Mathematical Methods of Particle Astrophysics

Both in gamma ray and neutrino astronomy, many experiments are “counting experiment”. I’ll center my discussion on this topic.

Reference:
Statistics for Nuclear and Particle Physics. Louis Lyons.

Probability and probability density functions

Throwing a dice results into a finite number of possible outcomes (six). You can calculate probabilities for each outcome.

For situations in which the outcome is a real number (non-countable and dense), the probability of specific value being measured can't be calculated. Instead you can calculate the probability of measuring a range of values:

$$P = \int_{x_i}^{x_f} dx\, f(x) \qquad\qquad 1 = \int_{-\infty}^{\infty} dx\, f(x)$$

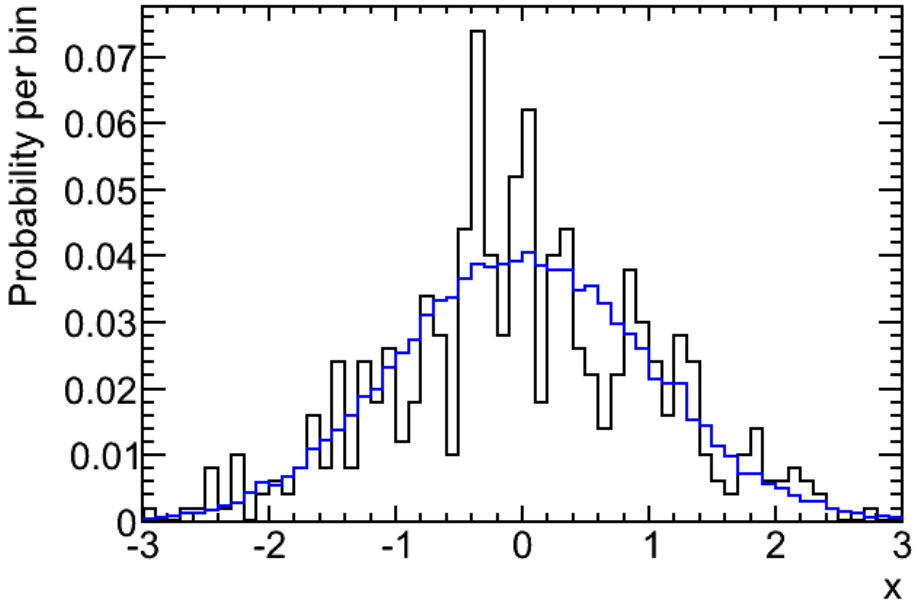
Example: the height of an infinite group of people. In fact, MDs report the "percentile" for height – which is the probability that you are at or above your measured height. Clearly this is an approximation, there is a finite number of people.

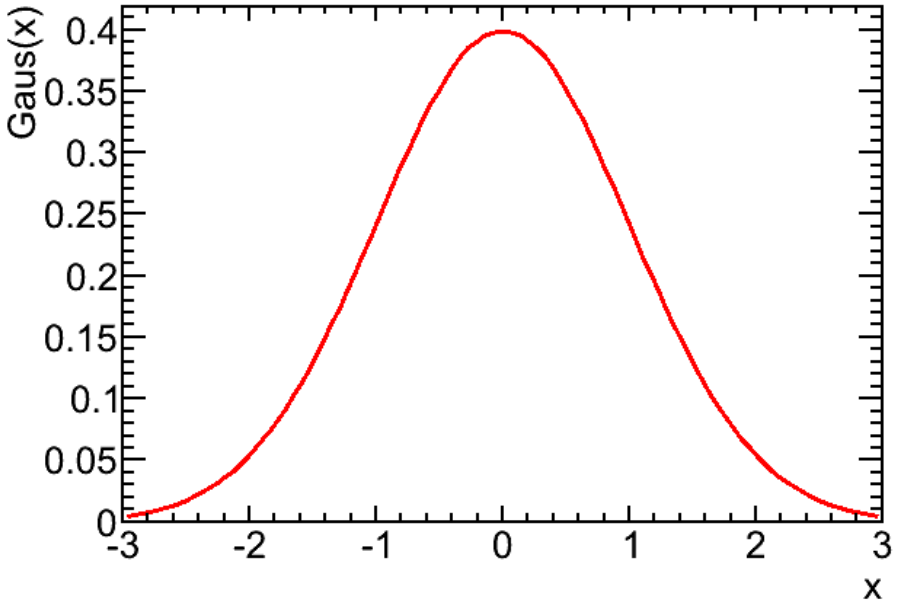
Probability density functions / distributions

Example: A Gaussian distribution is

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

The histogram below was generated for μ=0 and σ=1, with 500 and 25000 random samples. Note that in a histogram each bin has a well defined probability, i.e. each bin is an integral over the pdf.





