The Central Limit Theorem and the Method of Least Squares

Christian B. Molina

July 13, 2021

Abstract

Taking measurements and comparing the results to theoretical expectations is a major part of science and engineering. Non-ideal conditions, limited measurement instrumentation, the nature of the physical phenomena being measured, and dealing with unknown physical phenomena during experimentation and measurement results in uncertainties and data that may be imperfect and not precise. Because of this, it may be necessary to make incomplete conclusions or models of the problem. Some of these models may contain parameters that are determined empirically. Statistical analysis like the Central Limit Theorem and the Method of Least Squares make it possible to derive such parameters and make progressive conclusions.

The Central Limit Theorem states that if measured variables have a large enough number of independent degrees of freedom, the distribution of these variables goes asymptotically to a normal distribution. Distributions reflect the stochastic nature of the uncertainties associated with the measurements and the nature of the quantity itself. The Central Limit Theorem implies that the same probabilistic and statistical methods applicable to normal distributions can be applied to other types of distributions.

The Least Squares Method approximates the solution of over-determined systems by minimizing the sum of the squares of the residuals made within the results of the equations. This can be utilized to approximate the parameters of the resulting measurement. In this report, we shall discuss the application of the linear least-squares fit to a straight line.

1 Introduction

1.1 The Normal Distribution and The Central Limit Theorem

In order to understand the application of the Central Limit Theorem, we must first identify the key characteristics of the Normal Distribution. Distributions reflect the stochastic nature of the uncertainties associated with the measurements and the nature of the quantity itself.

The true value of a quantity is given by the average of a large number of measurements The average value in the limit where the number of measurements is infinite:

$$\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i \tag{1}$$

If the distribution is continuous and the probability of finding a value x in the interval dx is $p(x)dx$, the **mean** is given as an integral:

$$\mu = \int_a^b x * p(x)dx \tag{2}$$

Normalizing $p(x)$ would then result in:

$$\int_a^b x * p(x)dx = 1$$

Median($\mu_{1/2}$): the value where the probabilities of finding x below and above it are the same.

- $p(x < \mu_{1/2}) = p(x > \mu_{1/2})$

Most probable value($\mu_{\max}$): given by the value x where the probability distribution attains maximum value

- $p(x = \mu_{\max}) \geq p(x \neq \mu_{\max})$

variance (σ^2) for discrete distribution and continuous distribution:

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

$$\sigma^2 = \int_a^b (x - \mu)^2 * p(x)dx$$

standard deviation = $\sqrt{\sigma^2} = \sigma$

The central limit theorem states that if a variable x has a large number of independent degrees of freedom, the distribution of x goes asymptotically to a normal distribution.
Example: Consider a number x that is made of the sum of **n** random numbers. If n is large, the distribution of x has the form of the Normal Distribution, independent of the distribution of the n random numbers it is made of. Normal or Gaussian Distribution:

$$p(x)dx = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \tag{3}$$

If the mean and variance of each source of error are known, it is possible to deduce the corresponding quantities for the parent distribution. This is normally an impossible task, but with the following approach we can make a good estimate. If the sample measurements represents a random selection among the members in the parent distribution, the measured values must also distribute normally with the mean (μ) and variance (σ) that are good estimates of the corresponding values in the parent distribution.
Since we are assuming that the measured values follow a normal distribution, the probability for the mean of the parent distribution to be in the range $[\mu - \sigma, \mu + \sigma]$ is given by the integral A(x) for $x = \sigma$

$$A(\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} P_G(\mu, \sigma, t)dt = 0.68$$

This tells us that 68% of the measurements made will be within the range of $[\mu - \sigma, \mu + \sigma]$, while the remaining 32% will either be above $\mu + \sigma$ or below $\mu - \sigma$

1.2 The Method of Least Squares

The method of least squares is an application of the method of maximum likelihood. Given that y is a function of x with m parameters $a_1, a_2...a_m$, where

$$y = f(a_1, a_2...a_m; x)$$

What are the values of the parameters that give the best description to a set of measured values of y - in other words: knowing the measured values y but not knowing the parameters that allow y to be, how can we find those parameters?

