Central Limit Theorem: Statistical Descriptions of Data and Linear Least Square Fit

B. C. Chap [*]

Winter Quarter 2021

Course: Computational Methods of Mathematical Physics
Instructor: Professor D. Ferenc
University of California, Davis
One Shields Ave., Davis, CA 95616, USA

Abstract

When taking measurements, it is common to first determine what the expected behavior of the data is. We consider theory and previous results to do this. In physics, we derive or measure these values consistently under various circumstances, until an accepted value with uncertainty is established. We can determine to what degree of accuracy and precision given the uncertainties associated with the measurements given a large enough data set. It may be nearly impossible to deduce the distribution of data based on theory alone, thus multiple measurements are necessary to determine it. In this project we explore the Central Limit Theorem and its results to examine uniform and normal distributions of data and how the method of least squares is used to determine a linear fit of observed data with associated uncertainties.

1 Central Limit Theorem

The Central Limit Theorem (CLT) states that as sample size gets larger, regardless of how the source population is distributed, the sampling distribution of the sample means approaches a normal distribution. This entails that for a sufficiently large number of observed events, a normal distribution will be observed [1].

Probability distributions describe how likely different possible outcomes of an event are. In terms of data, they describe how measurements are spread. Probability distributions are characterized by a few parameters: the mean, the standard deviation, and the variance. The mean aka the average of the values, determines where the distribution is center as is denoted μ. The standard deviation, denoted σ, embodies how spread out the numbers are, and the variance denoted σ^2 is the average of the squared differences of the observed data from the mean [2]. The mean and standard deviation of the probability distribution for a variable x of sample size N can then be calculated as:

$$\mu_{\bar{x}} = \mu \tag{1}$$

[*] bcchap@ucdavis.edu

$$\sigma_{\bar{x}} = \sigma/\sqrt{N} \tag{2}$$

The CLT holds true when either of two conditions are met: the sample population is normal or the sample size is sufficiently large [1]. This will be examined later via Problem 6.3 of Computation Methods of Physics by S.M. Wong [2].

Suppose a series of measurements with their associated uncertainties are made ($x_i \pm \sigma_i$). To quantify how close the observed data is to theoretical predictions a_i, the χ^2 aka "chi-squared" can be calculated when uncertainties are unkown is:

$$\chi^2 = \sum_i \frac{(a_i - x_i)^2)}{x_i} \tag{3}$$

The closer the observed values are to the expected, the smaller the χ^2, thus large values of χ^2 entail that the results are unlikely to have occurred. The terms small and large can seem ambiguous, and this ambiguity is further examined in the following section discussing Problem 6.3 as well.

1.1 Problem 6.3

Problem 6.3 is as follows: **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. Collect 1000 such sums and plot their distribution. Compare the results with a normal distribution of the same mean and variance as the x collected. Calculate the χ^2 value. Repeat the calculations with n = 50. Compare the two chi^2 obtained.**

1.2 Procedure

To solve this problem, we must to create uniformly distributed data for n = 6 and n = 50, compare them both to Gaussian distributions, and calculate their χ^2 values. The procedure to accomplish this is as follows:

Argument List:

- n: number of values to be created
- nsum: the number of sums of n values
- lower_bound, upper_bound: interval for the random number generator
- bins: number of data groupings
- mu: mean value for the normal distribution
- sigma: standard deviation of the normal distribution

1. Uniform Probability Distribution: interval = [-1,1]

2. Create an array of n number of values with uniformly random generated numbers

3. Create an array where each value sums the n number of values previously generated $nsum$ times

4. Put the values of $nsum$ into a histogram with $bins$ bins

5. Plot the histogram

6. Normal Probability Distribution: $\mu = 0$, $\sigma =$ (upper_bound - lower_bound) / 4, interval = [-1,1]

7. Plot a normal distribution centered at μ with standard deviation σ on the same plot as the uniform histogram

8. Calculating χ^2

9. Use Equation 3 to determine the χ^2 value

1.3 Discussion

Problem 6.3 is exemplifies the CLT. While the problem asks us to analyze n = 6 and n = 50, the CLT is most apparent when examining an even smaller n value. For this reason, I have included n = 1, n = 6, and n = 50 to examine the results of the CLT in its entirety.