Moments

For a given pdf, you can define the nth-moment:

$$\mu'_n = \int_{-\infty}^{\infty} x^n f(x) dx$$

Recall that

$$1 = \int_{-\infty}^{\infty} f(x) dx$$

The first 4 moments have names:

these are right in spirit but wrong in detail. See "Moments, cumulants" in ChaosBook.org

$$\mu'_1 = \int_{-\infty}^{\infty} x f(x) dx$$

mean

$$\mu'_3 = \int_{-\infty}^{\infty} x^3 f(x) dx$$

skewness

$$\mu'_2 = \int_{-\infty}^{\infty} x^2 f(x) dx$$

variance

$$\mu'_4 = \int_{-\infty}^{\infty} x^4 f(x) dx$$

kurtosis

Mean and variance estimates

The situation arises that you don't know a distribution or it's moments. (If you know all moments, you know the distribution). You can estimate the mean μ and the variance σ^2 this way:

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

$$s^2 = \sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N-1}$$

You need at least 2 measurements to estimate s^2. Hence you have N-1 degres of freedom.

As $N \to \infty$, $\bar{x} \to \mu$ and $s^2 \to \sigma^2$

The standard deviation is s. The standard deviation is a common estimator for statistical error.
I'll use a bar to denote estimates …

Binomial Distribution

Imagine an experiment that can only have two outcomes. The success outcome has probability p and the fail outcome has probability 1-p.

The probability of obtaining r successes after N independent tries is given by the binomial distribution:

$$P(r) = \frac{N!}{r!(N-r)!} p^r (1-p)^{N-r}$$

The average is:

$$\bar{r} = \sum_{r=0}^{N} rP(r) = Np$$

And the variance is:

$$\sigma^2 = Np(1-p)$$

If p is unknown:

$$\bar{p} = \bar{r}/N \qquad\qquad s^2 = \frac{N}{N-1} N \frac{\bar{r}}{N} \left(1 - \frac{\bar{r}}{N}\right)$$

Binomial distribution

Example: Imagine a detector with 1000 channels. Each channel has a noise rate of 1 kHz. You want to know the probability of observing 1, 2, 3, etc. noise hits in separate channels in a time window of 1 microsecond.

Because the readout window is small, then p = 1 kHz x 1 μs = 10^{-3}.

Then: $P(0) = (1-p)^{1000} = 0.368$ $P(1) = \frac{1000!}{999!} p(1-p)^{999} = 0.368$

$P(2) = \frac{1000!}{2 \times 998!} p^2(1-p)^{998} = 0.184$ $P(3) = \frac{1000!}{3! \times 997!} p^3(1-p)^{997} = 0.061$

$\bar{r} = 10^{-3} \times 1000 = 1$ $\sigma^2 = 1000 \times 10^{-3}(1-10^{-3}) = 0.999$

(You can go ahead and try this with Veritas, IceCube, etc …)

Poisson distribution

In the limit $N \to \infty$ and $p \to 0$ such that $Np = \mu$ is constant, the binomial distribution becomes the Poisson distribution.

$$P(r) = \frac{\mu^r}{r!} e^{-\mu}$$

The average value of the Poisson distribution is μ and its variance is μ. (Both of this follow trivially from the binomial distribution values.

This is the basis for the $n \pm \sqrt{n}$ estimate of error in a counting experiment.

Poisson distribution

Example: Imagine a detector with 1000 channels. Each channel has a noise rate of 1 kHz. You want to know the probability of observing 1, 2, 3, etc. noise hits in separate channels in a time window of 1 microsecond.

Because the readout window is small, then p = 1 kHz x 1 μs = 10^{-3}. And thus μ = 1 and s^2 = 1

$$P(0) = e^{-\mu} = 0.368 \qquad P(1) = \mu e^{-\mu} = 0.368$$

$$P(2) = \frac{\mu^2}{2} e^{-\mu} = 0.184 \qquad P(3) = \frac{\mu^3}{3!} e^{-\mu} = 0.061$$

These are the same results than with binomial… You can go ahead and try with small N or with large p and check that the two distributions don't give the same result anymore.

