A PEDAGOGICAL NOTE ON THE CENTRAL LIMIT THEOREM

David J. Jeffery[1]

ABSTRACT

The central limit theorem (CLT) is a basic result of statistics. It is useful for students not specialists in statistics (e.g., astronomy and physics students) to consider some features of the theorem and see a non-rigorous proof of it that gives insight. Rigorous general proofs are found elsewhere.[2] Without experience with the techniques involved, they give little insight. Here the goal is insight, not rigor. Probably nothing in this note is novel: it is difficult to say anything novel given the long history history of the central limit theorem with origins going back Abraham de Moivre (1667–1754) in 1733 (e.g., Fischer 2011, p. 2).

Subject headings: methods: data analysis — methods: statistical

1. THE CENTRAL LIMIT THEOREM

The central limit theorem (CLT) has many versions all with their own special importances and limitations.[3] Here, we will just consider a commonly-thought-of version which we will just call the CLT without qualification for simplicity.

The CLT states

$$x = \frac{1}{n}\sum x_i \tag{1}$$

has a Gaussian distribution with mean

$$\mu = \frac{1}{n}\sum_i \mu_i \qquad \text{and standard deviation} \qquad \sigma = \sqrt{\frac{1}{n}\sum_i \sigma_i^2}\ , \tag{2}$$

[1]Department of Physics & Astronomy, University of Las Vegas, Nevada, 4505 S. Maryland Parkway Las Vegas, Nevada 89154, U.S.A.

[2]e.g., Wolfram MathWorld: Central Limit Theorem; Wikipedia: Central limit theorem.

[3]e.g., Wikipedia: Central limit theorem.

where the sums are over $n \to \infty$, the x_i are independent random variables which have arbitrary probability distributions, the μ_i are the means of the x_i distributions, and the σ_i are standard deviations of the x_i distributions.[4] The CLT holds approximately for finite n (i.e., x has approximately a Gaussian distribtion), except that it is exact for finite n in the special case that all the x_i have Gaussian distributions.

There are some limiting conditions for the CLT. The most obvious is that the standard deviations exist for the x_i. There are probability distributions where the variance (i.e., the square of the standard deviation) diverges to infinity: the best known one of these is probably the Lorentzian distribution (e.g., Bevington 1969, p. 50–51). A second condition is that the μ_i and σ_i are bounded, and so do not go infinity as $n \to \infty$. We will also assume for our non-rigorous proof that the x_i that make non-zero contributions repeat infinitely often. This assumption is not needed for the CLT in general. Given our assumption any probability distributions that do not repeat infinitely make zero contribution as $n \to \infty$. Of course, for the approximate CLT for finite n, non-infinitely repeating x_i do contribute to x. This is an essential point since in real applications of the CLT n will usually be finite, albeit usually large in some sense.

Why is the central limit theorem of interest? First, it makes sense of why many processes, which are themselves compounded of many elementary random processes, have Gaussian probability distributions at least approximately. Second, it allows us to trust the assignment of Gaussian errors to experimental results with confidence insofar as the results follow from processes alluded to in the first point. Third—for those who are fascinated by it—the history of the development of the central limit theorem is a significant part of the history of mathematics (e.g., Fischer 2011).

One also has to say that central limit theorem seems to be a basic aspect of reality that can be pondered on. One ponderation is that the CLT may or may not have a deep connection to quantum mechanics. The Gaussian wave function (i.e., the wave function that is a Gaussian) in quantum mechanics is the minimum uncertainty wave function: i.e., the one and only one for which the Heisenberg uncertainty principle

$$\sigma_x \sigma_p \geq \hbar/2 \tag{3}$$

holds as an equality (e.g., Griffiths 2005, p. 111, 113). To elucidate briefly: σ_x is the standard deviation in the x position of a particle, σ_p is the standard deviation in the x-direction momentum of the particle, and $\hbar$ is Planck's constant divided by 2π. The only stationary state of quantum mechanics that is a Gaussian is the ground state of the harmonic

[4] e.g., Wolfram MathWorld: Central Limit Theorem; Wikipedia: Central limit theorem.

oscillator potential (e.g., Griffiths 2005, p. 46, 113). The harmonic oscillator potential is one the most fundamental in nature.

