\theta} = (\theta_1, ..., \theta_k)^T$ are the parameters to be determined. The least squares (LS) estimators of $\vec{\theta}$, $\hat{\vec{\theta}}$, can be found by minimizing the LS function $Q^2(\vec{\theta})$ with respect to $\vec{\theta}$:

$$Q^2(\vec{\theta}) = (\vec{y} - \vec{\eta}(\vec{\theta}))^T V^{-1} (\vec{y} - \vec{\eta}(\vec{\theta})) = \Sigma_{i=1}^n \Sigma_{j=1}^n (y_i - \eta_i) V_{ij}^{-1} (y_j - \eta_j). \tag{A.42}$$

In the case of $y_i, i = 1, ..., n$ being n independent measurements, the LS function has simple form

$$Q^2(\vec{\theta}) = \Sigma_{i=1}^n \frac{(y_i - \eta_i)^2}{\sigma_i^2}, \tag{A.43}$$

where σ_i is the error of y_i. An usual case is y_i is a Poisson variable, then σ_i^2 can be approximated by y_i or its predicted value η_i. If $y_i, i = 1, ..., n$ are n independent Gaussians, $y_i \sim N(\eta_i, \sigma_i^2)$, the likelihood function of $\vec{y}$ is $L(\vec{\theta}) \propto exp[-\frac{1}{2}\sum_{i=1}^n (\frac{y_i - \eta_i}{\sigma_i})^2]$. In this case maximizing $L(\vec{\theta})$ with respect to parameters $\vec{\theta}$ is equivalent to minimizing the LS function $Q^2(\vec{\theta}) = \Sigma_{i=1}^n \frac{(y_i - \eta_i)^2}{\sigma_i^2}$, namely, the estimators of ML and LS methods for $\vec{\theta}$ are identical.

For the linear LS model, *i.e.* $f(x_i, \vec{\theta})$ is the linear function of $\vec{\theta}$:

$$f(x_i, \vec{\theta}) = \Sigma_{j=1}^k a_{ij} \theta_j, \qquad i = 1, ..., n, \quad k < n, \tag{A.44}$$

where a_{ij} equals x_i^{j-1} or is the $(j-1)$-th Legendre polynomial of x_i, minimizing the LS function $Q^2(\vec{\theta})$ simplified to solve a set of k linear equations. Define a_{ij} being the elements of a $n \times k$ matrix A, minimizing the LS function $Q^2(\vec{\theta})$ gives the LS estimator of parameters $\vec{\theta}$:

$$\hat{\vec{\theta}} = (A^T V^{-1} A)^{-1} A^T V^{-1} \vec{y}, \tag{A.45}$$

the covariance matrix of $\hat{\vec{\theta}}$ is

$$V(\hat{\vec{\theta}}) = (A^T V^{-1} A)^{-1}, \tag{A.46}$$

or equivalently

$$(V^{-1}(\hat{\vec{\theta}}))_{ij} = \frac{1}{2} \frac{\partial^2 Q^2}{\partial \theta_i \partial \theta_j} |_{\vec{\theta} = \hat{\vec{\theta}}} = \Sigma_{l,m=1}^n a_{li} a_{mj} (V^{-1})_{lm}, \qquad i, j = 1, ..., k. \tag{A.47}$$

If $y_i, i = 1, ..., n$ are independent each other, the non-diagonal elements equal to zeros, then the above equation is simplified to

$$(V^{-1}(\hat{\vec{\theta}}))_{ij} = \Sigma_{m=1}^n a_{mi} a_{mj} / \sigma_m^2, \qquad i, j = 1, ..., k. \tag{A.48}$$

The linear LS estimators provide the exact solutions for parameters $\vec{\theta}$, and they are unique and unbiased, and have minimum variances.

Expanding $Q^2(\vec{\theta})$ about $\hat{\vec{\theta}}$, one finds that the contour in parameter space defined by

$$Q^2(\vec{\theta}) = Q^2(\hat{\vec{\theta}}) + s^2 = Q^2_{min} + s^2 \tag{A.49}$$

has tangent planes located at plus or minus s- standard deviation from the LS estimates $\hat{\vec{\theta}}$.

For the linear LS model, if the observations $\vec{y}$ are multi-normal variables, the minimum of the LS function $Q^2(\vec{\theta})$

$$Q^2_{min}(\vec{\theta}) = \Sigma^n_{i=1}\Sigma^n_{j=1}(y_i - \hat{\eta}_i)V^{-1}_{ij}(y_j - \hat{\eta}_j) \tag{A.50}$$

is a χ^2 variable with degree of freedom of $n - k$. This means the Q^2_{min} obtained by LS method is a quantitative measure of the consistency between the measured values $\vec{y}$ and their fitted value $\hat{\vec{\eta}}$, *i.e.*, the Q^2_{min} represents the goodness of fit (see section A.3.1., goodness of fit tests).

For the non-linear LS model, *i.e.* $f(x_i, \vec{\theta})$ is the non-linear function of $\vec{\theta}$, usually the minimizing of the LS function $Q^2(\vec{\theta})$ is implemented via iteration procedure to obtain an approximate solution of $\hat{\vec{\theta}}$. The non-linear LS estimator is a biased estimator, its variance does not reach MVB, and the exact distribution of Q^2_{min} is unknown. However, if n is sufficiently large, the LS estimator is asymptotically unbiased, and its Q^2_{min} is approximately a χ^2 variable.

The LS method for binned data

For sufficiently large size n of data sample $\vec{x} = (x_1, ..., x_n)^T$ and the measurements expressed as a histogram binned data, assuming the observed number of measurements in i-th bin is $n_i, i = 1, ..., m$, and its corresponding expectation from assumed model is

$$f_i(\vec{\theta}) = np_i, \qquad p_i(\vec{\theta}) = \int_{\triangle x_i} g(x|\vec{\theta})dx, \tag{A.51}$$

where $g(x|\vec{\theta})$ is the *pdf* of observable x and $\vec{\theta}$ are the parameters to be determined. The normalization $\Sigma^m_{i=1}p_i = 1$ requires

$$\Sigma^m_{i=1}n_i = \Sigma^m_{i=1}f_i(\vec{\theta}) = n. \tag{A.52}$$

It can be proved for a given n, the LS function $Q^2(\vec{\theta})$ is of the form

$$Q^2(\vec{\theta}) = \Sigma^m_{i=1}\frac{(n_i - np_i)^2}{np_i} = \Sigma^m_{i=1}\frac{(n_i - f_i)^2}{f_i}. \tag{A.53}$$

The f_i in the denominator can be approximated by n_i. Minimizing this LS function leads to the LS estimates for parameters $\vec{\theta}$, which usually needs to be implemented by numerical iteration procedure. The n_i is a Poisson variable with the expectation np_i, if n is sufficiently large, n_i can be approximated by a Gaussian, then $(n_i - f_i)/\sqrt{f_i}$ or $(n_i - f_i)/\sqrt{n_i}$ is approximately a standard normal variable, therefore, the $Q^2_{min}(\vec{\theta})$ distributed approximately as a $\chi^2(m-1)$ variable, where the degree of freedom of $m - 1$ is due to the existence of a constraint equation A.52.