Normal (Gaussian) distribution

When μ is large, the Poisson distribution is well described by a Gaussian distribution of variance μ.

For arbitrary mean μ and variance σ^2:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Central Limit theorem

The mean of a sufficiently large number number N of independent random variables, each with finite mean μ and variance σ^2, will be normally distributed. The mean of the Gaussian will be μ and the variance σ^2/N.

Clearly the Central limit theorem explains why repeating measurements is a good idea, and why using a normal distribution is correct in estimating the spread of measurements.

A word of caution: measurement spreads are not always normally distributed.

Central Limit Theorem

Example:

Use a ruler to measure the length of a table and you get 99.7 cm you estimate the error of your measurement to be 0.1 cm.

You measure the table 3 times more, each measurement yielding 99.7 ± 0.1 cm.

Applying the central limit theorem yields a measurement for the table of 99.70 ± 0.05 cm.

Think of at least 3 reasons why this is an incorrect application of the central limit theorem.

Some properties of the normal distribution

The height of the curve at $x = \mu \pm \sigma$ is $e^{-1/2} = 0.607$, so the σ is roughly half width at half height for the normal distribution.

Common values of the fractional area under a normal are:

Range	Area	1 - Area
$\mu-\sigma < x < \mu+\sigma$	0.683	0.317
$\mu-1.644\sigma < x < \mu+1.644\sigma$	0.90	0.1
$\mu-2\sigma < x < \mu+2\sigma$	0.9545	0.455
$\mu-2.575\sigma < x < \mu+2.575\sigma$	0.99	0.01
$\mu-3\sigma < x < \mu+3\sigma$	0.99730	0.00270
$\mu-4\sigma < x < \mu+4\sigma$	0.9999366	6.334×10^{-5}
$\mu-5\sigma < x < \mu+5\sigma$	0.9999994	5.733×10^{-7}

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

Some properties of the normal distribution

Range	Area	1 - Area
$x < \mu+\sigma$	0.8413	0.1587
$x < \mu+2\sigma$	0.9772	0.0228
$x < \mu+3\sigma$	0.9987	0.0013
$x < \mu+4\sigma$	0.9999683	3.167×10^{-5}
$x < \mu+5\sigma$	0.999999713	2.867×10^{-7}

This is known as the cumulative distribution function for the normal distribution:

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

P-value

The *p-value* is the probability of observing a *test statistic* at least as extreme as the one actually observed assuming a null hypothesis. The p-value calculation assume that the null hypothesis IS true.

Example. In a counting experiment with large background B, the null hypothesis is well described by the pdf

$$P(x) = \frac{1}{\sqrt{2\pi B}} e^{-(x-B)^2/2B}$$

Recall that in this case the variance is B. If you observed N events, the p-value of N is:

$$p = \frac{1}{\sqrt{2\pi B}} \int_N^\infty e^{-(x-B)^2/2B}\, dx = \frac{1}{2}\left(1 - erf(\frac{N-B}{\sqrt{2B}})\right)$$

The test, statistics doesn't have to be a normal distribution, yet it is common to translate p-value into *sigmas* using the table in the previous slide.

P-value

The practice of using 5 s as a discovery threshold is widespread. This is an arbitrary threshold (which is fine).

You should however define the p-value for discovery a priori.

Sensitivity of a counting experiment

Imagine a detector in which the background, B, is large. Assume that you can somehow measure B experimentally using on/off-time techniques, then a given fluctuation in the on-time region has the significance:

$$Sig = \frac{N_{\rm on} - N_{\rm off}}{\sqrt{N_{\rm off}}} = \frac{S}{\sqrt{B}}$$

This sensitivity is motivated by comparing the standard deviation of the background, $\sqrt{B}$ to the signal.

Li and Ma (1983) have shown that this naïve formula is inappropriate because the uncertainties in the signal and background are ignored. See Li & Ma (1983) equation 17 for a more appropriate calculation.

However the naïve calculation is a very good approximation for small uncertainties. Li & Ma is the de facto standard in Gamma Ray astronomy.

Parameter fitting – least squares

Imagine that you have a set of measurement $y_i + \Delta y_i$ corresponding to a parameter x_i, e.g. resistance (y) as a function of temperature (x). You have hypothetical description of $y(x,\alpha_j)$, where α_j are Nf_{it} unknown parameters of the function.

