The Laws of Large Numbers Compared

Tom Verhoeff

July 1993

1 Introduction

Probability Theory includes various theorems known as *Laws of Large Numbers*; for instance, see [Fel68, Hea71, Ros89]. Usually two major categories are distinguished: *Weak Laws* versus *Strong Laws*. Within these categories there are numerous subtle variants of differing generality. Also the *Central Limit Theorems* are often brought up in this context.

Many introductory probability texts treat this topic superficially, and more than once their vague formulations are misleading or plainly wrong. In this note, we consider a special case to clarify the relationship between the Weak and Strong Laws. The reason for doing so is that I have not been able to find a concise formal exposition all in one place. The material presented here is certainly not new and was gleaned from many sources.

In the following sections, $X_1, X_2, \ldots$ is a sequence of *independent* and *identically distributed* random variables with *finite* expectation μ. We define the associated sequence $\bar{X}_i$ of partial *sample means* by

$$\bar{X}_n \quad = \quad \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

The Laws of Large Numbers make statements about the convergence of $\bar{X}_n$ to μ. Both laws relate bounds on sample size, accuracy of approximation, and degree of confidence. The Weak Laws deal with limits of probabilities involving $\bar{X}_n$. The Strong Laws deal with probabilities involving limits of $\bar{X}_n$. Especially the mathematical underpinning of the Strong Laws requires a careful approach ([Hea71, Ch. 5] is an accessible presentation).

2 The Weak Law of Large Numbers

Let's not beat about the bush. Here is what the Weak Law says about convergence of $\bar{X}_n$ to μ.

2.1 Theorem (*Weak Law of Large Numbers*) We have

$$\forall_{\varepsilon>0} \lim_{n\to\infty} \Pr\left(|\bar{X}_n - \mu| \le \varepsilon\right) = 1 \ . \qquad (1)$$

This is often abbreviated to

$$\bar{X}_n \quad \overset{P}{\to} \quad \mu \qquad \text{as } n \to \infty$$

or in words: $\bar{X}_n$ converges *in probability* to μ as $n \to \infty$. ▟

On account of the definition of limit and the fact that probabilities are at most 1, Equation (1) can be rewritten as

$$\forall_{\varepsilon>0} \, \forall_{\delta>0} \, \exists_{N>0} \, \forall_{n \geq N} \, \Pr\big(|\bar{X}_n - \mu| \leq \varepsilon\big) \geq 1 - \delta \ . \tag{2}$$

The proof of the Weak Law is easy when the X_i's have a finite variance. It is most often based on Chebyshev's Inequality.

2.2 Theorem (*Chebyshev's Inequality*) Let X be a random variable with finite mean μ and finite variance σ^2. Then we have

$$\Pr\left(|X - \mu| \geq a\right) \quad \leq \quad \frac{\sigma^2}{a^2}$$

for all $a > 0$. ▟

A slightly different way of putting it is this: For all $a > 0$, we have

$$\Pr\left(|X - \mu| \geq a\sigma\right) \quad \leq \quad \frac{1}{a^2} \ .$$

Thus, the probability that X deviates from its expected value by at least k standard deviations is at most $1/k^2$. Chebyshev's Inequality is sharp when no further assumptions are made about X's distribution, but for practical applications it is often too sloppy. For example, the probability that X remains within 3σ of μ is at least $\frac{8}{9}$, no matter what distribution X has. However, when X is known to have a normal distribution, this probability in fact exceeds 0.9986.

We now prove the Weak Law when the variance is finite. Let σ^2 be the variance of each X_i. In that case, we have $E\,\bar{X}_n = \mu$ and $\operatorname{Var}\bar{X}_n = \sigma^2/n$. Let $\varepsilon > 0$. Substituting $X, \mu, \sigma, a := \bar{X}_n, \mu, \sigma/\sqrt{n}, \varepsilon$ in Chebyshev's Inequality then yields

$$\Pr\big(|\bar{X}_n - \mu| \geq \varepsilon\big) \quad \leq \quad \frac{\sigma^2}{n\varepsilon^2} \ . \tag{3}$$

Hence, for $\delta > 0$ and for all $n \geq \max\{1, \sigma^2/\delta\varepsilon^2\}$ we have

$$\Pr\big(|\bar{X}_n - \mu| < \varepsilon\big) \quad > \quad 1 - \delta$$

which completes the proof.

