A Probabilistic Proof of the Lindeberg-Feller Central Limit Theorem

Larry Goldstein

1. INTRODUCTION. The Central Limit Theorem, one of the most striking and useful results in probability and statistics, explains why the normal distribution appears in areas as diverse as gambling, measurement error, sampling, and statistical mechanics. In essence, the Central Limit Theorem states that the normal distribution applies whenever one is approximating probabilities for a quantity which is a sum of many independent contributions all of which are roughly the same size. It is the Lindeberg-Feller Theorem [**4**] which makes this statement precise in providing the sufficient, and in some sense necessary, Lindeberg condition whose satisfaction accounts for the ubiquitous appearance of the bell-shaped normal.

Generally the Lindeberg condition is handled using Fourier or analytic methods and is somewhat hard to interpret. Here we provide a simpler, equivalent, and more easily interpretable probabilistic formulation of the Lindeberg condition and demonstrate its sufficiency and partial necessity in the Central Limit Theorem using more elementary means.

The seeds of the Central Limit Theorem, or CLT, lie in the work of Abraham de Moivre, who, in 1733, not being able to secure himself an academic appointment, supported himself consulting on problems of probability and gambling. He approximated the limiting probabilities of the binomial distribution, the one which governs the behavior of the number S_n of successes in an experiment which consists of n independent trials, each one having the same probability $p \in (0, 1)$ of success.

Each individual trial of such an experiment can be modelled by X, a (Bernoulli) random variable which records one for each success and zero for each failure,

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p,$$

and has mean $EX = p$ and variance $\mathrm{Var}(X) = p(1 - p)$. The record of successes and failures in n independent trials is then given by an independent sequence $X_1, X_2, \ldots, X_n$ of these Bernoulli variables, and the total number of successes S_n by their sum

$$S_n = X_1 + \cdots + X_n. \tag{1}$$

Exactly, S_n has the binomial distribution, which specifies that

$$P(S_n = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

For even moderate values of n managing the binomial coefficients $\binom{n}{k}$ becomes unwieldy, to say nothing of computing the sum which yields the cumulative probability

$$P(S_n \le m) = \sum_{k \le m} \binom{n}{k} p^k (1 - p)^{n-k}$$

that there will be m or fewer successes.

The great utility of the CLT is in providing an easily computable approximation to such probabilities that can be quite accurate even for moderate values of n. To state this result, let Z denote a standard normal variable, that is, one with distribution function $P(Z \le x) = \Phi(x)$ given by

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)du \quad \text{where} \quad \varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right), \tag{2}$$

and recall that we say a sequence of random variables Y_n is said to converge in distribution to Y, written $Y_n \to_d Y$, if

$$\lim_{n\to\infty} P(Y_n \le x) = P(Y \le x) \quad \text{for all continuity points } x \text{ of } P(Y \le x). \tag{3}$$

Letting $W_n = (S_n - np)/\sqrt{np(1-p)}$, the binomial standardized to have mean zero and variance one, obtained from S_n by subtracting its mean and dividing by its standard deviation, the CLT yields

$$W_n \to_d Z.$$

This result allows us to approximate the cumbersome cumulative binomial probability $P(S_n \le m)$ by the simpler $\Phi((m - np)/\sqrt{np(1-p)})$, noting that $\Phi(x)$ is continuous for all x, and therefore that the convergence of distribution functions in (3) here holds everywhere.

It was only for the special case of the binomial that normal approximation was first considered. Only many years later with the work of Laplace around 1820 did it begin to be systematically realized that the same normal limit is obtained when the underlying Bernoulli variables are replaced by any variables with a finite variance. The result was the classical Central Limit Theorem, which states that W_n converges in distribution to Z whenever

$$W_n = (S_n - n\mu)/\sqrt{n\sigma^2}$$

is the standardization of a sum S_n, as in (1), of independent and identically distributed random variables each with mean μ and variance σ^2. From this generalization it now becomes somewhat clearer why various distributions observed in nature, which may not be at all related to the binomial, such as the errors of measurement averages, or the heights of individuals in a sample, take on the bell-shaped form: each observation is the result of summing many small independent factors.

A further extension of the classical CLT could yet come. In situations where the summands do not have identical distributions, can the normal curve still govern? For an example, consider the symmetric group $\mathcal{S}_n$, the set of all permutations π on the set $\{1, 2, \ldots, n\}$. We can represent $\pi \in \mathcal{S}_7$, for example, by two-line notation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 4 & 3 & 7 & 6 & 5 & 1 & 2 \end{pmatrix}$$

from which one can read that $\pi(1) = 4$ and $\pi(4) = 6$. This permutation can also be represented in the cycle notation

$$\pi = (1, 4, 6)(2, 3, 7)(5)$$

with the meaning that π maps 1 to 4, 4 to 6, 6 to 1, and so forth. From the cycle representation we see that π has two cycles of length 3 and one of length 1, for a total

of three cycles. In general, let $K_n(\pi)$ denote the total number of cycles in a permutation $\pi \in \mathcal{S}_n$. If π is chosen uniformly from all the $n!$ permutations in $\mathcal{S}_n$, does the Central Limit Theorem imply that $K_n(\pi)$ is approximately normally distributed for large n?

To answer this question we will employ the Feller coupling [**3**], which constructs a random permutation π uniformly from $\mathcal{S}_n$ with the help of n independent Bernoulli variables $X_1, \ldots, X_n$ with distributions

$$P(X_i = 0) = 1 - \frac{1}{i} \quad \text{and} \quad P(X_i = 1) = \frac{1}{i}, \quad i = 1, \ldots, n. \tag{4}$$

Begin the first cycle at stage 1 with the element 1. At stage i, $i = 1, \ldots, n$, if $X_{n-i+1} = 1$ close the current cycle and begin a new one starting with the smallest number not yet in any cycle, and otherwise choose an element uniformly from those yet unused and place it to the right of the last element in the current cycle. In this way at stage i we complete a cycle with probability $1/(n - i + 1)$, upon mapping the last element of the current cycle to the one which begins it.