We can build the χ^2 :

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

Here σ_i are the estimated 1-σ errors from the theoretical estimate. Note that this is different that using $\sigma_i = \Delta y_i$. This latter choice is however often incorrectly used.

Least squares

The best possible values of α_j are obtained by minimizing χ^2 with respect to α_j.

$$\frac{\partial \chi^2}{\partial \alpha_j} = 0$$

In the case of a linear hypothesis (y=a+bx), the minimization is solving a sent of N-N_{fit} linear equations.

The χ^2 value has a probabilistic interpretation. But first note that there are N-Nf_{it} "degrees of freedom". There are N-Nfit independent terms in χ^2, so:

$$\chi^2 \approx N - N_{\mathrm{fit}}$$

Least squares

A very low value of χ^2 indicates suspiciously overestimated errors, a very high value of χ^2 indicates a hypothesis that doesn't describe the data. Many different hypothesis can result in reasonable χ^2 values!

The probability P that a value χ^2 obtained from an experiment with d degrees of freedom is due to chance is:

$$P_{\chi^2,d} = \left[2^{d/2}\Gamma(d/2)\right]^{-1} \int_{\chi^2}^{\infty} t^{d/2-1} e^{-t/2}\, dt$$

(this is actually a cumulative distribution function)

An online calculator is:

http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

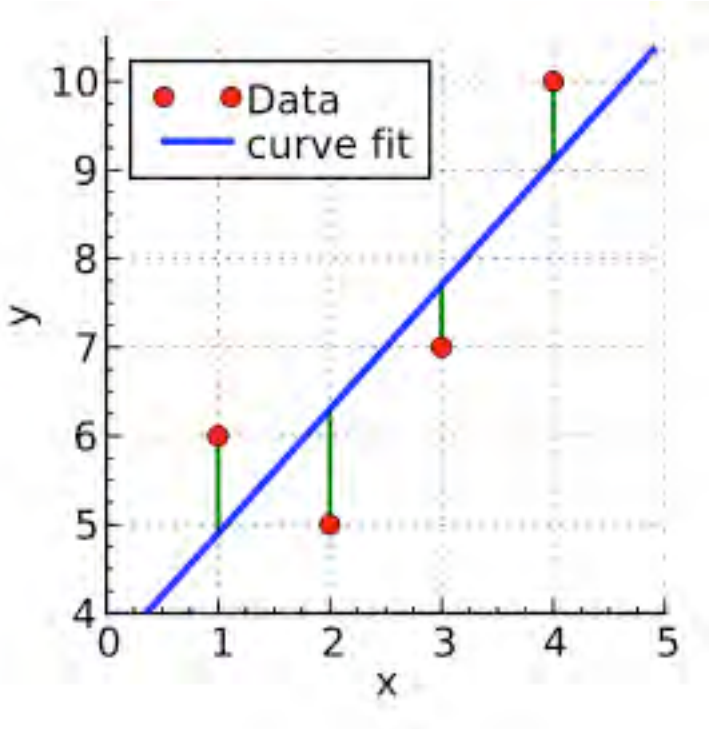
Least squares example

Measurements

X	Y
1	6
2	5
3	7
4	10

Hypothesis:
y = a + b x

Assume error is
$\sigma_i = 1$



The χ^2 is:

$$\chi^2(a,b) = [6-(a+1b)]^2 + [5-(a+2b)]^2 + [7-(a+3b)]^2 + [10-(a+4b)]^2$$

Minimizing with respect to a and b you have a set of 2 equations (N-Nfit), that can be solved (this is just linear algebra)

a = 3.5, b = 1.4, $\chi^2 = 4.2$

The probability of a χ^2 distribution exceeding 4.2 for 2 degrees of freedom is $P_{\chi^2=4.2,\, d=2} = 0.1224$

Parameter fitting – maximum likelihood

Let's study the example of a particle physics interaction leading to an angular distribution of the form:

$$\frac{dn}{d\cos\theta} = a + b\cos^2\theta$$

Let's assume that a and b are unknown.

As a first step we normalize this distribution and transform it into a probability density function:

$$y(a/b) = \frac{1}{2(1+b/3a)}\left(1 + b/a\cos^2\theta\right)$$

By doing this, we note that the pdf is a function of b/a. It's this parameter that we will be able to fit. The overall normalization of dn/dcosθ is not relevant here.