General LS estimation with constraints

Often in the estimation problem there exist a set of constraint equations between the true values of observations $\eta_i, i = 1, ..., n$. The typical example is in the kinematic analysis of a particle reaction or decay, the momentum and energy conservation laws constitute a set of restrictions relating the various momenta and angles for the particle combination defining the kinematic hypothesis. Some of the quantities have been measured to a certain accuracy (say, the momenta and angles of curved tracks), and some are completely unknown (the variables for an unseen particle). The purpose of the LS estimation is to investigate the kinematic hypothesis: for a successful minimization the constraint equations will supply estimates for the unmeasured variables as well as "improved measurements" for the measured quantities.

Assume $\vec{y} = (y_1, ..., y_n)^T$ are the measured values with covariant matrix $V(\vec{y})$, let the true values of $\vec{y}$ be $\vec{\eta}$. In addition, we have a set of J unmeasurable variables $\vec{\xi} = (\xi_1, ..., \xi_J)^T$. The n measurable and J unmeasurable variables are related and have to satisfy a set of K constraint equations

$$f_k(\vec{\eta}, \vec{\xi}) = 0, \qquad k = 1, ..., K.$$

According to the LS principle, we should adopt as the best estimates of the unknown $\vec{\eta}$ and $\vec{\xi}$ those values for which

$$Q^2(\vec{\eta}) = (\vec{y} - \vec{\eta})^T V^{-1}(\vec{y})(\vec{y} - \vec{\eta}) = minimum, \tag{A.54}$$

$$f(\vec{\eta}, \vec{\xi}) = \vec{0}. \tag{A.55}$$

Usually the method of the Lagrangian multipliers are used to solve above equations. We introduce K additional unknowns $\vec{\lambda} = (\lambda_1, ..., \lambda_K)^T$ and rephrase the problem by requiring

$$Q^2(\vec{\eta}, \vec{\xi}, \vec{\lambda}) = (\vec{y} - \vec{\eta})^T V^{-1}(\vec{y})(\vec{y} - \vec{\eta}) + 2\vec{\lambda}^T \vec{f}(\vec{\eta}, \vec{\xi}) = minimum. \tag{A.56}$$

We have now a total of $n + J + K$ unknowns. When the derivatives of Q^2 with respect to all unknowns are put equal to zero we get following set of equations

$$V^{-1}(\vec{y})(\vec{\eta} - \vec{y}) + F_\eta^T \vec{\lambda} = \vec{0}, \tag{A.57}$$

$$F_\xi^T \vec{\lambda} = \vec{0}, \tag{A.58}$$

$$\vec{f}(\vec{\eta}, \vec{\xi}) = \vec{0}, \tag{A.59}$$

where the matrices F_η (of dimension $K \times N$) and F_ξ (of dimension $K \times J$) are defined by

$$(F_\eta)_{ki} \equiv \frac{\partial f_k}{\partial \eta_i}, \qquad (F_\xi)_{kj} \equiv \frac{\partial f_k}{\partial \xi_j}. \tag{A.60}$$

The solution of this set of equations for the $n + J + K$ unknowns and their errors must in general case be found by iterations, producing successively better approximations.

In the linear LS estimation problem for the n measurable and J unmeasurable variables which are related and have to satisfy a set of K constraint equations, if the measured values $\vec{y} = (y_1, ..., y_n)^T$ is a multi-normal variable, the Q^2_{min} is a χ^2 variable with the degree of

freedom $(K - J)$. For the non-linear LS estimation problem of non-linear constraint equations, and/or $\vec{y}$ is not a multi-normal variable, the Q^2_{min} may be approximated by $\chi^2(K - J)$.

The momentum-energy conservation laws constitute a set of 4 constraint equations. If all the particle's parameters in a reaction or a decay process have been measured (no unmeasurable variables) and the momentum-energy conservation laws are applied to obtain better values of particles parameters (4C kinematic fit), the Q^2_{min} of the LS estimator is then an approximate $\chi^2(4)$ variable. If there exist J unmeasurable variables and r intermediate resonances which promptly decayed to observed final state particles, and the invariant masses of daughter particles of these resonances are constrained to their mother particles' masses, then the Q^2_{min} is approximately a $\chi^2(4 + r - J)$ variable.

A.2 Interval estimation, confidence interval and upper limit

The task of the interval estimation is to locate a region which contains the true value of the parameter θ to be studied with a probability γ. This region is called the confidence interval with coverage probability γ. When the goal of an experiment is to determine a parameter θ, the result is usually expressed by quoting, in addition to the point estimate, some sort of confidence interval which reflects the statistical precision of the measurement. In the simplest case this can be given by the parameter's estimated value $\hat{\theta}$ plus/minus an estimate of the standard deviation of $\hat{\theta}$, $\sigma_{\hat{\theta}}$. If the parameter θ has boundary (without losing generality, we assume it is lower boundary with the value zero throughout this Appendix), and the estimate of θ in an experiment is close to this boundary, then the determination of the interval estimation is difficult and needs to be treated in special way.

A.2.1 Frequentist confidence interval

Neyman method for confidence interval

Confidence interval refers to frequentist interval obtained with a procedure due to Neyman [7]. Consider a *pdf* $f(x;\theta)$ where x represents the measurement of the experiment and θ the unknown parameter for which we want to construct a confidence interval. The variable x could (and often does) represent an estimator of θ. Using $f(x;\theta)$ we can find for a pre-specified probability $\gamma = 1 - \alpha$ and for every value of θ a set of values $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ such that

$$P(x_1 < x < x_2;\theta) = 1 - \alpha \equiv \gamma = \int_{x_1}^{x_2} f(x;\theta)dx. \tag{A.61}$$

This is illustrated in Fig. A.1: a horizontal line segment $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ is drawn for representative values of θ. The union of such intervals for all values of θ, designated in the figure as $D(\alpha)$, is known as the confidence belt. Typically the curves $x_1(\theta,\alpha)$ and $x_2(\theta,\alpha)$ are monotonic functions of θ, which we assume for this discussion.

Upon performing an experiment to measure x and obtaining a value x_0, one draws a vertical line through x_0. The confidence interval for θ is the set of all values of θ for which

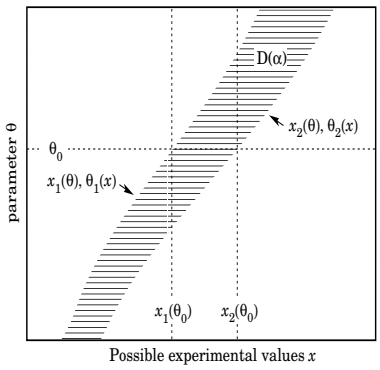

Figure A.1: Construction of the confidence belt

the corresponding line segment $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ is intercepted by this vertical line. Such confidence intervals are said to have a confidence level (CL) equal to $\gamma = 1-\alpha$.