As the total number $K_n(\pi)$ of cycles of π is exactly the number of times an element closes the loop upon completing its own cycle,

$$K_n(\pi) = X_1 + \cdots + X_n, \tag{5}$$

a sum of independent, but not identically distributed, random variables. Hence, despite the similarity of (5) to (1), the hypotheses of the classical Central Limit Theorem do not hold. Nevertheless, in 1922 Lindeberg [**8**] provided a general condition which can be applied in this case to show that $K_n(\pi)$ is asymptotically normal.

To explore Lindeberg's condition, first consider the proper standardization of $K_n(\pi)$ in our example. As any Bernoulli random variable with success probability p has mean p and variance $p(1 - p)$, the Bernoulli variable X_i in (4) has mean i^{-1} and variance $i^{-1}(1 - i^{-1})$ for $i = 1, \ldots, n$. Thus,

$$h_n = \sum_{i=1}^{n} \frac{1}{i} \quad \text{and} \quad \sigma_n^2 = \sum_{i=1}^{n} \left(\frac{1}{i} - \frac{1}{i^2} \right) \tag{6}$$

are the mean and variance of $K_n(\pi)$, respectively; the mean h_n is known as the nth harmonic number. In particular, standardizing $K_n(\pi)$ to have mean zero and variance 1 results in

$$W_n = \frac{K_n(\pi) - h_n}{\sigma_n},$$

which, by absorbing the scaling inside the sum, can be written as

$$W_n = \sum_{i=1}^{n} X_{i,n} \quad \text{where} \quad X_{i,n} = \frac{X_i - i^{-1}}{\sigma_n}. \tag{7}$$

In general, it is both more convenient and more encompassing to deal not with a sequence of variables but rather with a triangular array as in (7) which satisfies the following condition.

Condition 1.1. *For every $n = 1, 2, \ldots$, the random variables making up the collection $\mathbf{X}_n = \{X_{i,n} : 1 \le i \le n\}$ are independent with mean zero and finite variances*

$\sigma_{i,n}^2 = \mathrm{Var}(X_{i,n})$, *standardized so that*

$$W_n = \sum_{i=1}^{n} X_{i,n} \quad \textit{has variance} \quad \mathrm{Var}(W_n) = \sum_{i=1}^{n} \sigma_{i,n}^2 = 1.$$

Of course, even under Condition 1.1, some further assumptions must be satisfied by the summand variables for W_n to converge in distribution to the normal. For instance, if the first variable accounts for some nonvanishing fraction of the total variability, it will strongly influence the limiting distribution, possibly resulting in the failure of normal convergence. Ruling out such situations, the Lindeberg-Feller CLT, see [**4**], says that $W_n \to_d Z$ under the Lindeberg condition

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} L_{n,\epsilon} = 0 \quad \text{where} \quad L_{n,\epsilon} = \sum_{i=1}^{n} E\{X_{i,n}^2 \mathbf{1}(|X_{i,n}| \geq \epsilon)\}, \tag{8}$$

where for an event A, the 'indicator' random variable $\mathbf{1}(A)$ takes on the value 1 if A occurs, and the value 0 otherwise. The appearance of the Lindeberg condition is justified by explanations such as the one given by Feller [**4**], who roughly says that it requires the individual variances be due mainly to masses in an interval whose length is small in comparison to the overall variance. We present a probabilistic condition which is seemingly simpler, yet equivalent.

Our probabilistic approach to the CLT is through the so-called zero bias transformation introduced in [**6**], which was motivated by Stein's characterization [**10**] of $\mathcal{N}(0, \sigma^2)$, the mean zero normal distribution with variance σ^2. In particular, Stein proves that X has distribution $\mathcal{N}(0, \sigma^2)$ if and only if

$$\sigma^2 E[f'(X)] = E[Xf(X)] \tag{9}$$

for all absolutely continuous functions f for which these expectations exist.

To motivate the zero bias transformation, let B be a Bernoulli random variable with success probability $p \in (0, 1)$ and let X be the (nonnormal) centered Bernoulli variable $B - p$, having variance $\sigma^2 = p(1-p)$. Then, computing the right-hand side in Stein's identity (9), we find

$$\begin{aligned} E[Xf(X)] &= E[(B-p)f(B-p)] \\ &= p(1-p)f(1-p) - (1-p)pf(-p) \\ &= \sigma^2[f(1-p) - f(-p)] \\ &= \sigma^2 \int_{-p}^{1-p} f'(u)du \\ &= \sigma^2 Ef'(U), \end{aligned}$$

for U having uniform density over $[-p, 1-p]$. Hence, a version of Stein's identity holds for this nonnormal variable upon replacing the X appearing on the left-hand side of (9) by a variable with a distribution different from that of the given X.

This calculation gives one instance of the zero bias transformation: the centered Bernoulli $B - p$ is transformed to the uniform U over $[-p, 1-p]$. With $*$ indicating the transformation and $=_d$ the equality of two random variables in distribution, we write

$$(B-p)^* =_d U \quad \text{where } U \text{ has distribution } \mathcal{U}[-p, 1-p]. \tag{10}$$

In fact, it turns out [**6**] that for every X with mean zero and finite, nonzero variance σ^2, there exists a unique 'X-zero biased distribution' on an X^* which satisfies

$$\sigma^2 E f'(X^*) = E[Xf(X)] \tag{11}$$

for all absolutely continuous functions f for which these expectations exist. By Stein's characterization, X^* and X have the same distribution if and only if X has the $\mathcal{N}(0, \sigma^2)$ distribution, that is, the normal distribution is the zero bias transformation's unique fixed point. The Bernoulli example highlights the general fact that the distribution of X^* is always absolutely continuous, regardless of the nature of the distribution of X, and that the zero bias transformation moves a nonnormal distribution in some sense closer to normality.