In this note, the goal is to gain insight into the central limit theorem by considerations various features of it and by giving a non-rigorous proof: i.e., a proof with loopholes that are in closed in some way by rigorous proofs. As remarked in the abstract, rigorous proofs can be found elsewhere, but without experience with the techniques involved, they give little insight.

In § 2, we consider some vague questions about the CLT. Section 3 proves the CLT results given by equation (2). In § 4, we non-rigorously prove the CLT holds exactly for finite n when the random variables x_i have Gaussian probability distributions. In § 5, we show the connection of the CLT to the binomial probability distribution. Section 6 gives the non-rigorous proof of the CLT making use of results established in §§ 3, 4, and 5, and the Hagen derivation of the Gaussian. The conclusion is in § 7.

2. VAGUE QUESTIONS, PLAUSIBLE ANSWERS

At first sight, four questions about the CLT for $n \to \infty$ occur to the author:

1. Say all the x_i had a common probability distribution which was not Gaussian. How could the sum x have a Gaussian distribution and not the common distribution? Phrasing the question almost answers question: the particular x_i values that go to into the sum for the x come from different uncorrelated parts of the common distribution, and so do not just reproduce the common distribution.

2. Gaussians are symmetric. So what if all the x_i distributions are asymmetric with an apparent bias to high or low? The means mean something even for asymmetric distributions, and so it is plausible as $n \to \infty$ that that will override the asymmetries.

3. Say one of the x_i had a much larger scale (i.e., much larger μ_i and σ_i) than all the others? Would it not dominate behavior of the sum and enforce its distribution on the sum? No. When $n \to \infty$, no one large-scale x_i and, similarly, no finite number of large-scale x_i can dominate the sum. In fact, they would contribute nothing.

   Note, however, that if n is finite then one probability distribution or a few of the probability distributions can dominate the behavior of the sum x and prevent the emergence of Gaussian behavior.

4. Say a fixed fraction of identically-distributed x_i dominated the sum x. Would this class not enforce its distribution on the sum x? No. See the answer to question 1.

3. MEANS, VARIANCES, AND STANDARD DEVIATIONS

The CLT results of equation (2) are easy to prove. Average over

$$x = \sum x_i \qquad \text{and} \qquad (x-\mu)^2 = \sum (x_i - \mu_i)^2 \tag{4}$$

to obtain

$$\mu = \sum \mu_i \ , \qquad \text{and} \qquad \sigma^2 = \sum \sigma_I^2 \qquad \text{and} \qquad \sigma = \sqrt{\sum_i \sigma_i^2} \ . \tag{5}$$

One can for simplicity use alternate form $y_i = x_i - \mu_i$ for the set of random variables. Thus, one has

$$y = \sum y_i \qquad \text{and } \sigma \text{ and } \sigma^2 \text{ are unchanged.} \tag{6}$$

The results of this section hold for any n, and thus for $n \to \infty$ provided the sums converge. So the mean and standard deviation part of the CLT is proven generally.

4. GAUSSIAN RANDOM VARIABLES

Here we prove non-rigorously that the CLT holds exactly for both finite n and $n \to \infty$ for Gaussian random variables x_i.

Consider two independent random variables u and v both having Gaussian distributions with variances σ_u^2 and σ_v^2. For simplicity, we assume u and v are measured relative to their respective means. We sum them: $w = u + v$. Somewhat abstractly (and choosing to use v as the integration variable without loss of generality), we can write

$$P(w = u + v) = P(u|v)P(v) = P(w - v|v)P(v) \tag{7}$$

and now integrate over all v to get the probability of w for any general outcome. To be less abstract by using probability densities, we now write

$$g(u)h(v)\,dv\,du = g(w-v)h(v)\,dv\,dw \ , \qquad \text{and so} \qquad \rho(w)\,dw = \int_{-\infty}^{\infty} g(w-v)h(v)\,dv\,dw \ , \tag{8}$$

where ρ is the probability distribution for w,

$$g(u) = \left(\frac{1}{\sqrt{2\pi}\,\sigma_u}\right)\exp\left(-\frac{u^2}{2\sigma_u^2}\right) \qquad \text{and} \qquad h(v) = \left(\frac{1}{\sqrt{2\pi}\,\sigma_v}\right)\exp\left(-\frac{v^2}{2\sigma_v^2}\right) \tag{9}$$