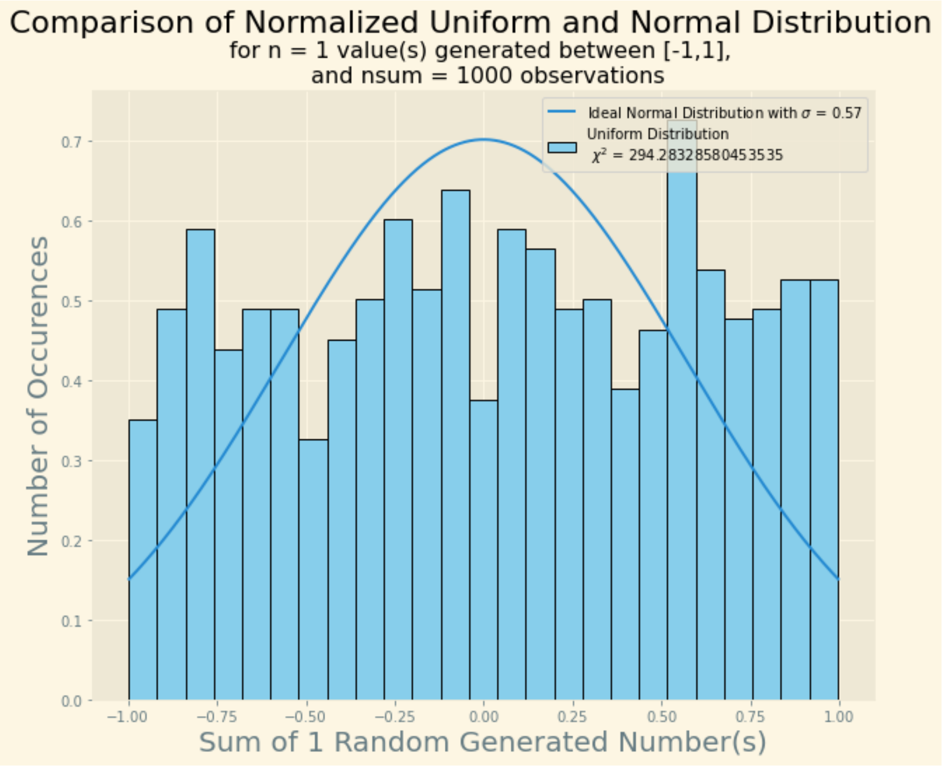

Figure 1: Normalized histogram showing the distribution of $nsum = 1000$ sums of $n = 1$ values