Let y_i be the measured value of y at $x = x_i$. The central limit theorem states that the more measurements we make, the results are expected to follow a normal distribution if the measurements for y_i are carried out many times.

$$p(y_i)dy_i = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\{-(y_i - f_i)^2/2\sigma_i^2\}dy_i$$

Therefore, the mean of the distribution $p(y_i)dy_yi$ can be described as:

$$f_i = f(a_1, a_2...a_m; x_i)$$

The last piece of $p(y_i)$ is the variance (σ_i^2), but we cannot deduce that without given any more information. For now we shall assume that σ_i^2 is given to us.

The likelihood function for the parameters $a_1, a_2, ..., a_m$ is given by the product of $p(y_1), p(y_2), ..., p(y_N)$:

$$\begin{aligned} L(a_1, a_2, ..., a_m) &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\{-(y_i - f_i)^2/2\sigma_i^2\} \\ &= \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}(\frac{y_i - f_i}{\sigma_i})^2\right\} \end{aligned}$$

The summation portion in the exponent is identified as Chi-square, a fundamental component to the method of least squares:

$$\chi^2 = \sum_{i=1}^{N}(\frac{y_i - f_i}{\sigma_i})^2$$

The partial derivative of Chi-square is the maximum of the likelihood distribution $L(a_1, a_2, ..., a_m)$ for $k = 1, 2, ..., m$

1.2.1 Least-Squares fit to a Straight Line

There needs to be an explicit functional depedence on f on the parameters a_k. For a linear function involving two parameters with $a_1 = a$ and $a_2 = b$ we have:

$$f(a, b; x) = a + bx$$

The partial derivatives of $f(a, b; x)$ with respect to the parameters a and b are then

$$\frac{\partial f}{\partial a} = 1 \qquad\qquad \frac{\partial f}{\partial b} = x$$

Therefore the maximum likelihood condition in this case becomes:

$$\sum_{i=1}^{N}(\frac{y_i - f_i}{\sigma_i^2}) = 0 \qquad\qquad \sum_{i=1}^{N}(\frac{y_i - f_i}{\sigma_i^2})x_i = 0$$

or:

$$a\sum_{i=1}^{N}\frac{1}{\sigma_i^2} + b\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} = \sum_{i=1}^{N}\frac{y_i}{\sigma_i^2} \qquad\qquad a\sum_{i=1}^{N}\frac{x_i}{\sigma_i^2} + b\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} = \sum_{i=1}^{N}\frac{x_iy_i}{\sigma_i^2}$$

This is known as linear regression, where a sample of N data points are fitted to the relation of $f(a, b; x) = a + bx$. This will also be known as *linear least-squares fit to a straight line.*

In order to solve for a and b in the two equations above, we can set them up in matrix form

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \theta \\ \phi \end{pmatrix}$$

The six elements in the equation are given by:

$$\alpha = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \qquad \beta = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \qquad \gamma = \beta$$

$$\delta = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad \theta = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2} \qquad \phi = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

The values of a and b can be solved using:

$$a = \frac{1}{D} \det \begin{vmatrix} \alpha & \beta \\ \phi & \delta \end{vmatrix} = \frac{1}{D}\{\theta\delta - \beta\phi\}$$

$$b = \frac{1}{D} \det \begin{vmatrix} \alpha & \theta \\ \gamma & \phi \end{vmatrix} = \frac{1}{D}\{\alpha\phi - \theta\gamma\}$$

The value of the determinant D is then:

$$D = \det \begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} = \alpha\delta - \beta\gamma = \alpha\delta - \beta^2$$

1.2.2 Uncertainties in the Parameters

In the previous section, we discussed the maximum likelihood values of a and b. The uncertainties of these parameters that arise from y_i must be included

In lesser detail than described above and in a similar fashion, we can obtain the uncertainty of a and b as:

$$\sigma_a^2 = \frac{\delta}{D}$$

$$\sigma_b^2 = \frac{\alpha}{D}$$

There is also the covariance between a and b:

$$\sigma_{a,b}^2 = -\frac{\beta}{D}$$

2 Methods

2.1 Algorithm: Normal Distribution and The Central Limit Theorem

In the algorithm below, we will use a random number generator with an even distribution in the range [-1,+1] to produce n=6 values and store the sum as x. 1000 such sums will be collected and their distribution will be plotted. We will compare the results with a normal distribution of the same mean and variance as the x collected, and then calculate the χ^2-value. The algorithm and calculations will be repeated with n=50, then we shall compare the two χ^2-values obtained. The following algorithm is initially written in python utilizing the tools *chisquare* and *norm* from the library *scipy.stats*