are, respectively, the properly normalized Gaussian probability distributions for u and v (e.g., Bevington 1969, p. 53), and the integral over v is formally a convolution of the two distributions.[5] Defining

$$A = \left(\frac{1}{\sqrt{2\pi}\,\sigma_u}\right)\left(\frac{1}{\sqrt{2\pi}\,\sigma_v}\right) \qquad \text{and} \qquad B = \frac{1}{2\sigma_u^2} + \frac{1}{2\sigma_v^2} \tag{10}$$

to simplify the algebra in evaluating $\rho(w)$, we proceed thusly

$$\begin{aligned}
\rho(w) &= \int_{-\infty}^{\infty} g(w-v)h(v)\,dv \\
&= A\exp\left(-\frac{w^2}{2\sigma_u^2}\right)\int_{-\infty}^{\infty}\exp\left[-\frac{(-2wv+v^2)}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right]dv \\
&= A\exp\left(-\frac{w^2}{2\sigma_u^2}\right)\int_{-\infty}^{\infty}\exp\left[-B\left(v^2 - \frac{wv}{B\sigma_u^2} + \frac{w^2}{4B^2\sigma_u^4} - \frac{w^2}{4B^2\sigma_u^4}\right)\right]dv \\
&= A\exp\left(-\frac{w^2}{2\sigma_u^2}\right)\int_{-\infty}^{\infty}\exp\left\{-B\left[\left(v - \frac{w}{2B\sigma_u^2}\right)^2 - \frac{w^2}{4B^2\sigma_u^4}\right]\right\}dv \\
&= \left(\frac{1}{\sqrt{2\pi}\,\sigma_u}\right)\left(\frac{1}{\sqrt{2\pi}\,\sigma_v}\right)\exp\left[-\frac{w^2}{2\sigma_u^2}\left(1 - \frac{1}{2B\sigma_u^2}\right)\right]\sqrt{\frac{\pi}{B}} \\
&= \frac{1}{\sqrt{2\pi}\,\sqrt{\sigma_u^2+\sigma_v^2}}\exp\left[-\frac{w^2}{2(\sigma_u^2+\sigma_v^2)}\right]
\end{aligned} \tag{11}$$

We define $\sigma_w^2 = \sigma_u^2 + \sigma_v^2$ and, restoring explicit means to u and v, we define $\mu_w = \mu_u + \mu_v$. We now have

$$\rho(w) = \frac{1}{\sqrt{2\pi}\,\sigma_w}\exp\left[-\frac{(w-\mu_w)^2}{2\sigma_w^2}\right]\ . \tag{12}$$

Thus, the convolution of two Gaussians is a Gaussian with the mean being sum of the component means and the variance being the sum of the component variances. And thus w has a Gaussian distribution.

Since the proof just given immediately generalizes to any number of component Gaussians, we have proven what we said we would in the preamble.

Now any probability distribution with a mean and standard deviation can be approximated by a Gaussian to some reasonable accuracy. One chooses the mean and standard

[5] e.g., Wikipedia: Convolution: Definition.

deviation of a to-be-fit probability distribution to be that of the Gaussian replacement or, if one has a continuous probability distribution, one can choose to fit a Gaussian to the maximum of said continuous probability distribution and distort the wings of the Gaussian to preserve normalization. Since one can always do this, it is clear that the finite-n CLT must always be approximately accurate even for a single-member set of random variables x_i (or y_i). Recall the CLT for general probability distributions (subject to some conditions) is only proven to be exactly accurate in the limit that the number of random variables $n \to \infty$.

5. A NON-RIGOROUS PROOF OF THE CENTRAL LIMIT THEOREM PART I: THE BINOMIAL PROBABILITY DISTRIBUTION

First, note for any of the random variables x_i with probability distribution $\rho(x_i)$ and mean μ_i that

$$\mu_i = \int_{-\infty}^{\infty} x_i \rho(x_i)\, dx_i \qquad \text{and so} \qquad 0 = \int_{-\infty}^{\infty} (x_i - \mu_i)\rho(x_i)\, dx_i \tag{13}$$

and so

$$0 = \int_{-\infty}^{\mu_i} (x_i - \mu_i)\rho(x_i)\, dx_i + \int_{\mu_i}^{\infty} (x_i - \mu_i)\rho(x_i)\, dx_i \tag{14}$$