The Central Limit Theorem

Note that $\sigma = 0$ is uninteresting because in that case we have $\Pr(X_n = \mu) = 1$ (on account of Chebyshev's Inequality and continuity of probability for monotonic sequences of events).

In the case of finite non-zero variance, the Central Limit Theorem provides a much stronger result.

2.3 Theorem (*Central Limit Theorem*) If the X_i's have finite non-zero variance σ^2, then for all $a \le b$,

$$\lim_{n\to\infty} \Pr\left(a \le \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le b\right) \quad = \quad \Phi(b) - \Phi(a) \tag{4}$$

where Φ is the standard normal distribution defined by

$$\Phi(z) \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2} dx \ .$$

Convergence in (4) is uniform in a and b. ▪

The Central Limit Theorem can be interpreted as stating that for large n, the random variable $\bar{X}_n$ approximately has a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

We now prove that the Central Limit Theorem implies the Weak Law of Large Numbers when $0 < \sigma < \infty$. First observe that substituting $a, b := -c/\sigma, c/\sigma$ in the Central Limit Theorem yields

$$\lim_{n\to\infty} \Pr\left(|\bar{X}_n - \mu| \le \frac{c}{\sqrt{n}}\right) \quad = \quad \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) \ . \tag{5}$$

Let $\varepsilon > 0$ and $\delta > 0$. Take $c > 0$ such that $\Phi(-c/\sigma) \le \delta/3$ (this is possible since $\Phi(z) \to 0$ as $z \to -\infty$) and take N such that $c/\sqrt{N} \le \varepsilon$ and the limit in (5) is approached closer than $\delta/3$ for all $n \ge N$. We derive for $n \ge N$ (with hints placed between braces):

$\Pr\left(|\bar{X}_n - \mu| \le \varepsilon\right)$

$\ge$ { monotonicity of Pr, using $c/\sqrt{n} \le c/\sqrt{N} \le \varepsilon$, on account of definition of N }

$\Pr\left(|\bar{X}_n - \mu| \le c/\sqrt{n}\right)$

$\ge$ { definition of N }

$\Phi(c/\sigma) - \Phi(-c/\sigma) - \delta/3$

$=$ { $\Phi(z) + \Phi(-z) = 1$ }

$1 - 2\Phi(-c/\sigma) - \delta/3$

$\ge$ { definition of c }

$1 - \delta$

This concludes the proof.

If convergence to the standard normal distribution is assumed to be 'good' (much better than δ), then we can take bound N such that

$$\Phi\left(\frac{\varepsilon}{\sigma}\sqrt{N}\right) \quad \geq \quad 1 - \frac{\delta}{2} \ . \tag{6}$$

Compare this to the bound $N \geq \sigma^2/\delta\varepsilon^2$ on account of Chebyshev's Inequality. As an example, consider the case where we want to be 95% certain that the sample mean falls within $\frac{1}{4}\sigma$ of μ; that is, $\delta = 0.05$ and $\varepsilon = \sigma/4$. Chebyshev's Inequality yields $N \geq 16/0.05 = 320$ and the standard normal approximation yields $\sqrt{N}/4 \geq 1.96$ or $N \geq 61.47$. Thus, if the standard normal approximation is 'good' then our need is already fulfilled by the mean of 62 samples, instead of the 320 required by Chebyshev's Inequality.

I would like to emphasize the following points concerning the Central Limit Theorem.

- There exist estimates of how closely the standard normal distribution approximates the distribution of the sample mean. Consult [Fel71, Hea71] for the Berry–Esséen bound.
- If the X_i's themselves have a normal distribution, then so does the sample mean and the 'approximation' in the Central Limit Theorem is in fact exact.
- The more general versions of the Weak Law are not derivable from (more general versions of) the Central Limit Theorem.

3 The Strong Law of Large Numbers

Let's start again with the theorem.

3.1 Theorem (*Strong Law of Large Numbers*) We have

$$\Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) \quad = \quad 1 \ . \tag{7}$$

This is often abbreviated to

$$\bar{X}_n \quad \overset{\text{a.s.}}{\to} \quad \mu \qquad \text{as } n \to \infty$$

or in words: $\bar{X}_n$ converges *almost surely* to μ as $n \to \infty$. ▟

One of the problems with such a law is the assignment of probabilities to statements involving infinitely many random variables. For that purpose, one needs a careful introduction of notions like *sample space*, *probability measure*, and *random variable*. See for instance [Tuc67, Hea71, Chu74a, LR79].