Now suppose that the true value of θ is θ_0, indicated in the figure.We see from the figure that θ_0 lies between $\theta_1(x)$ and $\theta_2(x)$ if and only if x lies between $x_1(\theta_0)$ and $x_2(\theta_0)$. The two events thus have the same probability, and since this is true for any value θ_0, we can drop the subscript 0 and obtain

$$\gamma = 1-\alpha = P(x_1(\theta) < x < x_2(\theta)) = P(\theta_2(x) < \theta < \theta_1(x)). \tag{A.62}$$

In this probability statement $\theta_1(x)$ and $\theta_2(x)$, *i.e.*, the endpoints of the interval, are the random variables and θ is an unknown constant. If the experiment were to be repeated a large number of times, the interval $[\theta_1, \theta_2]$ would vary, covering the fixed value θ in a fraction $\gamma = 1-\alpha$ of the experiments.

The condition of coverage probability does not determine x_1 and x_2 uniquely and additional criteria are needed. The most common criterion is to choose central intervals such that the probabilities below x_1 and above x_2 are each $\alpha/2$. In other cases one may want to report only an upper or lower limit, then the probability excluded below x_1 or above x_2 can be set to zero.

When the observed random variable x is continuous, the coverage probability obtained with the Neyman construction is $\gamma = 1-\alpha$, regardless of the true value of the parameter. If x is discrete, however, it is not possible to find segments $[x_1(\theta,\alpha), x_2(\theta,\alpha)]$ that satisfy Eq. A.62 exactly for all values of θ. By convention one constructs the confidence belt requiring the probability $P(x_1 < x < x_2)$ to be greater than or equal to $\gamma = 1-\alpha$. This gives confidence intervals that include the true parameter with a probability greater than or equal to $\gamma = 1-\alpha$.

Gaussian distributed measurements

An important example of constructing a confidence interval is when the data consist of a single random variable x that follow a Gaussian distribution; this is often the case when x represents an estimator for a parameter and one has a sufficiently large data sample. If there is more than one parameter being estimated, the multivariate Gaussian is used.

For the univariate case with known σ,

$$\gamma = 1 - \alpha = \frac{1}{\sqrt{2\pi}\sigma} \int_{\mu-\delta}^{\mu+\delta} e^{-(x-\mu)^2/2\sigma^2} dx = erf(\frac{\delta}{\sqrt{2}\sigma}) \tag{A.63}$$

is the probability that the measured value x will fall within $\pm\delta$ of the true value μ. From the symmetry of the Gaussian with respect to x and μ, this is also the probability for the interval $x \pm \delta$ to include μ. The choice $\delta = \sigma$ gives an interval called the standard error which has $\gamma = 1 - \alpha = 68.27\%$ if σ is known. Values of α for other frequently used choices of δ are given in Table A.1. The relation of α and δ can be also represented by the cumulated distribution function for the χ^2 distribution for $\chi^2 = (\delta/\sigma)^2$ and $n = 1$ degree of freedom:

$$\gamma = 1 - \alpha = F(\chi^2; n = 1). \tag{A.64}$$

For multivariate measurements of, say, n parameter estimates $\hat{\vec{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_n)^T$, one

Table A.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
0.0455	2σ	0.1	1.64σ
0.0027	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	0.0001	3.89σ

requires the full covariance matrix $V_{ij} = cov(\hat{\theta}_i, \hat{\theta}_j)$, which can be estimated by ML or LS method.

If the parameters $\vec{\theta}$ are estimated with the ML method, for sufficient large n and the likelihood function satisfies the regularity conditions, the likelihood function distributed asymptotically as a multi-Gaussian, then we have

$$lnL(\vec{\theta}) = lnL_{max} - Q(\vec{\theta})/2, \tag{A.65}$$

where $Q(\vec{\theta}) = (\vec{\theta} - \hat{\vec{\theta}})^T V^{-1}(\vec{\theta})(\vec{\theta} - \hat{\vec{\theta}})$ is asymptotically a $\chi^2(k)$ variable, and k is the dimension of $\vec{\theta}$. The intersection contour of super-plane $lnL = lnL_{max} - Q_\gamma/2$ and super-surface $lnL(\vec{\theta})$ forms the boundary of the confidence region of $\vec{\theta}$ with coverage probability of $\gamma = 1 - \alpha$, which is calculated by the cumulated χ^2 function

$$\gamma = 1 - \alpha = P(Q \le Q_\gamma) = \int_0^{Q_\gamma} \chi^2(Q; \nu = k) dQ = F_\alpha(Q_\gamma; \nu = k). \tag{A.66}$$

In the case that the parameters $\vec{\theta}$ are estimated with LS method, for linear LS estimator and multi-Gaussian measurements, we have

$$Q^2(\vec{\theta}) = Q^2_{min} + Q^2_{LS}, \tag{A.67}$$

where $Q^2_{LS}(\vec{\theta}) = (\vec{\theta} - \hat{\vec{\theta}})^T V^{-1}(\vec{\theta})(\vec{\theta} - \hat{\vec{\theta}})$ is a $\chi^2(k)$ variable and k is the dimension of $\vec{\theta}$ for non-constraint LS estimation, and the dimension of $\vec{\theta}$ minus the number of independent linear constraint equations for constraint LS case. The intersection contour of super-plane $Q^2(\vec{\theta}) = Q^2_{min} + Q_\gamma$ and super-surface $Q^2(\vec{\theta})$ forms the boundary of the confidence region of $\vec{\theta}$ with coverage probability of $\gamma \equiv 1 - \alpha$, which is also calculated by Eq. A.66. Values of Q_γ for $k = 1, 2, 3$ are given in Table A.2 for several values of the coverage probability $\gamma = 1 - \alpha$.

Table A.2: Q_γ for $k = 1, 2, 3$ corresponding to a coverage probability $\gamma = 1 - \alpha$ in the large data sample limit.

$\gamma(\%)$	k=1	k=2	k=3
68.27	1.00	2.30	3.53
90.	2.71	4.61	6.25
95.	3.84	5.99	7.82
95.45	4.00	6.18	8.03
99.	6.63	9.21	11.34
99.73	9.00	11.83	14.16

If the mentioned conditions are not fully satisfied, the confidence region determined by Eqs. A.65 and A.67 are not exact but an approximate one.

The ML method has an advantage that is easier to calculate the confidence region for combining several independent measurements of same parameters. Assume N independent measurements give likelihood functions $lnL_i(\vec{\theta}), i = 1, ..., N$, then the combined likelihood function is simply

$$lnL(\vec{\theta}) = \sum_i^N lnL_i(\vec{\theta}). \tag{A.68}$$

Then using this likelihood functions in Eq. A.65 can give the confidence region with coverage probability γ for combined estimate $\hat{\vec{\theta}}$.