and so

$$\gamma_{i,\mathrm{L}} = \frac{\int_{-\infty}^{\mu_i} |x_i - \mu_i|\rho(x_i)\, dx_i}{1/2} = \frac{\int_{\mu_i}^{\infty} |x_i - \mu_i|\rho(x_i)\, dx_i}{1/2} = \gamma_{i,\mathrm{R}} \tag{15}$$

where $\gamma_{i,\mathrm{L}}$ and $\gamma_{i,\mathrm{R}}$ are mean (absolute) deviations, respectively, to the left and right of the mean. Thus, the mean deviation itself satisfies

$$\gamma_i = \gamma_{i,\mathrm{L}} = \gamma_{i,\mathrm{R}} \ . \tag{16}$$

Consider a trial with outcome x_k which is the sum k values of the x_i to the right of their respective means (right steps) and $(n-k)$ values of the x_i to the left of their respective means (left steps) as indicated by subscripts:

$$x_k - \mu = \sum_{i,k} (x_i - \mu_i) + \sum_{i,(n-k)} (x_i - \mu_i) \ . \tag{17}$$

We now replace each x_i but what it does on average (which is equal contributions from left and right of its mean with equal probability) and then average over infinitely many trials for x_k (recalling that the x_i are independent, and so uncorrelated) to obtain

$$\mu_k - \mu = k\gamma_n - (n-k)\gamma_n = \left(k - \frac{n}{2}\right)(2\gamma_n) \ , \tag{18}$$

where μ_k is the average of x_k and we have defined

$$\gamma_n = \frac{1}{n}\sum_i \gamma_i \tag{19}$$

which must be true since all the x_i are treated equally in the averaging. The probability distribution P_k of the means μ_k for infinitely many outcomes must be the (symmetric) binomial probability distribution—the situation is analogous to flipping a true coin n times. Explicitly,

$$P_k = \binom{n}{k}\left(\frac{1}{2}\right)^n = \frac{n!}{k!(n-k)!}\left(\frac{1}{2}\right)^n \tag{20}$$

with $\langle k\rangle = n/2$ and k variance $\sigma_k^2 = n/4$ (e.g., Bevington 1969, p. 53).

Applying the binomial distribution to the $(\mu_k - \mu)$ from equation (18), we obtain

$$\mu = \langle\mu_k\rangle \qquad \text{and} \qquad \sigma_{\mu_k}^2 = \frac{n}{4}(2\gamma_n)^2 = \frac{1}{n}\left(\sum_i \gamma_i\right)^2 \neq \sum_i \sigma_i^2 = \sigma^2 \ . \tag{21}$$

The fact that $\sigma_{\mu_k}^2 \neq \sigma^2$ is not surprising: $\sigma_{\mu_k}^2$ is the variance of averages of individual outcomes and σ^2 is the variance of the individual outcomes. The two variances are comparable. To show this, we evaluate them replacing the individual values γ_i and σ_i (which for most distributions will be comparable) by their respective means $\langle\gamma_i\rangle$ and $\langle\sigma_i\rangle$. We find

$$\sigma_{\mu_k}^2 \sim n\langle\gamma_i\rangle^2 \sim n\langle\sigma_i\rangle^2 \sim \sigma^2 \tag{22}$$

But note $\sigma_{\mu_k}^2$ and σ^2 will never be equal except for very special cases. The easiest imagined of the those special cases is where all x_i have a common double-Dirac-function probability distribution with all the γ_i and σ_i equal.

The analysis so far shows that the outcomes are grouped in groups that are proportional to the binomial probabilities. But the μ_k are not adequate for predicting the average behavior with k. The group of the outcomes that yield a specific μ_k actually have a large spread along the x dimension.

A simple way to correct them to do so is to rescale the μ_k to what we will call the y_k which are given by

$$y_k = \left(k - \frac{n}{2}\right)(2\sigma_n) \ , \tag{23}$$

where have defined

$$\sigma_n^2 = \frac{\sigma^2}{n} \ . \tag{24}$$

The y_k are the rescaled mean outcomes for k right steps and $(n-k)$ left steps by the random variables x_i. With the rescaled y_k, we find

$$\sigma_{y_k}^2 = \frac{n}{4}(2\sigma_n)^2 = \sigma^2 \tag{25}$$

which is the correct variance for the individual outcomes. It is reasonable to conclude that the rescaling causes the y_k to yield the average results of the individual outcomes for all properties. It has to be said that the fact that we get right result for the CLT with the rescaling is the rigorous justification for the rescaling.