One way to see the 'if' direction of Stein's characterization, that is, why the zero bias transformation fixes the normal, is to note that the density function $\varphi_{\sigma^2}(x) = \sigma^{-1}\varphi(\sigma^{-1}x)$ of a $\mathcal{N}(0, \sigma^2)$ variable, with $\varphi(x)$ given by (2), satisfies the differential equation with a form 'conjugate' to (11),

$$\sigma^2 \varphi'_{\sigma^2}(x) = -x\varphi_{\sigma^2}(x),$$

and now (11), with $X^* = X$, follows for a large class of functions f by integration by parts.

It is the uniqueness of the fixed point of the zero bias transformation, that is, the fact that X^* has the same distribution as X only when X is normal, that provides the probabilistic reason behind the CLT. This 'only if' direction of Stein's characterization suggests that a distribution which gets mapped to one nearby is close to being a fixed point of the zero bias transformation, and therefore must be close to the transformation's only fixed point, the normal. Hence the normal approximation should apply whenever the distribution of a random variable is close to that of its zero bias transformation.

Moreover, the zero bias transformation has a special property that immediately shows why the distribution of a sum W_n of comparably sized independent random variables is close to that of W_n^*: a sum of independent terms can be zero biased by randomly selecting a single summand, with each summand chosen with probability proportionally to its variance, and replacing it with one of comparable size. Thus, by differing only in a single summand, the variables W_n and W_n^* are close, making W_n an approximate fixed point of the zero bias transformation, and therefore approximately normal. This explanation, when given precisely, becomes a probabilistic proof of the Lindeberg-Feller CLT under a condition equivalent to (8) which we call the 'small zero bias condition'.

To set the stage for the introduction of the small zero bias condition we first consider more precisely this special property of the zero bias transformation on independent sums. Given $\mathbf{X}_n$ satisfying Condition 1.1, let $\mathbf{X}_n^* = \{X_{i,n}^* : 1 \le i \le n\}$ be a collection of random variables so that $X_{i,n}^*$ has the $X_{i,n}$ zero bias distribution and is independent of $\mathbf{X}_n$. Further, let I_n be a random index, independent of $\mathbf{X}_n$ and $\mathbf{X}_n^*$, with distribution

$$P(I_n = i) = \sigma_{i,n}^2, \tag{12}$$

and write the variables $X_{I_n,n}$ and $X_{I_n,n}^*$ which are selected by I_n, that is, the mixtures, using indicator functions as

$$X_{I_n,n} = \sum_{i=1}^{n} \mathbf{1}(I_n = i) X_{i,n} \quad \text{and} \quad X_{I_n,n}^* = \sum_{i=1}^{n} \mathbf{1}(I_n = i) X_{i,n}^*. \tag{13}$$

Then, upon replacing in the sum W_n the variable selected by I_n by the corresponding variable having its zero biased distribution, we obtain

$$W_n^* = W_n - X_{I_n,n} + X_{I_n,n}^*, \tag{14}$$

a variable which has the W_n zero bias distribution.

The proof of this fact is simple. For all absolutely continuous functions f for which the following expectations exist,

$$\begin{aligned} E[W_n f(W_n)] &= E\left[\sum_{i=1}^n X_{i,n} f(W_n)\right] \\ &= \sum_{i=1}^n E[X_{i,n} f((W_n - X_{i,n}) + X_{i,n})] \\ &= \sum_{i=1}^n E[\sigma_{i,n}^2 f'((W_n - X_{i,n}) + X_{i,n}^*)] \\ &= E\left[\sum_{i=1}^n \mathbf{1}(I_n = i) f'(W_n - X_{i,n} + X_{i,n}^*)\right] \\ &= E[f'(W_n - X_{I_n,n} + X_{I_n,n}^*)], \end{aligned}$$

where in the third equality we have used the fact that $X_{n,i}$ and $W_n - X_{i,n}$ are independent, and in the fourth equality that I_n is independent of $\mathbf{X}_n$ and $\mathbf{X}_n^*$. In view of (11), we now have the equality of the expectations of W_n^* and $W_n - X_{I_n,n} + X_{I_n,n}^*$ on a large enough class of functions sufficient to guarantee that these two random variables have the same distribution.

From (14) we see that the CLT should hold when the random variables $X_{I_n,n}$ and $X_{I_n,n}^*$ are both small asymptotically, since then the distribution of W_n is close to that of W_n^*, making W_n an approximate fixed point of the zero bias transformation. The following theorem shows that properly formalizing the notion of smallness results in a condition equivalent to Lindeberg's. Recall that we say a sequence of random variables Y_n converges in probability to Y, and write $Y_n \to_p Y$, if

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0 \quad \text{for all } \epsilon > 0.$$

Theorem 1.1. *If $\mathbf{X}_n$, $n = 1, 2, \ldots$ is a collection of random variables satisfying Condition* 1.1*, then the small zero bias condition*

$$X_{I_n,n}^* \to_p 0 \tag{15}$$

and the Lindeberg condition (8) *are equivalent.*

Our probabilistic proof of the Lindeberg-Feller Theorem develops by first showing that the small zero bias condition implies that

$$X_{I_n,n} \to_p 0,$$

and hence that

$$W_n^* - W_n = X_{I_n,n}^* - X_{I_n,n} \to_p 0.$$

Theorem 1.2 confirms that this convergence in probability to zero, mandating that W_n have its own zero bias distribution in the limit, is sufficient to guarantee that W_n converges in distribution to the normal.

