Statistical Analysis For Physicists Basics

Priyashree Roy
FSU Weekly Group Meeting

Weekly Group Meeting

10/07/2014

Outline

Types of Errors

Two types -

- Random errors - They occur simply from the inability of any measuring device to give infinitely accurate answers. This leads to random fluctuations in the measurements. They affect the **PRECISION** of the measurement.
- Systematic errors - They are more in the nature of mistakes. They can come from faulty experiment, calibration or techniue. They affect the **ACCURACY** of the experiment.
- Categorize as random or systematic - detector resolution error, detector calibration error, time interval of taking a measurement (w/ clocks properly calibrated), detector inefficiency, statistical error in counting.

Types of Errors

Continued -

- Random - detector resolution error, time interval of taking a measurement (w/ clocks properly calibrated), statistical error in counting.
  Systematic - detector calibration error, detector inefficiency.
- Punch line -
  Factors that affect precision - random errors.
  Factors that affect accuracy - systematic errors.
- Random error (i.e. the spread in the mean, not the variance of the distribution (following slides)) dec. as $\frac{1}{\sqrt{N}}$, whereas systematic errors don't get affected by sample size (N). They can only be eliminated.
- So, usually the goal is to dec. random errors (by inc. N) till it is of the same order of magnitude as systematic errors.

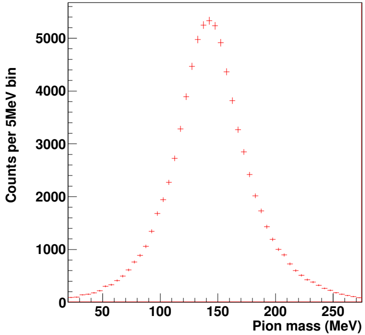
Outline

Random Errors

- We will always obtain a distribution of random observations for our experiments. The distributions are usually characterized by their mean (μ) and standard deviation (σ). This is the distribution of the hypothetical infinite set of data points, called as parent distribution.
- Since in reality we have only a finite set of data points, we can only get estimated mean ($\bar{x}$) and estimated standard deviation (s).
- $\bar{x} = \sum x_i P(x_i),\ s^2 = \sum (x_i - \bar{x})^2 P(x_i)$, P is probability function.
- 3 most common distribution functions - Binomial, Poisson and Gaussian.
- Poisson and Gaussian are limiting cases binomial distribution. For large sample size, Poisson tends to Gaussian.
- **Note :** We could have defined the average deviation $|x_i - \bar{x}|$ as the error. But, the absolute sign makes calculations difficult. So, we use standard deviation (σ) instead !

Poisson Distribution

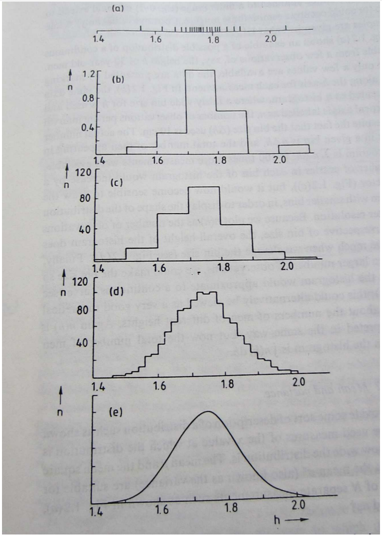
- This distribution describes the statistical fluctuations in the collection of a finite no. of counts over a finite interval of time. The observed counts will be distributed about the mean in a Poisson distribution instead of a Gaussian distribution.
- Estimated mean = N, mean counting rate or mean count.
  $s = \sqrt{N}$. This is what we see in our root histograms.
- As we increase the sample size, the fractional error goes down.
  $\sigma_f = \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}}$



Gaussian Distribution

- This distribution describes the distribution of random observations for many experiments.
- Probability density function, f(x)= $\frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{(x-\mu)^2}{2\sigma^2}]$
- ◇ Using the formula for s^2 (slide 2) will give s= σ.
- ◇ At $x = x \pm \sigma$, $f = \frac{f_{max}}{\sqrt{e}}$