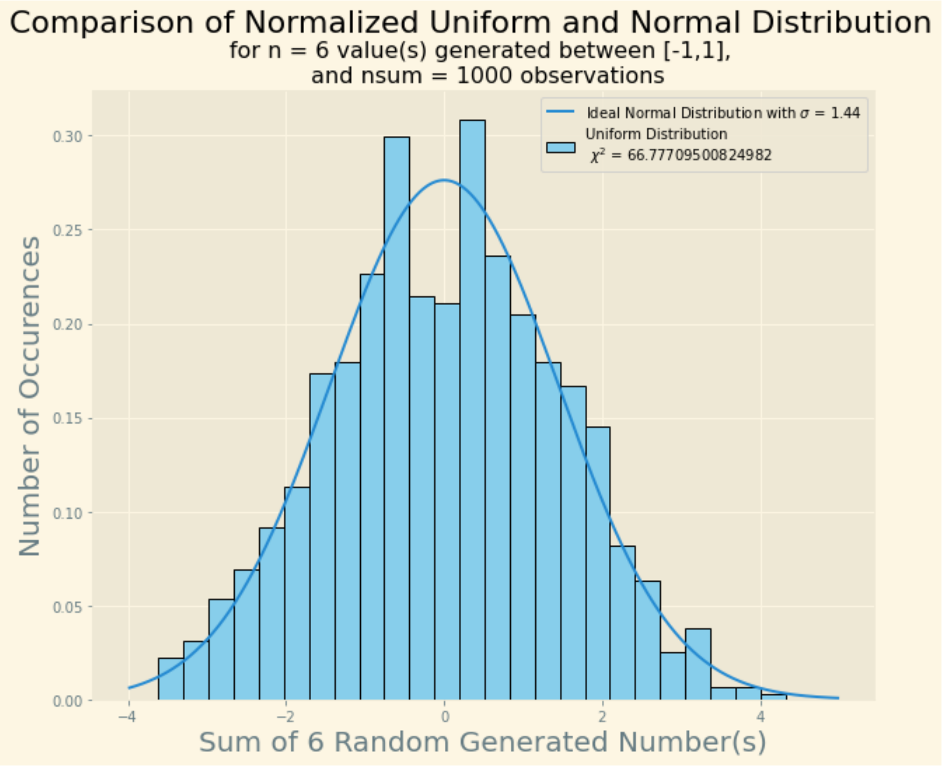

Figure 2: Normalized histogram showing the distribution of $nsum = 1000$ sums of $n = 6$ values

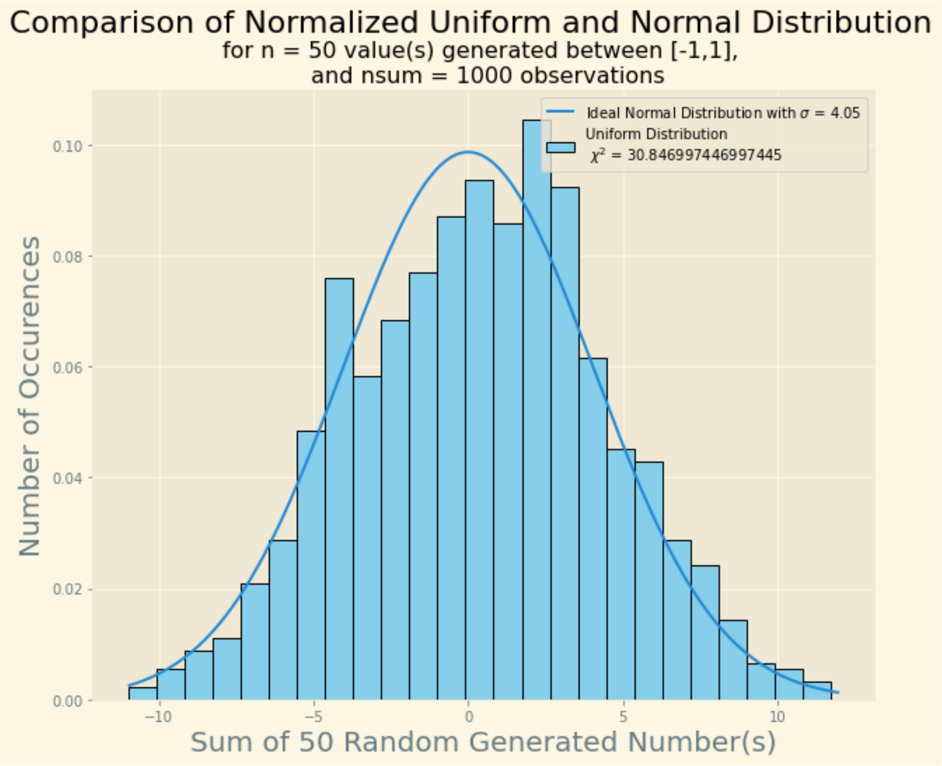

Figure 3: Normalized histogram showing the distribution of $nsum = 1000$ sums of $n = 50$ values