Poisson distributed measurements

If n represents the number of events produced in a reaction with cross section σ, say in a fixed integrated luminosity L, then it follows a Poisson distribution with mean $s = \sigma L$ in the case there is no background. Therefore, to determine the cross section of a reaction or the branching ratio of a decay process in terms of the number of observed events, the interval estimation of Poisson distributed data must be met. The probability of observing n events of the Poisson distribution with the mean s is

$$P(n, s) = \frac{s^n e^{-s}}{n!}. \tag{A.69}$$

The upper and lower (one sided) limits on the mean s can be found from the Neyman procedure to be

$$s_{lo} = \frac{1}{2} F^{-1}_{\chi^2}(\alpha_{lo}; 2n), \tag{A.70}$$

$$s_{up} = \frac{1}{2} F^{-1}_{\chi^2}(1-\alpha_{up}); 2(n+1)), \tag{A.71}$$

where the upper and lower limits are at confidence levels of $1-\alpha_{lo}$ and $1-\alpha_{up}$, respectively, and $F^{-1}_{\chi^2}$ is the quantile of the χ^2 distribution (inverse of the cumulative distribution). The quantiles $F^{-1}_{\chi^2}$ can be obtained from standard tables or from the CERNLIB routine CHISIN. For central confidence intervals at CL $1-\alpha$, set $\alpha_{lo} = \alpha_{up} = \alpha/2$. Values for confidence levels of 90% and 95% are shown in Table A.3.

Table A.3: Lower and upper (one-sided) limits for the mean s of a Poisson variable given n observed events in the absence of background, for CL of 90% and 95%

.

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

If the number of observed events n contains both signal and background events, which are Poisson variables with mean s and b, respectively, then we have

$$P(n,s) = \frac{(s+b)^n e^{-(s+b)}}{n!}. \tag{A.72}$$

For a specific value of s, the upper and lower limit of the central confidence region, $[n_l, n_u]$, and the lower limit of the upper confidence belt, n_{l_0}, at given confidence level $\gamma = 1-\alpha$ can be determined by

$$\sum_{n=0}^{n_l} P(n,s) \leq \frac{\alpha}{2}, \qquad \sum_{n=n_u+1}^{\infty} P(n,s) \leq \frac{\alpha}{2}, \tag{A.73}$$

$$\sum_{n=0}^{n_{lo}} P(n,s) \leq \alpha, \tag{A.74}$$

respectively. For all s values, such calculations give the confidence belts for central region and upper confidence belt. The inequality sign is to ensure the actual coverage greater or equal to the given coverage in the discrete variable case.

Confidence interval near the physics boundary

A number of issues arise in the construction and interpretation of confidence intervals when the parameter can only take on values in a restricted range. An important sample is where the mean of a Gaussian variable is constrained on physical grounds to be non-negative. This arises, for example, when the square of the neutrino mass is estimated from $\hat{m}^2 = \hat{E}^2 - \hat{p}^2$, where $\hat{E}$ and $\hat{p}$ are independent, Gaussian distributed estimates of the energy and momentum. Although the true m^2 is constrained to be positive, random errors in $\hat{E}$ and $\hat{p}$ can easily lead to negative values for the estimate $\hat{m}^2$.

If one uses the prescription given above for Gaussian distributed measurements, which says to construct the interval by taking the estimate plus/minus one standard deviation, then this can give intervals that are partially or entirely in the unphysical region. In fact, by following strictly the Neyman construction for the central confidence interval, one finds that the interval is truncated below zero; nevertheless an extremely small or even a zero-length interval can result.

An additional important example is where the experiment consists of counting a certain number of events, n, which is assumed to be Poisson distributed. Suppose the expectation value $E(n) = \mu$ is equal to $s + b$, where s and b are the means for signal and background processes, and assume further that b is a known constant. Then $\hat{s} = n - b$ is an unbiased estimator for s. Depending on true magnitudes of s and b, the estimate $\hat{s}$ can easily fall in the negative region. Similar to the Gaussian case with the positive mean, the central confidence interval or even the upper limit for s may be of zero length.

An additional difficulty arises when a parameter estimate is not significantly far away from the boundary, in which case it is natural to report a one-sided confidence interval (often an upper limit). It is straightforward to force the Neyman prescription to produce only an upper limit by setting $x_2 = \infty$ in Eq. A.61 . Then x_1 is uniquely determined and the upper limit can be obtained. If, however, the data come out such that the parameter estimate is not so close to the boundary, one might wish to report a central (*i.e.*, two-sided) confidence interval. As pointed out by Feldman and Cousins [8], however, if the decision to report an upper limit or two-sided interval is made by looking at the data ("flip-flopping"), then the resulting intervals will not in general cover the parameter with the probability $1 - \alpha$.

With the confidence intervals suggested by Feldman and Cousins [8], the prescription determines whether the interval is one- or two-sided in a way which preserves the coverage probability. Intervals with this property are said to be unified. Furthermore, this prescription is such that null intervals do not occur. For a given choice of $1 - \alpha$, if the parameter estimate is sufficiently close to the boundary, then the method gives an one-sided limit. In the case of a Poisson variable in the presence of background, for example, this would occur if the number of observed events is compatible with the expected background. For parameter estimates increasingly far away from the boundary, *i.e.*, for increasing signal significance, the interval makes a smooth transition from one- to two-sided, and far away from the boundary one obtains a central interval. The intervals according to this method for the mean of Poisson variable in the absence of background are given in Table A.4.

The intervals constructed according to the unified procedure in Ref. [8] for a Poisson variable n consisting of signal and background have the property that for $n = 0$ observed

Table A.4: Unified confidence interval $[s_1, s_2]$ for a mean s of a Poisson variable given n observed events in the absence of background, for CL of 90% and 95%

.

	$1-\alpha = 90\%$		$1-\alpha = 95\%$	
n	s_1	s_2	s_1	s_2
0	0.00	2.44	0.00	3.09
1	0.11	4.36	0.05	5.14
2	0.53	5.91	0.36	6.72
3	1.10	7.42	0.82	8.25
4	1.47	8.60	1.37	9.76
5	1.84	9.99	1.84	11.26
6	2.21	11.47	2.21	12.75
7	3.56	12.53	2.58	13.81
8	3.96	13.90	2.94	15.29
9	4.36	15.30	4.36	16.77
10	5.50	16.50	4.75	17.82

events, the upper limit decreases for increasing expected background. This is counter-intuitive, since it is known that if $n = 0$ for the experiment in question, then no background was observed, and therefore one may argue that the expected background should not be relevant. Roe and Woodroofe [9] proposed a solution to this problem by using such a fact that, given an observation n, the background b can not be larger than n in any case. Therefore, the usual Poisson *pdf* should be replaced by a conditional *pdf*, and then this conditional *pdf* is used to construct the confidence intervals following Feldman and Cousins' procedure.

Confidence interval incorporating systematic uncertainties

A modification of the Neyman method incorporating systematic uncertainty of the signal detection efficiency has been proposed by Highland and Cousins [10], in which a "semi-Bayesian" approach is adopted, where an average over the probability of the detection efficiency is performed. This method is of limited accuracy in the limit of high relative systematic uncertainties. On the other hand, an entirely frequentist approach has been proposed for the uncertainty in the background rate prediction [11]. This approach is based on a two-dimensional confidence belt construction and likelihood ratio hypothesis testing and treats the uncertainty in the background as a statistical uncertainty rather than as a systematic one. Recently, Conrad *etal* extend the method of confidence belt construction proposed in [12] to include systematic uncertainties in both the signal and background efficiencies as well as systematic uncertainty of background expectation prediction. It takes into account the systematic uncertainties by assuming a *pdf* which parameterizes our knowledge on the uncertainties and integrating over this *pdf*. This method, combining classical and Bayesian elements, is referred to as semi-Bayesian approach. A FORTRAN program, POLE, has been coded to calculate the confidence intervals for a maximum of observed events of 100 and a maximum signal expectation of 50 [13].