Effect of Sample size on Mean and Variance



- My misconception - As we increase statistics (N), the Gaussian distribution (which represents the statistical fluctuations) will become narrower. This is wrong.
- With inc. N, s^2 will not change much. It will get closer to σ^2 but they don't differ by much anyway if N is not too small. From the formula for s^2, it is evident that with inc. N, both numerator and denominator will increase.
- On the other hand, **the spread in the determination of the mean goes down as $\frac{s}{\sqrt{N}}$.**

Example and Proof

Suppose we are measuring the pion mass and the detector resolution is 5 MeV. Let's assume that each datapoint has this error only. Then, $\sigma_i = \sigma = 5MeV$ for all i datapoints.

Then, $\bar{x} = \sum x_i P(x_i) = \frac{\sum x_i}{N}$ (since all datapoints have the same error, $P = 1/N$) and $s^2 = \sum (x_i - \bar{x})^2 P(x_i) = \frac{N\sigma}{N} = \sigma$

Using the error propagation equation,

${\sigma_\mu}^2 = \sum \sigma_i^2 w_i(x_i)$, where w_i is the properly normalized weight of each data in the calculation of the mean.

$= \sum \sigma_i^2 (\frac{\partial \mu}{\partial x_i})^2$

$= \sum \sigma_i^2 (\frac{1}{N})^2$

$= \frac{s^2}{N} = \frac{(5MeV)^2}{N}$, i.e. the spread in the mean dec. as sample size inc.

Outline

Introduction

- We will discuss the error estimation in 2 types of fitting techniques-
  - ◇ The maximum likelihood method
  - ◇ The method of least squares
- Method of least squares is a special case of the maximum likelihood method, basically when there is a lot of statistics.

The Maximum Likelihood Method

- In this method we write the likelihood function, $L = \prod P(y_i(\alpha))$ and maximize it. Corresponding value of the fit parameter α is the true value.
- In most cases this function is Gaussian distributed with respect to its fit parameter. E.g., the likelihood function for ϕ distribution and for a single meson production with pol. beam and unpol. target and Σ observable as the fit parameter. Then let's assume that $L = f(\Sigma)$ is Gaussian distributed.
- The error in Σ is then the standard deviation of the Gaussian distribution. So, at $\Sigma \pm \sigma$, $L = \frac{L_{max}}{e}$. Or, $-logL = -logL_{max} + 0.5$
  Or, $l = l_{min} + 0.5$
- We use Minuit to minimize l. The error in the fit parameter Σ is the change in the value of Σ that will step-up l by 0.5 from l_{min}. One way is to do it numerically.

The Maximum Likelihood Method Continued

Another way that Minuit can use to calculate error in the fit parameter - Suppose $l = l_{min}$ at $\Sigma = \Sigma_0$.

Taylor expansion about Σ_0, $l(\Sigma) = l(\Sigma_0 + \sigma)$

or, $l(\Sigma_0) + 0.5 = l(\Sigma)_0 + 0 + \frac{1}{2}\frac{\partial^2 l}{\partial \Sigma^2}|_{\Sigma_0}\sigma_{\Sigma}^2$

or, $\sigma_{\Sigma} = [\frac{\partial^2 l}{\partial \Sigma^2}]^{-\frac{1}{2}}$ evaluated at Σ_0.

The Method of Least Squares

- In this method we write the function, $S = \sum(\frac{y_i^{obs} - y_i^{fit}(\alpha)}{\sigma_i})^2$ and minimize it using Minuit. Corresponding value of the fit parameter α is the true value.
- In this method, the error σ in the fit parameter is the change in the parameter that will make S go from S_{min} to $S_{min} + 1$. One way is to find it numerically.
- The other way is to calculate $[\frac{1}{2}\frac{\partial^2 S}{\partial \alpha^2}]^{-\frac{1}{2}}$ at the minimum. We can derive this by using the Taylor expansion method as used in the previous slide.