It was previously stated that either of the distribution of values will approach a normal distribution if one of two conditions are met: either the population data is normal or there is a sufficiently large population. Using a uniform random number generator, we are able to examine the latter. For n = 1, we see that the distribution is not Gaussian at all, it is uniform as expected. However, as n increases, we see that the distribution becomes more Gaussian. This is also exemplified by the χ^2 value. Larger χ^2 correspond to the data being less Gaussian, whereas the smaller χ^2 correspond to more Gaussian data. Overall, this confirms the CLT. Regardless of the distribution of the population, for a sufficiently large population, the distribution approaches a Gaussian distribution.

2 Methods of Least Squares

Supposed we have a sample of measured values of y for different values of x with uncertainties, as is the case for Box 6.1 of Wong. To fit a linear line to this data, we first assume that there is a linear relationship between the two variables:

$$y = m * x + b \tag{4}$$

To find the parameters m and b that would best fit the data, we can use the method of least squares. Similarly to examining probability distributions, the best fit values of m and b yield the minimum χ^2 possible, 0, thus satisfying what is known as the maximum likelihood condition. For known uncertainties, χ^2 is as follows:

$$\chi^2 = \sum_i \frac{(a_i - x_i)^2)}{\sigma_i^2} = 0 \tag{5}$$

which leads to a system of equations that can be solved to determine the best fit parameters m and b which will be. This system in matrix form is as follows:

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

where

$$\alpha = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \qquad \beta = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \qquad \gamma = \beta$$
$$\delta = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad \theta = \sum_{i=1}^{N} \frac{y_i^2}{\sigma_i^2} \qquad \phi = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \tag{6}$$

The values of m and b can then be determined by:

$$m = \frac{1}{D} det \begin{vmatrix} \theta & \beta \\ \phi & \delta \end{vmatrix} = \frac{1}{D}(\theta\delta - \beta\phi) \tag{7}$$

$$b = \frac{1}{D} det \begin{vmatrix} \alpha & \beta \\ \gamma & \phi \end{vmatrix} = \frac{1}{D}(\alpha\phi - \theta\gamma) \tag{8}$$

and the value of the determinant D is:

$$D = \begin{vmatrix} \alpha & \beta \\ \gamma & \delta \end{vmatrix} = \alpha\delta - \beta\gamma = \alpha\delta - \beta^2 \tag{9}$$

Note that the uncertainties used are associated with y_i, but it is entirely possible to determine the uncertainties of the fit parameters. For the scope of this project, determining the uncertainties of the fit parameters is the extent, however a greater linear fit can be determined by also accounting for the uncertainties in each variable. Assuming that the uncertainties in each variable are uncorrelated and that the covariance between the two of them vanishes, the following are the uncertainties of m and b:

$$\sigma_m^2 = \frac{\delta}{D} \tag{10}$$

$$\sigma_b^2 = \frac{\alpha}{D} \tag{11}$$

with a covariance of:

$$\sigma_{a,b}^2 = -\frac{\beta}{D} \tag{12}$$

This method is exemplified in the following section through completing Box 6.1 of Wong [2].

2.1 Box 6.1

The following table is taken Table 6-2 taken of Wong, which is a sample of measured values of y for different values of x:

Table 1: Measured x and y data with corresponding uncertainties from Table 6-2 Wong [2]

x_i	0.25	1.05	2.25	2.88	2.97	3.64	3.92	4.94	5.92
y_i	0.86	2.18	4.84	5.8	6.99	8.84	8.71	11.98	12.4
σ_i	0.27	1.16	1.14	0.93	0.31	0.66	0.98	0.93	0.6

and the procedure to apply the linear least-squares fit to a straight line for this data is outlined in the following section.