A.2.2 Bayesian confidence interval

In Bayesian approach one has to assume a prior *pdf* of an unknown parameter and then perform an experiment to update the prior distribution. The prior *pdf* reflects the experimenter's subjective degree of belief about unknown parameter before the measurement was carried out. The updated prior, called posterior *pdf*, is used to draw inference on unknown parameter. This updating is done with the use of Bayes theorem [14]. Assuming that n represents the number of observed events, s is the expectation of the number of signal events which is unknown and to be inferred, $p(s|n)$ is the conditional *pdf* of observing n events with given signal s, $\pi(s)$ is the prior *pdf*, the Bayes theorem gives the posterior *pdf*:

$$h(s|n) = \frac{p(s|n)\pi(s)}{\int_0^\infty p(s|n)\pi(s)ds}. \tag{A.75}$$

Here the lower limit of the integral is zero, which is the possible minimum of the signal expectation. Using this posterior *pdf*, one can calculate a Bayesian confidence interval for the signal expectation s at given confidence level $CL = 1 - \alpha$:

$$1 - \alpha = \int_{s_L}^{s_U} h(s|n)ds. \tag{A.76}$$

However, such intervals are not uniquely determined. Often, the highest posterior density (HPD) confidence interval I is chosen, which is determined in following way:

$$1 - \alpha = \int_I h(s|n)ds, \quad h(s_1|n) \geq h(s_2|n) \quad for\ any\ s_1 \in I\ and\ s_2 \notin I. \tag{A.77}$$

The upper limit of the signal expectation s at given confidence level $CL = 1 - \alpha$, s_{UP}, is naturally given by:

$$1 - \alpha = \int_0^{s_{UP}} h(s|n)ds. \tag{A.78}$$

The nice feature of the Bayesian approach is that the zero value of an upper limit s_{UP} always corresponds to the zero value of confidence level $CL = 1 - \alpha$, which is not necessarily true for the classical approach. The most important issue is to determine a prior *pdf* of the parameter. This is an issue which brings most of controversies into Bayesian methods. An important question is that if one should use an $informative$ prior, *i.e.*, a prior which incorporates results of previous experiments, or a $non - informative$ prior, *i.e.*, a prior which claims total ignorance. The major objection against informative prior is based on such argument: if we assume a prior which incorporates results of previous experiments, then our measurement will not be independent, hence, we will not be able to combine our results with previous results by taking a weighted average.

Thus, we only discuss the Bayesian inference that assumes a non-informative prior *pdf* for the non-negative parameter of a Poisson distribution. For the case that in the "signal region" where the signal events resides, the number of signal events is a Poisson variable with unknown expectation s , and the number of background events is a Poisson variable with expectation b, the conditional *pdf* of observing total events n, $p(s|n)$, can be represented by

$$P(s|n) = \frac{(s+b)^n e^{-(s+b)}}{n!}. \tag{A.79}$$

To deduce the posterior *pdf*, one has to assume a prior *pdf*. Bayes stated that, the non-informative prior for any parameter must be flat [14]. This statement does not based on any strict mathematical argument, but merely his intuition. The obvious weakness of Bayes prior *pdf* is that if one can assume a flat distribution of an unknown parameter, then one can also assume a flat distribution for any function of this parameter, and these two prior functions are apparently not identical. Jeffreys [15] resolves this problem by introducing an invariate non-informative prior *pdf*, which can be deduced from Fisher information. For the *pdf* shown by Eq. A.79, the Jeffreys prior *pdf* is proportional to $1/\sqrt{s+b}$. In general, we can use a prior *pdf* of

$$\pi(s) \propto \frac{1}{(s+b)^m}, \qquad s \geq 0, \quad 0 \leq m \leq 1. \tag{A.80}$$

$m = 0$ corresponds to Bayes prior, and $m = 0.5$ to Jeffreys prior. One can choose m value as he/she thinks appropriate, however, it should always be kept in mind that different m value will give different answer for the confidence interval or upper limit. Substituting $p(n|s)$ of Eq. A.79 and $\pi(s)$ of Eq. A.80 into Eq. A.75, the posterior *pdf* is then given by

$$h(s|n) = \frac{(s+b)^{n-m} e^{-(s+b)}}{\Gamma(n-m+1, b)}, \tag{A.81}$$

where

$$\Gamma(x, b) = \int_0^\infty s^{x-1} e^{-x} ds, \qquad x > 0, b > 0 \tag{A.82}$$

is an incomplete gamma function.

In the case that the systematic uncertainties of the signal efficiency and background expectation can be neglected, the signal expectation s is an unknown constant and the background expectation b is a known value. Substituting the posterior *pdf* of Eq. A.81 into Eq. A.78, we obtain

$$\alpha = \frac{\Gamma(n-m+1, s_{UP}+b)}{\Gamma(n-m+1, b)}. \tag{A.83}$$

If the flat prior$m = 0$is used, Eq. A.83 turns into

$$\alpha = e^{-s_{UP}} \cdot \frac{\sum_{k=0}^{n} \frac{(s_{UP}+b)^k}{k!}}{\sum_{k=0}^{n} \frac{b^k}{k!}}. \tag{A.84}$$

The upper limit s_{UP} at given confidence level $1-\alpha$ can be acquired by solving Eq. A.83 or Eq. A.84 numerically from measured values of n and b. Eq. A.84 has been recommended by PDG [4], therefore, widely used in particle physics experiments. However, from statistics point of view, the Jeffreys prior seems to be a more appropriate non-informative prior as mentioned above, therefore, using Eq. A.83 with $m = 0.5$ to determine s_{UP} seems a reasonable choice.

Now we turn to the question of inclusion of systematic uncertainties. First we consider only the uncertainty of background expectation is present, and the distribution of the background expectation is represented by a pdf $f_{b'}(b, \sigma_b)$ with the mean b and standard deviation σ_b . The conditional pdf expressed by Eq. A.79 now is modified to

$$q(n|s)_b = \int_0^\infty p(n|s)_{b'} \cdot f_{b'}(b, \sigma_b) db', \tag{A.85}$$

where $p(n|s)_{b'}$ has the same expression in Eq. A.79 with b replaced by b'.