Theorem 1.2. *If* $\mathbf{X}_n$, $n = 1, 2, \ldots$ *satisfies Condition* 1.1 *and the small zero bias condition* $X^*_{I_n,n} \to_p 0$, *then* $W_n \to_d Z$, *a standard normal random variable.*

We return now to the number $K_n(\pi)$ of cycles of a random permutation in $\mathcal{S}_n$, with mean h_n and variance σ_n^2 given by (6). Since $\sum_{i=1}^{\infty} 1/i^2 < \infty$, by upper and lower bounding the nth harmonic number h_n by integrals of $1/x$, we have

$$\lim_{n\to\infty} \frac{h_n}{\log n} = 1 \quad \text{and therefore} \quad \lim_{n\to\infty} \frac{\sigma_n^2}{\log n} = 1. \tag{16}$$

In view of (7) and (4) we note that in this case $W_n = \sum_{i=2}^{n} X_{i,n}$, as $X_1 = 1$ identically makes $X_{1,n} = 0$ for all n. Now by the linearity relation

$$(aX)^* =_d aX^* \quad \text{for all } a \neq 0,$$

which follows directly from (11), by (10) we have

$$X^*_{i,n} =_d U_i/\sigma_n, \text{ where } U_i \text{ has distribution } \mathcal{U}[-1/i, 1 - 1/i],\ i = 2, \ldots, n.$$

In particular, $|U_i| \leq 1$ with probability one for all $i = 2, \ldots, n$, and therefore

$$|X^*_{I_n,n}| \leq 1/\sigma_n \to 0 \tag{17}$$

by (16). Hence the small zero bias condition is satisfied, and Theorem 1.2 may be invoked to show that the number of cycles of a random permutation is asymptotically normal.

More generally, the small zero bias condition will hold for an array $\mathbf{X}_n$ with elements $X_{i,n} = X_i/\sigma_n$ whenever the independent mean zero summand variables $X_1, X_2, \ldots$ satisfy $|X_i| \leq C$ with probability one for some constant C, and the variance σ_n^2 of their sum S_n tends to infinity. In particular, from (11) one can verify that $|X_i| \leq C$ with probability one implies $|X_i^*| \leq C$ with probability one, and hence (17) holds with C replacing 1. In such a case, the Lindeberg condition (8) is also not difficult to verify: for any $\epsilon > 0$ one has $C/\sigma_n < \epsilon$ for all n sufficiently large, and all terms in the sum in (8) are identically zero.

Next consider the verification of the Lindeberg and small zero bias conditions in the identically distributed case, showing that the classical CLT is a special case of the Lindeberg-Feller theorem. Let $X_1, X_2, \ldots$ be independent with $X_i =_d X$, $i = 1, 2, \ldots$, where X is a random variable with mean μ and variance σ^2. By replacing X_i by $(X_i - \mu)/\sigma$, it suffices to consider the case where $\mu = 0$ and $\sigma^2 = 1$. Now set

$$X_{i,n} = \frac{1}{\sqrt{n}} X_i \quad \text{and} \quad W_n = \sum_{i=1}^{n} X_{i,n}.$$

For the verification of the classical Lindeberg condition, first use the fact that the distributions are identical and the $\sqrt{n}$-scaling to obtain

$$L_{n,\epsilon} = nE\{X_{1,n}^2 \mathbf{1}(|X_{1,n}| \geq \epsilon)\} = E\{X^2 \mathbf{1}(|X| \geq \sqrt{n}\epsilon)\}.$$

Now note that $X^2\mathbf{1}(|X| \geq \sqrt{n}\epsilon)$ tends to zero as $n \to \infty$, and is dominated by the integrable variable X^2; hence, the dominated convergence theorem may be invoked to provide the needed convergence of $L_{n,\epsilon}$ to zero.

Verification that the small zero bias condition is satisfied in this case is more mild. Again using that $(aX)^* = aX^*$, we have

$$X^*_{I_n,n} =_d \frac{1}{\sqrt{n}} X^*,$$

the mixture on the left being of these identical distributions. But now for any $\epsilon > 0$

$$\lim_{n\to\infty} P(|X^*_{I_n,n}| \geq \epsilon) = \lim_{n\to\infty} P(|X^*| \geq \sqrt{n}\epsilon) = 0,$$

that is, $X^*_{I_n,n} \to_p 0$.

Though Theorems 1.1 and 1.2 show that the Lindeberg condition is sufficient for normal convergence, it is easy to see, and well known, that it is not necessary. In particular, consider the case where for all n the first summand $X_{1,n}$ of W_n has the mean zero normal distribution σZ with variance $\sigma^2 \in (0, 1)$, and the Lindeberg condition is satisfied for the remaining variables, that is, that the limit is zero when taking the sum in (8) over all $i \neq 1$. Since the sum of independent normal variables is again normal, W_n will converge in distribution to Z, but (8) does not hold, since for all $\epsilon > 0$

$$\lim_{n\to\infty} L_{n,\epsilon} = E\{X^2_{1,n}\mathbf{1}(|X_{1,n}| \geq \epsilon)\} = \sigma^2 E\{Z^2\mathbf{1}(\sigma|Z| \geq \epsilon)\} > 0.$$

Defining

$$m_n = \max_{1\leq i\leq n} \sigma^2_{i,n} \tag{18}$$

to use for excluding such cases, we have the following partial converse to Theorem 1.2 (see also [**4**, Theorem 2, page 492] proved independently by Feller and Lévy, and also [**7**] for interesting CLT history).

Theorem 1.3. *If* $\mathbf{X}_n$, $n = 1, 2, \ldots$ *satisfies Condition* 1.1 *and*

$$\lim_{n\to\infty} m_n = 0, \tag{19}$$

then the small zero bias condition is necessary for $W_n \to_d Z$.

We prove Theorem 1.3 in Section 5 by showing that $W_n \to_d Z$ implies that $W^*_n \to_d Z$, and that (19) implies $X_{I_n,n} \to_p 0$. But then $W_n + X^*_{I_n,n} = W^*_n + X_{I_n,n} \to_d Z$ also, and we can then prove that

$$W_n \to_d Z \quad \text{and} \quad W_n + X^*_{I_n,n} \to_d Z \quad \text{imply that} \quad X^*_{I_n,n} \to_p 0.$$

This argument formalizes the probabilistic reason that the small zero bias condition, or Lindeberg condition, is necessary for normal convergence under (19).

Section 2 draws a parallel between the zero bias transformation and the one better known for size biasing, and there we consider its connection to the differential equation method of Stein using test functions. In Section 3 we prove the equivalence of the classical Lindeberg condition and the small zero bias condition, and then, in Sections 4 and 5, we prove its sufficiency and partial necessity for normal convergence.