We now rewrite equation (23) to get

$$\frac{1}{k-(n/2)} = \frac{1}{f_k\sqrt{n}/2} = \frac{2\sigma_n}{y_k} \ , \tag{26}$$

where f_k is the number k standard deviations $\sqrt{n}/2$ by which k differs from its mean $n/2$. When $n \to \infty$ for fixed f_k and the corresponding y_k, σ_n becomes a infinitesimal provided it does not grow. In fact, to derive a Gaussian from our binomial distribution for the y_k (and hence for x), we will let $n \to \infty$, $\sigma_n \to 0$, and hold $\sqrt{n}\sigma_n = \sigma$ constant. Holding σ constant means we are shrinking the σ_i and the size scale of the x_i. This is just the alternative perspective to letting σ grow asymptotically linearly with $\sqrt{n}$ (as we specified in § 1), and thus holding the σ_i and the size scale of the x_i asymptotically constant. Our derivation follows that of Gotthilf Hagen (Hagen 1837) in a version reconstructed from memory from a long-ago book on experimental data analysis—which, memory notwithstanding, is not Squires (1968).

The main trick of the derivation is to find a differential equation for P_k in terms of y_k. We note that

$$\Delta y_k = 2\sigma_n \ , \qquad k = \frac{n + y_k/\sigma_n}{2} \ , \qquad P_{k+1} = \left(\frac{n-k}{k+1}\right) P_k \ , \tag{27}$$

and

$$\Delta P_k = \left(\frac{n-k}{k+1} - 1\right) P_k = \left(\frac{n-2k-1}{k+1}\right) P_k = \left[\frac{-y_k/\sigma_n - 1}{n/2 + y_k/(2\sigma_n) + 1}\right] P_k \ . \tag{28}$$

Now we write

$$\frac{\Delta P_k}{\Delta y_k} = \frac{P_k}{2\sigma_n}\left[\frac{-y_k/\sigma_n - 1}{n/2 + y_k/(2\sigma_n) + 1}\right] P_k \ . \tag{29}$$

And now we assume $n >> y_k/\sigma_n$ and $y_k/\sigma_n >> 1$, and keep only leading terms to get

$$\frac{\Delta P_k}{\Delta y_k} = -\frac{P_k y_k}{n\sigma_n^2} \ , \tag{30}$$

We now drop the k subscript, let $n \to \infty$ and $\sigma_n \to 0$ with $n\sigma_n^2 = \sigma$ staying constant, and take the limit as $\Delta y \to 0$. We obtain differential equation

$$\frac{dP}{dy} = -\frac{Py}{\sigma^2} \ , \tag{31}$$

The normalized solution of the differential equation with y replaced by $(x - \mu)$ is

$$P = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{32}$$

which actually completes the proof of the CLT: QED.

The above proof is non-rigorous in that we have treated limiting processes cavalierly and in that we buttressed our argument for the rescaling by saying it gives the right answer. But the proof gives insight into why the central limit theorem is true at a level suitable for astronomy and physics students.

6. CONCLUSION

Given its significance and just fame, it's welcome that the central limit theorem can be understood from simple considerations.

Support for this work has been provided the Department of Physics & Astronomy of the University of Nevada, Las Vegas and the Homer L. Dodge Department of Physics & Astronomy of the University of Oklahoma.

REFERENCES

Bevington, P. R. 1969, Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill Book Company)

Fischer, H. 2011, Central Limit Theorem: From Classical to Modern Probability Theory (New York: Springer Science+Business Media, LLC)

Griffiths, D. J. 2005, *Introduction to Quantum Mechanics* (Upper Saddle River, New Jersey: Pearson/Prentice Hall), (Gr)

Hagen, G. 1837, Grundzüge der Wahrscheinlichkeitsrechnung (Berlin: Dümmler)

Squires, G. L. 1968, Practical Physics (New York: McGraw-Hill Book Company)