Next we take into account the uncertainties of the signal efficiency and background expectation simultaneously, and consider they are independent each other. The distribution of the signal relative efficiency ε (with respect to the nominal signal detection efficiency) is described by a *pdf* $f_\varepsilon(1, \sigma_\varepsilon)$ with the mean 1 and standard deviation σ_ε. The conditional *pdf* described by Eq. A.79 is then further modified to

$$q(n|s)_b = \int_0^\infty \int_0^\infty p(n|s\varepsilon)_{b'} f_{b'}(b, \sigma_b) f_\varepsilon(1, \sigma_\varepsilon) db' d\varepsilon, \tag{A.86}$$

where $p(n|s\varepsilon)_{b'}$ represents that in Eq. A.79 b is replaced by b' , and s by $s\varepsilon$. One notices that the lower limits of integrals in Eqs. A.85, A.86 are all zeros, which are the possible minimum value of any efficiencies and number of background events.

Using $q(s|n)_b$ in Eqs. A.85, A.86 to construct posterior *pdf*

$$h(s|n) = \frac{q(n|s)_b \pi(s)}{\int_0^\infty q(n|s)_b \pi(s) ds}, \tag{A.87}$$

one can calculate the confidence interval or upper limit s_{UP} on s at any given confidence level with inclusion of systematic uncertainties in terms of Eq. A.76 or A.78.

An algorithm for calculating the upper limit at given confidence level with or without inclusion of systematic uncertainties in pure Bayesian approach has been coded. It has been implemented as a FORTRAN program, BPULE (Bayesian Poissonian Upper Limit Estimator) [16], where an iterative procedure is carried out by minimizing the difference between the given confidence level and the calculated value in terms of Eq. A.78 until a convergence is reached.

A.3 Tests of hypotheses

In addition to estimating parameters, one often wants to assess the validity of certain statements concerning the data's underlying distribution. Hypothesis tests provide a rule for accepting or rejecting hypotheses depending on the outcome of a measurement. We restrict ourselves here to discuss the Goodness-of-fit tests - one of the non-parametric tests, which deals with questions of the functional form for the distribution of the data and gives the probability to obtain a level of incompatibility with a certain hypothesis that is greater than or equal to the level observed with the actual data. Two methods will be stated: Pearson's χ^2 test and Kolmogorov-Smirnov test, which is applicable for the large and small size of the measured data sample, respectively. Finally, we have a section to discuss an important concept in particle physics experiment-the statistical significance of signal.

A.3.1 Goodness-of-fit test

Pearson's χ^2 test

We assume that n observations on the variable x belong to N mutually exclusive classes, such as successive intervals in a histogram, non-overlapping regions in two-dimentional

plot, etc. The number of events $n_1, n_2, ..., n_N$ in the different classes will then be multinomially distributed, with probabilities p_i for the individual classes as determined by the underlying distribution $f(x)$ for continuous variable x:

$$p_i = \int_{\triangle x_i} f(x)dx, \qquad i = 1, 2, ..., N,$$

or $q_j = P(x = x_j), j = 1, 2, \cdots$ for discrete variable x:

$$p_i = \sum_{j, x_j \in \triangle x_i} q_j, \qquad i = 1, 2, ..., N,$$

where $\triangle x_i$ represents the i-th interval. The hypothesis we wish to test specifies the class probabilities according to a certain prescription,

$$H_0 : p_i = p_{0i}, \qquad i = 1, 2, ..., N, \tag{A.88}$$

where

$$\sum_{i=1}^{N} p_{0i} = 1, \tag{A.89}$$

is the overall normalization and

$$p_{0i} = \int_{\triangle x_i} f_0(x)dx, \quad or \quad p_{0i} = \sum_{j, x_j \in \triangle x_i} q_{0j}.$$

Therefore, what we wish to test is if the distribution of the observation $f(x)$ or q_j is consistent with the assigned distribution $f_0(x)$ or q_{0j}, or equivalently, if the hypothesis H_0 is accepted by the observed data, given that the total number in all classes is n? To test whether the set of predicted numbers np_{0i} is compatible with the set of observed numbers n_i we take as our test statistic the quantity

$$X^2 = \sum_{i=1}^{N} \frac{(n_i - np_{0i})^2}{np_{0i}} = \frac{1}{n} \sum_{i=1}^{N} \frac{n_i^2}{p_{0i}} - n. \tag{A.90}$$

When H_0 is true this statistic is approximately $\chi^2(N-1)$ distributed. This is called the Pearson theorem.

If H_0 is true and the experiment is repeated many times under the same conditions with n observations, the actual values obtained for X^2, X^2_{obs}, will therefore be distributed nearly like $\chi^2(N-1)$; in particular, the mean value for X^2_{obs} will be $\simeq N-1$ and the variance $\simeq 2(N-1)$. If, on the other hand, H_0 is not true, the expectation for each n_i is not np_{0i}, and the sum of terms $(n_i - np_{0i})^2/np_{0i}$ will tend to become on the average larger than if H_0 were true. Hence it seems reasonable to reject H_0 if X^2_{obs} becomes too large. The criteria to reject H_0 at *significance level* α is

$$X^2_{obs} > \chi^2_\alpha(N-1), \tag{A.91}$$

where $\chi^2_\alpha(N-1)$ is determined by the $\chi^2(N-1)$ *pdf* $f(y; N-1)$ such that

$$\alpha = \int_{\chi^2_\alpha(N-1)}^{\infty} f(y; N-1)dy.$$

Often, the model which to describe the distribution of the measured data includes L unknown parameters. For a Least-Square estimation we know that the comparison between data and fitted model is made using the χ^2 distribution with a number of degrees of freedom equal to the number of independent observations minus the number of independent parameters estimated. This procedure is exact only in the limit of infinitely many observations and with a linear parameter dependence; otherwise it is an approximation. Thus, if there are L parameters in H_0 which are estimated by the LS method and N classes subject to an overall normalization condition, Pearson's χ^2 test for goodness-of-fit consists in comparing the fitted (minimum) value X^2_{min} to the χ^2 distribution with $(N-1-L)$ degrees of freedom.

Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) test avoids the binning of individual observations and may be more sensitive to the data, and is superior to the χ^2 test in particular for small samples and has many nice properties when applied to problems in which no parameters are estimated.

Given n independent observations on the variable x we form an *ordered sample* by arranging the observations in ascending order of magnitude, $x_1, x_2, \cdots, x_n$. The cumulative distribution for this sample of size n is now defined by

$$S_n(x) = \begin{cases} 0, & x < x_i, \\ \frac{i}{n}, & x_i \leq x \leq x_{i+1}, \\ 1, & x \geq x_n. \end{cases} \tag{A.92}$$

Thus $S_n(x)$ is an increasing step function with a step of height $1/n$ at each of the observational points $x_1, x_2, \cdots, x_n$.

The KS test involves a comparison between the observed cumulative distribution function $S_n(x)$ for the data sample and the cumulative distribution function $F_0(x)$ which is determined by some theoretical model. We state the null hypothesis as

$$H_0: \quad S_n(x) = F_0(x). \tag{A.93}$$

For H_0 true one expects that the difference between $S_n(x)$ and $F_0(x)$ at any point should be reasonably small. The KS test looks at the difference $S_n(x) - F_0(x)$ at all observed points and takes the maximum of the absolute value of this quantity, D_n, as a test statistic

$$D_n = max|S_n(x) - F_0(x)|. \tag{A.94}$$

It can be shown that provided no parameter in $F_0(x)$ has been determined from the data, and assuming H_0 true, the variable D_n has a distribution which is independent of $F_0(x)$, *i.e.* D_n is *distribution free*. This holds irrespective of the sample size.