Some pains have been taken to keep the treatment as elementary as possible, in particular by avoiding the use of characteristic functions. Though some technical argument is needed, only real functions are involved, and the development remains at a level as basic as the material permits. With the help of two general results in Section 6, the presentation is self-contained with the exception of the existence of the zero bias distribution in the generality already specified, and the following two results stated in Section 2: the bounds (24) to the Stein equation, applied in Theorem 1.2, and the well-known fact involving convergence in distribution, that (3) implies (21) when $\mathcal{C} = \mathcal{C}_b$, used in Theorem 1.3.

2. BIASING AND THE STEIN EQUATION. Relation (11) characterizing the zero bias distribution of a mean zero random variable with finite variance is quite similar to the better known identity which characterizes the size bias distribution of a nonnegative random variable X with finite mean μ. In particular we say that X^s has the X-size biased distribution if

$$\mu E f(X^s) = E[Xf(X)] \tag{20}$$

for all functions f for which these expectations exist. For example, if X is a nonnegative integer valued random variable, it is easy to verify that

$$P(X^s = k) = \frac{kP(X = k)}{\mu} \quad \text{for } k = 0, 1, \ldots,$$

that is, the probability that X^s takes on the value k is 'biased' by the size k of the value taken. Size biasing can appear unwanted in various sampling contexts, and is also implicated in generating the waiting time paradox ([**4**, sec. I.4]).

We note that the size biasing relation (20) is of the same form as the zero biasing relation (11), but with the mean μ replacing the variance σ^2, and f rather than f' evaluated on the biased variable. Hence zero biasing is a kind of analog of size biasing on the domain of mean zero random variables. In particular, the two transformations share the property that a sum of independent terms can be biased by replacing a single summand by one having that summand's biased distribution; in zero biasing the summand is selected with probability proportional to its variance, and in size biasing with probability proportional to its mean.

The biasing relations (11) and (20) are in terms of 'test functions', while definition (3) of convergence in distribution is expressed using distribution functions. The connection between these frameworks (see [**1**]) is the fact that convergence in distribution of Y_n to Y implies the convergence of expectations of functions of Y_n to those of Y, precisely, $Y_n \to_d Y$ implies

$$\lim_{n\to\infty} Eh(Y_n) = Eh(Y) \quad \text{for all } h \in \mathcal{C} \tag{21}$$

when $\mathcal{C} = C_b$, the collection of all bounded, continuous functions. Clearly, (21) holding with $\mathcal{C} = C_b$ implies that it holds with $\mathcal{C} = C_{c,0}^\infty$, the set of all functions with compact support which integrate to zero and have derivatives of all orders, since $C_{c,0}^\infty \subset C_b$. In Section 6 we prove the following result, making all three statements equivalent.

Theorem 2.1. *If* (21) *holds with* $\mathcal{C} = C_{c,0}^\infty$, *then* $Y_n \to_d Y$.

In light of the characterization (9), a strategy first suggested in [**9**] (see also [**11**]) for proving $W_n \to_d Z$ is to choose a class of test functions $\mathcal{C}$, such as $C_{c,0}^\infty$, which is

rich enough to guarantee convergence in distribution, and then for given $h \in \mathcal{C}$ to find a function f which solves the 'Stein equation'

$$f'(w) - wf(w) = h(w) - Eh(Z). \tag{22}$$

Now demonstrating $Eh(W_n) \to Eh(Z)$ can be accomplished by showing

$$\lim_{n\to\infty} E\left[f'(W_n) - W_n f(W_n)\right] = 0.$$

It is easy to verify that when $Eh(Z)$ exists an explicit solution to (22) is given by

$$f(w) = \varphi^{-1}(w) \int_{-\infty}^{w} [h(u) - Eh(Z)]\varphi(u)du, \tag{23}$$

where $\varphi(u)$ is the standard normal density given in (2). Stein [**11**] showed that when h is a bounded differentiable function with bounded derivative h', the solution f is twice differentiable and satisfies

$$||f'|| \le 2||h|| \quad \text{and} \quad ||f''|| \le 2||h'||, \tag{24}$$

where for any function g,

$$||g|| = \sup_{-\infty<x<\infty} |g(x)|.$$

These bounds will be applied in Section 4.

3. EQUIVALENCE: PROOF OF THEOREM 1.1. Since the random index I_n is independent of $\mathbf{X}_n$ and $\mathbf{X}_n^*$, taking expectations in (13) and using (12) yields

$$Ef(X_{I_n,n}) = \sum_{i=1}^{n} \sigma_{i,n}^2 Ef(X_{i,n}) \quad \text{and} \quad Ef(X^*_{I_n,n}) = \sum_{i=1}^{n} \sigma_{i,n}^2 Ef(X^*_{i,n}). \tag{25}$$

Let $\epsilon > 0$ be given and set

$$f(x) = \begin{cases} x + \epsilon & \text{if } x \le -\epsilon, \\ 0 & \text{if } -\epsilon < x < \epsilon, \\ x - \epsilon & \text{if } x \ge \epsilon. \end{cases}$$

The function f is absolutely continuous with

$$f'(x) = \mathbf{1}(|x| \ge \epsilon) \quad \text{almost everywhere.}$$

Hence, using the zero bias relation (11) for the second equality,

$$\begin{aligned} \sigma_{i,n}^2 P(|X^*_{i,n}| \ge \epsilon) &= \sigma_{i,n}^2 Ef'(X^*_{i,n}) \\ &= E\left[\left(X_{i,n}^2 - \epsilon|X_{i,n}|\right)\mathbf{1}(|X_{i,n}| \ge \epsilon)\right] \\ &\le E\left[\left(X_{i,n}^2 + \epsilon|X_{i,n}|\right)\mathbf{1}(|X_{i,n}| \ge \epsilon)\right] \\ &\le 2E\left(X_{i,n}^2 \mathbf{1}(|X_{i,n}| \ge \epsilon)\right), \end{aligned} \tag{26}$$

and summing over $i = 1, \ldots, n$ and applying identity (25) for the indicator function $\mathbf{1}(|x| \geq \epsilon)$ we obtain

$$P(|X^*_{I_n,n}| \geq \epsilon) = \sum_{i=1}^{n} \sigma^2_{i,n} P(|X^*_{i,n}| \geq \epsilon) \leq 2L_{n,\epsilon}.$$

Hence the Lindeberg condition (8) implies the small zero bias condition (15).