Algorithm 1: The Central Limit Theorem and Normal Distribution

```
Result: ChiSquared
INIT: x, sigma, mu, sigma;
seed: to initialize random.random();
n_bins = 25;
for i in range(1000) do
    j,rand_sum = (0,0);
    while j < n do
        rand_sum = rand_sum + r.uniform(-1,1);
        j+=1
    end
    x.append(rand_sum)
end
# Plot histogram and calculate chi-squared;
mu, sigma = norm.fit(x);
hist_data = plt.hist(x,n_bins);
ChiSquared = chisquare(hist_data[0]);
# Generate Normal Distribution;
xmin, xmax = plt.xlim();
xval = np.linspace(xmin, xmax, 100);
pdf = norm.pdf(xval, mu, sigma);
plt.plot(xval,pdf);
```

The code above begins by initializing the seed for the random number generator and the variables x, σ, and μ. For the histogram, 25 bins will suffice to show the normalization of the randomly generated sums. The `for-loop` collects 1000 summations of n random values, appending each sum to x.

`plt.hist`, `norm.fit`, and `chisquare` are tools from the libraries `matplotlib.pyplot` and `scipy.stats`. `norm.fit` is used to obtain μ and σ. `plt.hist` generates observed frequency within the data x and separates them into 25 bins. With the observed frequency, `chisquare` can be used to run a one-way chi-square test.

Lastly for the normal distribution, we can use `norm.pdf` to generate a normalized probability distribution function using μ and σ.

2.2 Algorithm: Linear Least-Squares Fit to a Straight Line

Algorithm 2 will implement the method of linear least-squares fit to a set of data $[x_i, y_i]$ to find the parameters a and b of $y_i = ax_i + b$. Each point will be given a standard deviation σ_i. $\alpha, \beta, \delta, \gamma, \phi$ and θ will be calculated in order to find the parameters and then a line fitted to the data will be plotted.

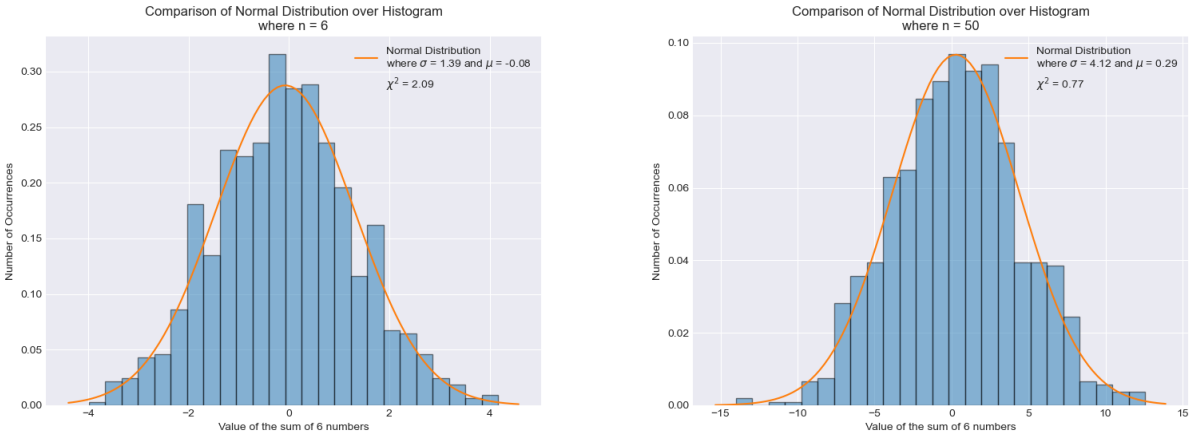
Algorithm 2: Linear Least-Squares Fit to a Straight Line

```
Result: a and b
integralResults: list of zeros with size 3;
x, y and sigma are given;
INIT omega;
for i in sigma do
    if i==0 then
        omega.append(1)
    else
        omega.append(1/i**2)
    end
end
alpha, delta = (np.sum(1 / sigma**2), np.sum((x / sigma)**2));
gamma = beta = np.sum(x / sigma**2);
theta, phi = (np.sum(y / sigma**2), np.sum(x*y / sigma**2));
det_D = alpha*delta - beta**2;
a, b = ((theta*delta - beta*phi) / det_D, (alpha*phi - theta*gamma) / det_D);
sigma_a = (delta / det_D)**(1/2) ;
sigma_b = (alpha / det_D)**(1/2);
sigma_ab = -beta / det_D ;
```

Algorithm 2 is a direct translation of the equations introduced in subsection 1.2.1.