For continuous variable x and finite n, the D_n has the distribution of [17]

$$P(D_n < z + \frac{1}{2n}) = \begin{cases} 0, & z \leq 0, \\ \int_{\frac{1}{2n}-z}^{\frac{1}{2n}+z} \int_{\frac{3}{2n}-z}^{\frac{3}{2n}+z} \cdots \int_{\frac{2n-1}{2n}-z}^{\frac{2n-1}{2n}+z} f(y_1, \cdots, y_n) dy_1 \cdots dy_n, & 0 < z < 1 - \frac{1}{2n} \\ 1, & z \geq 1 - \frac{1}{2n}, \end{cases} \tag{A.95}$$

where

$$f(y_1, \cdots, y_n) = \begin{cases} n!, & when\ 0 < y_1 < \cdots < y_n < 1, \\ 0, & others. \end{cases} \tag{A.96}$$

For large n the D_n has the cumulative distribution of

$$\lim_{n\to\infty} P(D_n \leq \frac{z}{\sqrt{n}}) = 1 - 2\sum_{r=1}^{\infty}(-1)^{r-1}e^{-2r^2z^2}, \qquad (z > 0). \tag{A.97}$$

This relation is approximately valid at $n \simeq 80$.

If H_0 is true, the D_n tends to be small, while if H_0 is not true, the D_n tends to be larger than if H_0 were true. Hence it seems reasonable to reject H_0 if D_n becomes too large. The criteria to reject H_0 at significance level α is

$$P(D_n > D_{n,\alpha}). \tag{A.98}$$

A table in the Appendix of the book [2] or [3] gives the critical values $D_{n,\alpha}$ at 5 different significance level α for $n \leq 100$, and the approximate expression for $n > 100$.

A.3.2 Statistical significance of signal

The statistical significance of a signal in an experiment of particle physics is to quantify the degree of confidence that the observation in the experiment either confirm or disprove a null hypothesis H_0, in favor of an alternative hypothesis H_1. Usually the H_0 stands for known or background processes, while the alternative hypothesis H_1 stands for a new or a signal process plus background processes with respective production cross section. This concept is very useful for usual measurements that one can have an intuitive estimation, to what extent one can believe the observed phenomena are due to backgrounds or a signal. It becomes crucial for measurements which claim a new discovery or a new signal. As a convention in particle physics experiment, the "5σ" standard, namely the statistical significance $S \geq 5$ is required to define the sensitivity for discovery; while in the cases $S \geq 3$ ($S \geq 2$), one may claim that the observed signal has strong (weak) evidence.

However, as pointed out in Ref. [18], the concept of the statistical significance has not been employed consistently in the most important discoveries made over the last quarter century. Also, the definitions of the statistical significance in different measurements differ from each other. Listed below are various definitions for the statistical significance in counting experiment (see, for example, refs. [19] [20] [21]):

$$S_1 = (n - b)/\sqrt{b}, \tag{A.99}$$

$$S_2 = (n - b)/\sqrt{n}, \tag{A.100}$$

$$S_{12} = \sqrt{n}/\sqrt{b}, \tag{A.101}$$

$$S_{B1} = S_1 - k(\alpha)\sqrt{n/b}, \tag{A.102}$$

$$S_{B12} = 2S_{12} - k(\alpha), \tag{A.103}$$

$$\int_{-\infty}^{S_N} N(0,1)dx = \sum_{i=0}^{n-1} e^{-b}\frac{b^i}{i!}, \tag{A.104}$$

where n is the total number of the observed events, which is the Poisson variable with the expectation $s+b$, s is the expected number of signal events to be searched, while b is the known expected number of Poisson distributed background events. All numbers are counted in the "signal region" where the searched signal events are supposed to appear. In equations A.102 and A.103 the $k(\alpha)$ is a factor related to the α that the corresponding statistical significance assumes $1-\alpha$ acceptance for positive decision about signal observation, and $k(0.5)=0, k(0.25)=0.66, k(0.1)=1.28, k(0.05)=1.64$, etc [20]. In equation A.104, $N(0,1)$ is a notation for the standard normal function. On the other hand, the measurements in particle physics often examine statistical variables that are continuous in nature. Actually, to identify a sample of events enriched in the signal process, it is often important to take into account the entire distribution of a given variable for a set of events , rather than just to count the events within a given signal region of values. In this situation, I. Narsky [21] gives a definition of the statistical significance via likelihood function

$$S_L = \sqrt{-2\ln L(b)/L(s+b)} \tag{A.105}$$

under the assumption that $-2\ln L(b)/L(s+b)$ distributes as χ^2 function with degree of freedom of 1.

Upon above situation, it is clear that we desire to have a self-consistent definition for statistical significance, which can avoid the ambiguity that the same S value in different measurements may imply virtually different statistical significance, and can be suitable for both counting experiment and continuous test statistics.

Definition of the statistical significance

In the PDG [4], the $p-$value is defined to quantify the level of agreement between the experimental data and a hypothesis. Assume an experiment makes a measurement for test statistic t being equal to t_{obs}, and t has a probability density function $g(t|H_0)$ if a null hypothesis H_0 is true. We further assume that large t values correspond to poor agreement between the null hypothesis H_0 and data, then the $p-$value of an experiment would be

$$p(t_{obs}) = P(t > t_{obs}|H_0) = \int_{t_{obs}}^{\infty} g(t|H_0)dt. \tag{A.106}$$

A very small $p-$value tends to reject the null hypothesis H_0.

Since the $p-$value of an experiment provides a measure of the consistency between the H_0 hypothesis and the measurement, Zhu [22] define the statistical significance S in terms of the $p-$value in the following form

$$\int_{-S}^{S} N(0,1)dx = 1 - p(t_{obs}) \tag{A.107}$$

under the assumption that the null hypothesis H_0 represents that the observed events can be described merely by background processes. Because a small $p-$value means a small probability of H_0 being true, corresponds to a large probability of H_1 being true, one would get a large signal significance S by this expression. The left side of equation A.107 represents the integral probability of the normal distribution in the region within S standard deviation ($S\sigma$). In such a definition, some correlated S and $p-$values are listed in Table A.5.