For the implication in the other direction use that for all x,

$$x^2 \mathbf{1}(|x| \geq \epsilon) \leq 2\left(x^2 - \frac{\epsilon}{2}|x|\right)\mathbf{1}\left(|x| \geq \frac{\epsilon}{2}\right).$$

Applying (26) and (25) to obtain the second and third equalities, respectively, we have

$$\begin{aligned} L_{n,\epsilon} = \sum_{i=1}^{n} E\{X^2_{i,n}\mathbf{1}(|X_{i,n}| \geq \epsilon)\} &\leq 2\sum_{i=1}^{n} E\left(X^2_{i,n} - \frac{\epsilon}{2}|X_{i,n}|\right)\mathbf{1}\left(|X_{i,n}| \geq \frac{\epsilon}{2}\right) \\ &= 2\sum_{i=1}^{n} \sigma^2_{i,n} P\left(|X^*_{i,n}| \geq \frac{\epsilon}{2}\right) \\ &= 2P\left(|X^*_{I_n,n}| \geq \frac{\epsilon}{2}\right), \end{aligned}$$

proving that the small zero bias condition (15) implies the Lindeberg condition (8). ∎

4. SUFFICIENCY: PROOF OF THEOREM 1.2. With the help of two lemmas we prove Theorem 1.2, the 'small zero bias' version of the Lindeberg-Feller theorem.

Lemma 4.1. *Let* $\mathbf{X}_n$, $n = 1, 2, \ldots$ *satisfy Condition* 1.1 *and* m_n *be given by* (18). *Then*

$$X_{I_n,n} \to_p 0 \quad \textit{whenever} \quad \lim_{n\to\infty} m_n = 0.$$

Proof. From (25) with $f(x) = x$ we find $EX_{I_n,n} = 0$, and hence $\mathrm{Var}(X_{I_n,n}) = EX^2_{I_n,n}$. Now (25) again, with $f(x) = x^2$, yields

$$\mathrm{Var}(X_{I_n,n}) = \sum_{i=1}^{n} \sigma^4_{i,n}.$$

Since for all i, $\sigma^4_{i,n} \leq \sigma^2_{i,n} \max_{1\leq j\leq n} \sigma^2_{j,n} = \sigma^2_{i,n} m_n$, for any $\epsilon > 0$ we have

$$P(|X_{I_n,n}| \geq \epsilon) \leq \frac{\mathrm{Var}(X_{I_n,n})}{\epsilon^2} = \frac{1}{\epsilon^2}\sum_{i=1}^{n} \sigma^4_{i,n} \leq \frac{1}{\epsilon^2} m_n \sum_{i=1}^{n} \sigma^2_{i,n} = \frac{1}{\epsilon^2} m_n,$$

the first inequality being Chebyshev's, and the last equality following from Condition 1.1. As $m_n \to 0$ by hypotheses, the proof is complete. ∎

Lemma 4.2. *If* $\mathbf{X}_n$, $n = 1, 2, \ldots$ *satisfies Condition* 1.1 *and the small zero bias condition, then*

$$X_{I_n,n} \to_p 0.$$

Proof. For all n, $1 \le i \le n$, and $\epsilon > 0$,

$$\sigma_{i,n}^2 = E(X_{i,n}^2 \mathbf{1}(|X_{i,n}| < \epsilon)) + E(X_{i,n}^2 \mathbf{1}(|X_{i,n}| \ge \epsilon)) \le \epsilon^2 + L_{n,\epsilon}.$$

Since the upper bound does not depend on i,

$$m_n \le \epsilon^2 + L_{n,\epsilon},$$

and now, since $\mathbf{X}_n$ satisfies the small zero bias condition, by Theorem 1.1 we have

$$\limsup_{n\to\infty} m_n \le \epsilon^2$$

and therefore

$$\lim_{n\to\infty} m_n = 0.$$

The claim now follows by Lemma 4.1. ∎

We are now ready to prove the forward direction of the Lindeberg-Feller CLT.

Proof of Theorem 1.2. Let $h \in C_{c,0}^\infty$ and let f be the solution to the Stein equation, for that h, given by (23). Substituting W_n for w in (22), taking expectation, and using (11) we obtain

$$E\left[h(W_n) - Eh(Z)\right] = E\left[f'(W_n) - W_n f(W_n)\right] = E\left[f'(W_n) - f'(W_n^*)\right] \tag{27}$$

with W_n^* given by (14). Since

$$W_n^* - W_n = X_{I_n,n}^* - X_{I_n,n},$$

the small zero bias condition and Lemma 4.2 imply

$$W_n^* - W_n \to_p 0. \tag{28}$$

By (24) f' is bounded with a bounded derivative f'', hence its global modulus of continuity

$$\eta(\delta) = \sup_{|y-x|\le\delta} |f'(y) - f'(x)|$$

is bounded and satisfies $\lim_{\delta\to 0} \eta(\delta) = 0$. Now, by (28),

$$\eta(|W_n^* - W_n|) \to_p 0, \tag{29}$$

and by (27) and the triangle inequality

$$\begin{aligned} |Eh(W_n) - Eh(Z)| &= \left|E(f'(W_n) - f'(W_n^*))\right| \\ &\le E\left|f'(W_n) - f'(W_n^*)\right| \\ &\le E\eta(|W_n - W_n^*|). \end{aligned}$$

Therefore

$$\lim_{n\to\infty} Eh(W_n) = Eh(Z)$$

by (29) and the bounded convergence theorem. Invoking Theorem 2.1 finishes the proof. ∎

5. PARTIAL NECESSITY. In this section we prove Theorem 1.3, showing in what sense the Lindeberg condition, in its equivalent zero bias form, is necessary. We begin with Slutsky's lemma, see [**5**], which states that

$$U_n \to_d U \text{ and } V_n \to_p 0 \quad \text{implies} \quad U_n + V_n \to_d U. \tag{30}$$

When independence holds, we have the following kind of reverse implication, whose proof is deferred to Section 6.