Parameter fitting – maximum likelihood

Let there be i=1,…,N events, each with a measured θ_i angle. Then for each event we can calculate

$$y_i = \frac{1}{2(1+b/3a)}\left(1 + b/a\cos^2\theta_i\right)$$

We define the likelihood as the joint pdf for all events:

$$\mathcal{L}(b/a) = \prod_{i=1}^{N} y_i$$

Maximizing $\mathcal{L}$, provides for the best possible value of b/a assuming that the hypothesis y(b/a) is correct. Observe that the normalization constant of dn/dcosθ depends on b/a – so using a normalized pdf, instead of just any distribution is critical.

Parameter fitting – maximum likelihood

There's no straight forward probabilistic interpretation for the likelihood. If a fit is a good description of the data, then $\mathcal{L}_{max}$ is "large", if it's bad, then $\mathcal{L}_{max}$ is "small". The difficulty relies on determining what is large and what is small.

In some simple cases, a "good" value of $\mathcal{L}_{max}$ can be estimated directly, in others, it is done brute force by finding the distribution of $\mathcal{L}_{max}$ for events that fit the hypothesis.

Note that in practice, the maximization of $\mathcal{L}$ is done numerically. It is usually better to minimize $-\log\mathcal{L}$, that to maximize $\mathcal{L}$ directly. But this is only for numerical convenience.

Parameter fitting – maximum likelihood

Imagine a likelihood function of one parameter, i.e. $\mathcal{L}_{\text{max}}$ (p). The best value of p is found via

$$\frac{d\mathcal{L}}{dp} = 0$$

Near the maximum, the likelihood function is well described by a second order parabola (this follows trivially from Taylor series expansion). The uncertainty in p can be found by how wide or narrow the likelihood is near the maximum

$$\sigma = (-\frac{d^2\mathcal{L}}{dp^2})^{-1/2}$$

Relationship between least squares and likelihood

Let (x_i, y_i) be a data set and y(x) a hypothesis. Assume that the uncertainty of y(x) is normally distributed with constant variance σ^2. The pdf evaluated at x_i is:

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - y(x))^2/2\sigma^2}$$

You can now write a likelihood function for the data set:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - y(x))^2/2\sigma^2}$$

From which it follows:

$$-log(\mathcal{L}) = -N \log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \frac{1}{2} \sum_{i=1}^{N} \frac{(x_i - y(x))^2}{\sigma^2}$$

This looks familiar …

Relationship between least squares and likelihood

The maximization of $\mathcal{L}$ is equivalent to minimizing $-\log\mathcal{L}$. Since the first term in $-\log\mathcal{L}$ is constant, it doesn't matter for minimization.

Maximizing $\mathcal{L}$ is equivalent to (I dropped the factor of ½) minimizing:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - y(x))^2}{\sigma^2}$$

Least squares is mathematically the same as a the likelihood method if you assume normally distributed errors.

Hypothesis testing with least squares

Assume that you have j = 1, …, M (large) experiments, each experiments with a data set of I = 1,…, N data points ($x_{i,j}$,$y_{i,j}$). Assume a hypothesis y(x). Each experiment has a minimum value for least squares χ^2_j, with d degrees of freedom. You can calculate the list of probabilities of of observed c2j exceeding that value.

$$P_{\chi^2_j,d} = \left[2^{d/2}\Gamma(d/2)\right]^{-1} \int_{\chi^2_j}^{\infty} t^{d/2-1} e^{-t/2}\, dt$$

Assuming that the hypothesis y(x) is good and assuming that errors are distributed normally, then P should be distributed uniformly.

Likelihood ratio test

Assume two hypothesis for a counting experiment. For the null hypothesis, and the alternative hypothesis. The null hypothesis is a special case of the alternative hypothesis. Let N be the observed events, n_s (unknown) signal events and N-n_s the background events.

Let $\mathcal{L}$(n_s,N-n_s) be the likelihood for the alternative hypothesis. Then the null hypothesis likelihood is $\mathcal{L}$(0,N). You can define the likelihood ratio:

$$\Lambda = \frac{\mathcal{L}(n_s, N - n_s)}{\mathcal{L}(0, N)}$$

Likelihood ratio test

Maximizing Λ with respect to n_s, yields the most likely value of n_s. The distribution of Λ allows the calculation of a p-value for the observed Λ_{obs}, and thus determining a criteria for which hypothesis is more likely. In practice the distribution of Λ is obtained via simulations or from data known to be well described by background only.

Numerically, it is often more convenient to minimize $-\log\Lambda$ than to maximize Λ.