Minuit and Factor of 2 needed for MLM Parameter Error Estimation

As is evident, the least squares method (LSM) and the likelihood method(MLM) fit error estimation differ by a factor of 2. The step size is 0.5 in MLM whereas it is 1.0 in LSM. That's why **in Minuit we need to multiply the $-logL$ expression by 2 to get the errors right if the step size is set to 1. Or, do not multiply $-logL$ by 2 but make the step size** $= 0.5$.

When can Minuit Error Calculation go Wrong ?

Following can be the causes -

- The value of step-up - it can be different from 0.5 if the likelihood is not gaussian distributed with respect to its parameters. This can happen for low statistics.
- Improper normalization of the χ^2 or the likelihood function -
  $\chi^2 = \sum \frac{(x_i - y_i(\alpha))^2}{e_i^2}$
  The terms $\frac{1}{e_i^2}$ should be inverse of variances. If they are only relative weights then the absolute values of the errors will not be correct.
  The likelihood function should be properly normalized. i.e.,
  $\int l(x,\alpha)dx = constant$, x being the observed datapoints. The normalization constant does not affect the value or the error of the fit parameters but it affects the convergence of the fit.
- Non-linear dependence on the fit parameter - This brings in a technical issue. Different techniques like MIGRAD, MINOS, HESSE will give different errors.

Outline

Basics

- This method is based on the assumption that each datapoint in the histogram is gaussian distributed with mean $y^{theory}(x_i)$ and std. deviation σ_i (the vertical error bar in the datapoints).
- It is thus necessary to have enough events per bin in counting experiments. **Rule of thumb is that N>10 per bin** since Poisson distribution tends to Gaussian in this case.
- With Gaussian distribution assumption for each datapoint, the probability to make an observed set of measurements is the product of the probabilities for each observation:
$P(\alpha) = \prod(\frac{1}{\sigma_i\sqrt{2\pi}}) exp\left\{-\frac{1}{2}\sum[\frac{y_i - y_i^{theory}(\alpha)}{\sigma_i}]^2\right\}$ where α is the fit parameter.
- The probability becomes maximum when the sum in the exponential becomes minimum. This sum is the goodness-of-fit parameter χ^2. Its value is affected by uncertainties in σ_i, functional form of the fit function $y_i{}^{theory}$ etc.

χ^2 Test for Goodness-of-fit

Recall that $s^2 = \sum (x_i - \bar{x})^2 w(x_i)$ is the estimated variance.

The parent or true invariance, $\sigma^2 = \sum \sigma_i^2 w(x_i)$

The normalized weight, $w(x_i) = \frac{1/\sigma_i^2}{\sum(1/\sigma_i^2)}$. Therefore, $\sigma^2 = \frac{N}{\sum(1/\sigma_i^2)}$

The estimated variance, which is charateristic of both the spread of the data and the accuracy of the fit is given by,

$s^2 = \frac{N}{N-m} \sum (x_i - \bar{x})^2 w(x_i) = (\frac{1}{N-m})(\frac{N}{\sum(1/\sigma_i^2)})(\sum \frac{(x_i-\bar{x})^2}{\sigma_i^2})$
or, $s^2 = (\frac{1}{N-m})\sigma^2\chi^2$

Here $N - m$ is the number of degrees of freedom for fitting N datapoints with an m parameter fit.

Reduced $\chi^2 = \frac{\chi^2}{N-m} = \frac{s^2}{\sigma^2} = \frac{Variance_{estimated}}{Variance_{true}}$

χ^2 Test for Goodness-of-fit and Effect of Bin Size

- For a good fit, $s^2 \sim \sigma^2$, so $\chi^2{}_{red} \sim 1$.
- For a bad fit, $s^2 > \sigma^2$, so $\chi^2{}_{red} > 1$. e.g. when the bin size is too big, the data distribution will differ from the fit curve.
- When there is error in the assignment of the σ_i of the datapoints (such as when they are not the Gaussian variances), then $\chi^2{}_{red} < 1$, for e.g. when bin size is too small and so counts per bin < 10.

References

- Data reduction and error analysis for the physical sciences, P.R. Bevington, D.K. Robinson.
- Statistics for nuclear and particle physicists, Louis Lyons
- Minuit manual