Lemma 5.1. *Let U_n and V_n, $n = 1, 2, \ldots$ be two sequences of random variables such that U_n and V_n are independent for every n. Then*

$$U_n \to_d U \text{ and } U_n + V_n \to_d U \quad \text{implies} \quad V_n \to_p 0.$$

Next, we show that the zero bias transformation enjoys the following continuity property.

Lemma 5.2. *Let Y and Y_n, $n = 1, 2, \ldots$ be mean zero random variables with finite, nonzero variances $\sigma^2 = \mathrm{Var}(Y)$ and $\sigma_n^2 = \mathrm{Var}(Y_n)$, respectively. If*

$$Y_n \to_d Y \quad \text{and} \quad \lim_{n\to\infty} \sigma_n^2 = \sigma^2,$$

then

$$Y_n^* \to_d Y^*.$$

Proof. Let $f \in C_{c,0}^\infty$ and $F(y) = \int_{-\infty}^{y} f(t)dt$. Since Y and Y_n have mean zero and finite variances, their zero bias distributions exist, so in particular,

$$\sigma_n^2 Ef(Y_n^*) = E\big[Y_n F(Y_n)\big] \quad \text{for all } n.$$

By (21), since $yF(y)$ is in C_b, we obtain

$$\sigma^2 \lim_{n\to\infty} Ef(Y_n^*) = \lim_{n\to\infty} \sigma_n^2 Ef(Y_n^*) = \lim_{n\to\infty} E[Y_n F(Y_n)] = E[YF(Y)] = \sigma^2 Ef(Y^*).$$

Hence $Ef(Y_n^*) \to Ef(Y^*)$ for all $f \in C_{c,0}^\infty$, so $Y_n^* \to_d Y^*$ by Theorem 2.1. ∎

We now provide the proof of the partial converse to the Lindeberg-Feller theorem.

Proof of Theorem 1.3. Since $W_n \to_d Z$ and $\mathrm{Var}(W_n) \to \mathrm{Var}(Z)$, the sequence of variances and the limit being identically one, Lemma 5.2 implies $W_n^* \to_d Z^*$. But Z is a fixed point of the zero bias transformation, hence $W_n^* \to_d Z$.

Since $m_n \to 0$, Lemma 4.1 yields that $X_{I_n,n} \to_p 0$, and Slutsky's lemma (30) now gives that

$$W_n + X_{I_n,n}^* = W_n^* + X_{I_n,n} \to_d Z.$$

Hence

$$W_n \to_d Z \quad \text{and} \quad W_n + X_{I_n,n}^* \to_d Z.$$

Since W_n is a function of $\mathbf{X}_n$, which is independent of I_n and $\mathbf{X}_n^*$ and therefore of $X_{I_n,n}^*$, invoking Lemma 5.1 yields $X_{I_n,n}^* \to_p 0$. ∎

6. APPENDIX. Here we provide the proofs that convergence of expectations over the class of functions $C^{\infty}_{c,0}$ implies convergence in distribution, and for the converse of Slutsky's lemma under an additional independence assumption.

Proof of Theorem 2.1. Let $a < b$ be continuity points of $P(Y \leq x)$. Billingsely [**1**] exhibits an infinitely differentiable function ψ taking values in [0, 1] such that $\psi(x) = 1$ for $x \leq 0$ and $\psi(x) = 0$ for $x \geq 1$. Hence, for all $u > 0$ the function

$$\psi_{a,b,u}(x) = \psi(u(x-b)) - \psi(u(x-a)+1)$$

is infinitely differentiable, has support in $[a - 1/u, b + 1/u]$, equals 1 for $x \in [a, b]$, and takes values in [0, 1] for all x. Furthermore,

$$\int_{-\infty}^{\infty} \psi_{a,b,u}(x)dx = \frac{1}{u} + (b-a),$$

so for every $\epsilon \in (0, 1]$, letting

$$d = -\frac{1}{u} + \epsilon^{-1}\left(\frac{1}{u} + (b-a)\right)$$

the function

$$\psi_{a,b,u,\epsilon}(x) = \psi_{a,b,u}(x) - \epsilon\psi_{b+2/u,b+2/u+d,u}(x)$$

is an element of $C^{\infty}_{c,0}$. Furthermore, for all $u > 0$ and $\epsilon \in (0, 1]$, $\psi_{a,b,u,\epsilon}(x)$ equals 1 on $[a, b]$, lies in [0, 1] for $x \in [a - 1/u, b + 1/u]$, and lies in $[-\epsilon, 0]$ for all other x. Hence

$$\begin{aligned}\limsup_{n\to\infty} P(Y_n \in (a,b]) &\leq \limsup_{n\to\infty} E\psi_{a,b,u,\epsilon}(Y_n) + \epsilon \\ &= E\psi_{a,b,u,\epsilon}(Y) + \epsilon \\ &\leq P\left(Y \in \left(a - \frac{1}{u}, b + \frac{1}{u}\right]\right) + \epsilon.\end{aligned}$$