3 Discussion

3.1 Normal Distribution and The Central Limit Theorem



(a) n = 6 (b) n = 50

Figure 1: Histogram of randomized data versus their normalized probability distribution functions

In figure 1 two instances of Algorithm 1 are run - n = 6 and n = 50. When $n = 6$, we can see slight discrepancies in comparison to the calculated normal distribution. $n = 50$ on the other hand shows the distribution of the data following closely to that of its normal distribution. The standard deviation of $n = 6$ is $\sigma = 1.39$ and the mean is $\mu = -0.08$, while the standard deviation of $n = 50$ is $\sigma = 4.12$ and the mean is $\mu = 0.29$ Most importantly, minimizing the value of χ^2 indicates that the observed data with no dependence on chance follows closely towards the expected data. For $n = 6$, $\chi^2 = 2.09$ and for $n = 50$, $\chi^2 = 0.77$. Thus taking more measurements decreases the value of χ^2.

This also shows the claim of the Central Limit Theorem, in which if measured variables have a large enough number of independent degrees of freedom, the distributions of these variables goes asymptotically to a normal distribution.

3.2 Linear Least-Squares Fit to a Straight Line

Table 6-2: A sample of measured values of y for different values of x.

i	x_i	y_i	σ_i	i	x_i	y_i	σ_i
1	0.25	0.86	0.27	6	3.64	8.84	0.66
2	1.05	2.18	1.16	7	3.92	8.71	0.98
3	2.25	4.84	1.14	8	4.94	11.98	0.93
4	2.88	5.80	0.93	9	5.92	12.40	0.60
5	2.97	6.99	0.31				

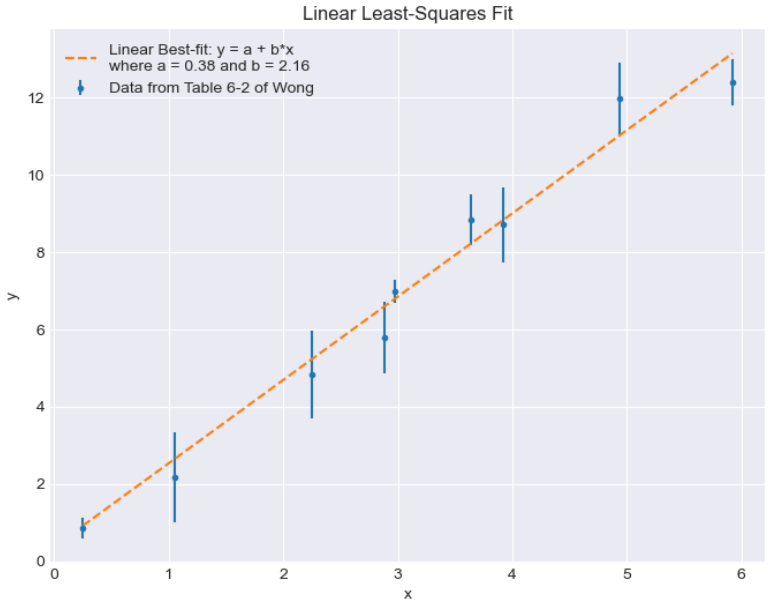


Figure 2: Linear Least-Squares fit to a set of data

For the application of Algorithm 2, we will be using the data from Table 6-2 of the Wong textbook. Table 6-2 presents the measured values x_i and y_i as well as the errors in measurements σ_i. In Figure 2, we can see the results of the linear least-squares fit of the data. This gives us a complete picture of the methods utilized in section 1.2 of the Introduction.

4 Conclusion

In this report, we presented and outlined Normal Distributions and the Central Limit Theorem, and the Method of Least Squares. We then presented algorithms that translate these methods into viable code that can be applied to analyze measurements to generate information and data that expands the utility of the measurements.

References

[1] S.S.M Wong.*Computational Methods in Physics and Engineering*