2.2 Procedure

Argument List:

- x_i: array of x values
- y_i: the number of sums of n values
- sigma: uncertainty of y_i

1. Weight each point

    2. This is unneccesary for this problem as there are no sigma values of 0

3. Calculate α, β, δ, γ, θ, and , ϕ using Equations (6)
4. Compute the determinant D using Equations (9)
5. Calculate the best values of m and b using Equations in (8) and their associated uncertainties using Equations (10) and (11)
6. Calculate the covariance using Equation (12)

2.3 Discussion

The following table outlines the resulting values calculated via the procedure for the data outline in Table 1.

Table 2: Calculated Values of the Linear Least Squares Fit for Table 1

α	34.0630
β	74.7702
γ	74.7702
δ	278.9344
θ	174.2428
ϕ	630.1284
D	3910.7600
m	2.1571
b	0.3804
σ_m	0.0933
σ_b	0.2671
$\sigma^2_{m,b}$	-0.0191

It is also useful to examine the linear fit to the data graphically, which is depicted as:

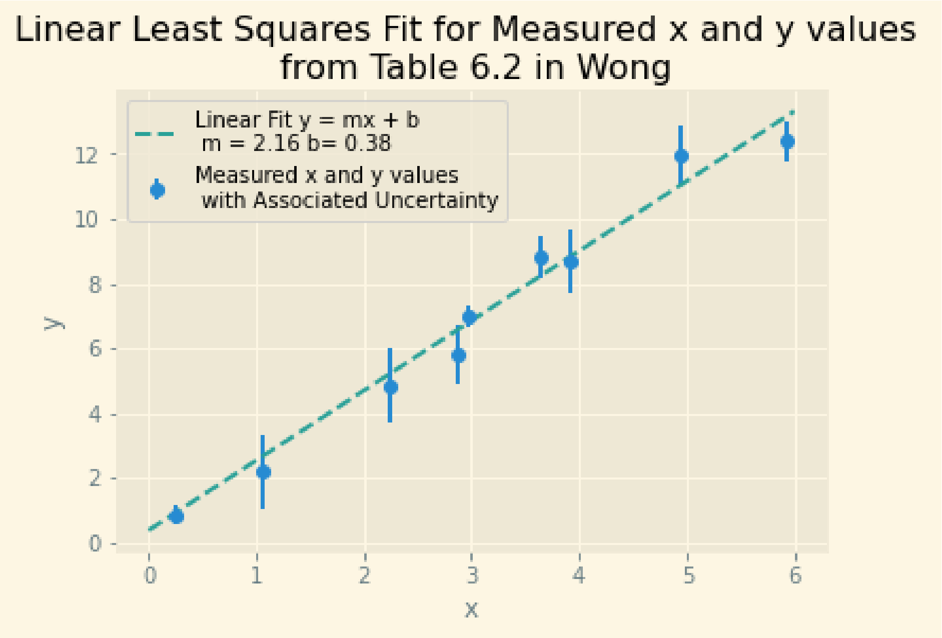

Figure 4: Data from Wong Table 6.2 fitted to a linear line using the method of least squares fit

Based on the graphics alone, the method of least square fits is a straightforward and useful approach to determining the relationship between two dependent variables which also accounts for their uncertainties.

3 Conclusion

In this project, we examined the CLT and the Method of Least Squares Fit to determine a linear fit to data and their associated uncertainties. As expected, the CLT holds true for large values of n and regardless of how the initial population is distributed. Specifically, the distribution of data approaches a Gaussian distribution regardless of the sample population distribution. For the Method of Least Squares Fit, a linear line with parameters m and b was able to be determined given a data set of measured x and y values with known uncertainties, which is what was expected. Overall, the CLT and Method of Least Square Fit provide us with ways into examining distributions of data and determing appropriate fits for them.

References

[1] LaMorte Wayne W. *Central Limit Theorem.* URL: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/BS704_Probability12.html.

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