Letting ϵ tend to zero and u to infinity, since a and b are continuity points,

$$\limsup_{n\to\infty} P(Y_n \in (a,b]) \leq P(Y \in (a,b]).$$

A similar argument using $\psi_{a+1/u,b-1/u,u,\epsilon}(x)$ shows that the reverse inequality holds with lim inf replacing lim sup, so for all continuity points a and b,

$$\lim_{n\to\infty} P(Y_n \in (a,b]) = P(Y \in (a,b]).$$

For b any continuity point and $\epsilon \in (0, 1]$, there exist continuity points $a < b < c$ with $P(Y \notin (a, c]) < \epsilon$. Since

$$P(Y \leq a) \leq P(Y \notin (a,c]) < \epsilon,$$

for all n sufficiently large $P(Y_n \leq a) \leq P(Y_n \notin (a, c]) \leq \epsilon$, and we have

$$|P(Y_n \leq b) - P(Y \leq b)| \leq |P(Y_n \in (a,b]) - P(Y \in (a,b])| + 2\epsilon,$$

yielding

$$\lim_{n\to\infty} P(Y_n \le b) = P(Y \le b). \qquad \blacksquare$$

Proof of Lemma 5.1. We first prove Lemma 5.1 for the special case where $U_n =_d U$ for all n, that is, we prove that if

$$U + V_n \to_d U \quad \text{with } U \text{ independent of } V_n \tag{31}$$

then $V_n \to_p 0$. By adding to U an absolutely continuous random variable A, independent of U and V_n, (31) holds with U replaced by the absolutely continuous variable $U + A$; we may therefore assume without loss of generality that U possesses a density function.

If V_n does not tend to zero in probability, there exist positive ϵ and p such that for infinitely many n

$$2p \le P(|V_n| \ge \epsilon) = P(V_n \ge \epsilon) + P(-V_n \ge \epsilon),$$

so either V_n or $-V_n$ is at least ϵ with probability at least p. Assume that there exists an infinite subsequence $\mathbf{K}$ such that

$$P(V_n \ge \epsilon) \ge p \quad \text{for all } n \in \mathbf{K}, \tag{32}$$

a similar argument holding in the opposite case.

Since U has a density the function $s(x) = P(x \le U \le x + 1)$ is continuous, and as the limits of $s(x)$ at plus and minus infinity are zero, $s(x)$ attains its maximum value, say s, in a bounded region. In particular,

$$y = \inf\{x : s(x) = s\}$$

is finite, and, by definition of y and the continuity of $s(x)$,

$$\sup_{x \le y-\epsilon} s(x) = r < s.$$

Since U and V_n are independent

$$P(y \le U + V_n \le y + 1 | V_n) = s(y - V_n) \quad \text{for all } n. \tag{33}$$

Therefore, on the one hand we have

$$P(y \le U + V_n \le y + 1 | V_n \ge \epsilon) \le r \quad \text{for all } n \in \mathbf{K},$$

but by conditioning on $V_n \ge \epsilon$ and its complement, using (33), (32), (31), and the fact that U is absolutely continuous, we obtain the contradiction

$$\begin{aligned} \liminf_{n\to\infty} P(y \le U + V_n \le y + 1) &\le rp + s(1 - p) \\ &< s = P(y \le U \le y + 1) \\ &= \lim_{n\to\infty} P(y \le U + V_n \le y + 1). \end{aligned}$$

To generalize to the situation where $U_n \to_d U$ through a sequence of distributions which may depend on n, we use Skorohod's construction (see Theorem 11.7.2 of [**2**]),

which implies that whenever $Y_n \to_d Y$, there exist $\overline{Y}_n$ and $\overline{Y}$ on the same space with $\overline{Y}_n =_d Y_n$ and $\overline{Y} =_d Y$ such that $\overline{Y}_n \to_p \overline{Y}$. In particular, $\overline{Y}_n$ and $\overline{Y}$ can be taken to be the inverse distribution functions of Y_n and Y, respectively, evaluated on the same uniform random variable. In this way we may construct $\overline{U}_n$ and $\overline{U}$ on the same (but now perhaps enlarged) space as V_n. Then, by the hypotheses and Slutsky's lemma (30), we obtain

$$\overline{U} + V_n = (\overline{U}_n + V_n) + (\overline{U} - \overline{U}_n) \to_d \overline{U} \quad \text{with } \overline{U} \text{ independent of } V_n.$$

Since this is the situation of (31), we conclude $V_n \to_p 0$. ∎

ACKNOWLEDGMENT. We enthusiastically thank the generous efforts of an anonymous reviewer, whose helpful comments, of both a technical and nontechnical nature, greatly improved this work in all respects.

REFERENCES

1. P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
2. R. Dudley, *Real Analysis and Probability*, Cambridge University Press, Cambridge, 1989.
3. W. Feller, The fundamental limit theorems in probability, *Bull. Amer. Math. Soc.* **51** (1945) 800–832.
4. ———, *An Introduction to Probability Theory and Its Applications*, vol. 2, Wiley, New York, 1967.
5. T. Ferguson, *A Course in Large Sample Theory*, Chapman & Hall, New York, 1996.
6. L. Goldstein and G. Reinert, Stein's method and the zero bias transformation with application to simple random sampling, *Ann. Appl. Probab.* **7** (1997) 935–952.
7. L. Le Cam, The central limit theorem around 1935, *Statist. Sci.* **1** (1986) 78–96.
8. J. Lindeberg, Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung, *Math. Z.* **15** (1922) 211–225.
9. C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 2, University of California Press, Berkeley, (1972) 583–602.
10. ———, Estimation of the mean of a multivariate normal distribution, *Ann. Statist.* **9** (1981) 1135–1151.
11. ———, *Approximate Computation of Expectations*, Institute of Mathematical Statistics, Hayward, CA, 1986.

LARRY GOLDSTEIN received his M.A. in Mathematics (1979), M.S. in Electrical Engineering (1982), and Ph.D. in Mathematics (1984), all from the University of California, San Diego. Since 1984 he has been in the mathematics department at the University of Southern California, where he is now full professor. His interests lie in probability and statistics, presently in Stein's method and cohort sampling schemes in epidemiology. He has applied the Central Limit Theorem while serving as a shipboard scientist on an exploration of the Mid-Atlantic Ridge, and as a consultant for VISA's Grinch sweepstakes to estimate the odds of winning and the average prize payout, which it did to surprising accuracy.
Department of Mathematics, KAP 108, University of Southern California, Los Angeles, CA 90089-2532
larry@math.usc.edu