Table A.5: *Statistical Significance S and correlated $p-$value.*

S	$p-$value
1	0.3173
2	0.0455
3	0.0027
4	6.3×10^{-5}
5	5.7×10^{-7}
6	2.0×10^{-9}

Statistical significance in counting experiment

A group of particle physics experiment involves the search for new phenomena or signal by observing a unique class of events that can not be described by background processes. One can address this problem to that of a "counting experiment", where one identifies a class of events using well-defined criteria, counts up the number of observed events, and estimates the average rate of events contributed by various backgrounds in the signal region, where the signal events (if exist) will be clustered. Assume in an experiment, the number of signal events in the signal region is a Poisson variable with the expectation s, while the number of events from backgrounds is a Poisson variable with a known expectation b, then the observed number of events distributes as the Poisson variable with the expectation $s+b$. If the experiment observed n_{obs} events in the signal region, then the $p-$value is

$$\begin{aligned} p(n_{obs}) &= P(n > n_{obs}|H_0) = \sum_{n=n_{obs}}^{\infty} \frac{b^n}{n!} e^{-b} \\ &= 1 - \sum_{n=0}^{n_{obs}-1} \frac{b^n}{n!} e^{-b}. \end{aligned} \tag{A.108}$$

Substituting this relation to equation A.107, one immediately has

$$\int_{-S}^{S} N(0,1)dx = \sum_{n=0}^{n_{obs}-1} \frac{b^n}{n!} e^{-b} \tag{A.109}$$

Then, the signal significance S can be easily determined. Comparing this equation with equation A.104 given by Ref. [21], we found the lower limit of the integral is different.

Statistical significance in continuous test statistics

The general problem in this situation can be addressed as follows. Suppose we identify a class of events using well-defined criteria, which are characterized by a set of n observations $x_1, \cdots, x_n$ for a random variable x. In addition, one wish to test a hypothesis which predicts the probability density function of x, say $f(x|\vec{\theta})$, where $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ is a set of parameters which need to be estimated from the data. Then the problem is to define a statistic that gives a measure of the consistency of the distribution of data with the distribution given by the hypothesis.

To be concrete, we consider the random variable x is, say, an invariant mass, and the n observations $x_1, \cdots, x_n$ give an experimental distribution of x. Assuming parameters $\vec{\theta} = (\theta_1, \cdots, \theta_k) \equiv (\vec{\theta_s}; \vec{\theta_b})$, where $\vec{\theta_s}$ and $\vec{\theta_b}$ represent the parameters belong to signal (say, a resonance) and backgrounds contribution, respectively. We assume the null hypothesis H_0 stands for that the experimental distribution of x can be described merely by the background processes, namely, the null hypothesis H_0 specifies fixed values for a subset of parameters $\vec{\theta_s}$. Therefore, the parameters $\vec{\theta}$ are restricted to lie in a subspace ω of its total space Ω. On the basis of a data sample of size n from $f(x|\vec{\theta})$ we want to test the hypothesis $H_0 : \vec{\theta}$ belongs to ω. Given the observations $x_1, \cdots, x_n$, the likelihood function is $L = \prod_{i=1}^{n} f(x_i|\vec{\theta})$. The maximum of this function over the total space Ω is denoted by $L(\hat{\Omega})$; while within the subspace ω the maximum of the likelihood function is denoted by $L(\hat{\omega})$, then we define the likelihood-ratio $\lambda \equiv L(\hat{\omega})/L(\hat{\Omega})$. It can be shown that for H_0 true, the statistic

$$t \equiv -2 \ln \lambda \equiv 2(\ln L_{max}(s+b) - \ln L_{max}(b)) \tag{A.110}$$

is distributed as $\chi^2(r)$ (r is the number of parameters specified by H_0) when the sample size n is large [1]. In equation A.110 we use $\ln L_{max}(s+b)$ and $\ln L_{max}(b)$ denoting $L(\hat{\Omega})$ and $L(\hat{\omega})$, respectively. If λ turns out to be in the neighborhood of 1, the null hypothesis H_0 is such that it renders $L(\hat{\omega})$ close to the maximum $L(\hat{\Omega})$, and hence H_0 will have a large probability of being true. On the other hand, a small value of λ will indicates that H_0 is unlikely. Therefore, the critical region of λ is in the neighborhood of 0, corresponding to large value of statistic t. If the measured value of t in an experiment is t_{obs}, from equation A.106 we have $p-$value

$$p(t_{obs}) = \int_{t_{obs}}^{\infty} \chi^2(t; r) dt. \tag{A.111}$$

Therefore, in terms of equation A.107, one can calculate the signal significance according to following expression:

$$\int_{-S}^{S} N(0,1) dx = 1 - p(t_{obs}) = \int_{0}^{t_{obs}} \chi^2(t; r) dt. \tag{A.112}$$

For the case of $r = 1$, we have

$$\begin{aligned} \int_{-S}^{S} N(0,1) dx &= \int_{0}^{t_{obs}} \chi^2(t; 1) dt \\ &= 2 \int_{0}^{\sqrt{t_{obs}}} N(0,1) dx. \end{aligned} \tag{A.113}$$

and immediately obtain

$$\begin{aligned} S &= \sqrt{t_{obs}} \\ &= [2(\ln L_{max}(s+b) - \ln L_{max}(b))]^{1/2}, \end{aligned} \tag{A.114}$$

which is identical to the equation A.105 given by Ref. [21].

Bibliography

[1] W.T. Eadie *et al*, Statistical methods in experimental physics, North-Holland publishing Co., Amsterdam, 1971.

[2] A. Frodesen *et al*, Probability and Statistics in particle physics, Universitetsforlaget, Bergen-Oslo-Tromsø, 1979.

[3] Yongsheng Zhu, Probability and statistics in experimental physics, 2nd edition, Academic Press, Beijing, 2006.

[4] Particle Data Group, W.-M. Yao *et al*, Journal of Physics G, **33**, 1 (2006).

[5] G. Cowan, Statistical data analysis, Oxford Univ. Press Inc., New York, 1998.

[6] R. Barlow, arXiv Physics/0406120, 2004.

[7] J. Neyman, Phil. Trans. Royal Soc. London, Series A, **236**, 333 (1937)

[8] G. Feldman and R.D. Cousins, Phys. Rev. **D57**, 3873 (1998).

[9] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D63**, 013009 (2001).

[10] R.D. Cousins and V.L. Highland, Nucl. Instr. Meth. in Phys. Res. **A320**, 331 (1992).

[11] W.A. Rolke and A.M. Lopez, Nucl. Instr. Meth. in Phys. Res. **A458**, 745 (2001).

[12] J. Conrad *et al*, Phys. Rev. **D67**, 012002 (2003).

[13] http://www3.tsl.uu.se/~conrad/pole.html.

[14] T. Bayes, Phil. Trans. Roy. Soc., **53**,370(1763).

[15] H. Jeffreys, Theory of probability, 3rd edition, Clarendon, Oxford, 1961.

[16] http://psip.ihep.ac.cn/~dusx/zhuys/Bayes/

[17] V. Rohatgi, An introduction to probability theory and mathematical statistics, John Wiley & Sons, New York, 1976.

[18] P.K. Sinervo, Proc of Conf. "Advanced statistical techniques in particle physics", Durham, UK, 18-22 March, 2002, p64.

[19] S.I. Bityukov *et al*, Nucl.Intr.Meth. **A452**, 518 (2000)

[20] S.I. Bityukov *et al*, Proc of Conf. "Advanced statistical techniques in particle physics", Durham, UK, 18-22 March, 2002, p77.

[21] I. Narsky, Nucl.Intr.Meth. **A450**, 444 (2000)

[22] Yongsheng Zhu, HEP&NP, **30**, 331 (2006)