Using some Probability Theory, the Strong Law can be rewritten into a form with probabilities involving finitely many random variables only. We rewrite Equation (7) in a chain of equivalences:

$$
\begin{array}{ll}
 & \Pr\left(\lim\limits_{n\to\infty} \bar{X}_n = \mu\right) = 1 \\
\Leftrightarrow & \qquad \{ \text{ definition of limit } \} \\
 & \Pr\left(\forall_{\varepsilon>0}\, \exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\right) = 1 \qquad (8) \\
\Leftrightarrow & \qquad \{ \text{ Note 1 below } \} \\
 & \forall_{\varepsilon>0}\, \Pr\left(\exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\right) = 1 \qquad (9) \\
\Leftrightarrow & \qquad \{ \text{ Note 2 below } \} \\
 & \forall_{\varepsilon>0}\, \forall_{\delta>0}\, \exists_{N>0}\, \Pr\left(\forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\right) \geq 1 - \delta \qquad (10) \\
\Leftrightarrow & \qquad \{ \text{ Note 3 below } \} \\
 & \forall_{\varepsilon>0}\, \forall_{\delta>0}\, \exists_{N>0}\, \forall_{r\geq 0}\, \Pr\left(\forall_{N\leq n\leq N+r}\, |\bar{X}_n - \mu| \leq \varepsilon\right) \geq 1 - \delta \qquad (11)
\end{array}
$$

Comparing Equations (2) and (10) we immediately infer the Weak Law from the Strong Law, which explains their names.

In order to supply the notes to above derivation, let $(\Omega, \mathcal{F}, P)$ be an appropriate probability space for the random variables X_i, and define events A_ε, B_N, and C_r for $\varepsilon > 0$, $N > 0$, and $r \geq 0$ by

$$
\begin{array}{rcl}
A_\varepsilon & = & \{\omega \in \Omega \mid \exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n(\omega) - \mu| \leq \varepsilon\} \\
B_N & = & \{\omega \in \Omega \mid \forall_{n\geq N} |\bar{X}_n(\omega) - \mu| \leq \varepsilon\} \\
C_r & = & \{\omega \in \Omega \mid \forall_{N\leq n\leq N+r}\, |\bar{X}_n(\omega) - \mu| \leq \varepsilon\}\ .
\end{array}
$$

These events satisfy the following monotonicity properties:

$$
\begin{array}{rcll}
A_\varepsilon & \supseteq & A_{\varepsilon'} & \text{for } \varepsilon \geq \varepsilon' \\
B_N & \subseteq & B_{N+1} & \\
C_r & \supseteq & C_{r+1}\ . &
\end{array}
$$

Therefore, on account of the continuity of probability measure P for monotonic chains of events, we have

$$
\begin{array}{rcll}
P(\bigcap_{m=1}^{\infty} A_{1/m}) & = & \lim\limits_{m\to\infty} P(A_{1/m}) & (12) \\
P(\bigcup_{N=1}^{\infty} B_N) & = & \lim\limits_{N\to\infty} P(B_N) & (13) \\
P(\bigcap_{r=0}^{\infty} C_r) & = & \lim\limits_{r\to\infty} P(C_r)\ . & (14)
\end{array}
$$

Note 1. We derive

$$
\begin{array}{ll}
 & \Pr\left(\forall_{\varepsilon>0}\, \exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\right) = 1 \\
\Leftrightarrow & \qquad \{ \text{ definitions of Pr and } A_\varepsilon \ \} \\
 & P(\bigcap_{\varepsilon>0} A_\varepsilon) = 1 \\
\Leftrightarrow & \qquad \{ \text{ monotonicity of } A_\varepsilon, \text{ using } 1/m \to 0 \text{ as } m \to \infty \ \} \\
 & P(\bigcap_{m=1}^{\infty} A_{1/m}) = 1 \\
\Leftrightarrow & \qquad \{ \text{ (12) } \}
\end{array}
$$

$$\lim_{m\to\infty} P(A_{1/m}) = 1$$

$\Leftrightarrow$ { property of limits, using that $P(A_{1/m})$ is descending and at most 1 }

$$\forall_{m>0}\, P(A_{1/m}) = 1$$

$\Leftrightarrow$ { see first two steps, also using monotonicity of P }

$$\forall_{\varepsilon>0}\, \Pr\big(\exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\big) = 1$$

Note 2. We derive for $\varepsilon > 0$

$$\Pr\big(\exists_{N>0}\, \forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\big) = 1$$

$\Leftrightarrow$ { definitions of Pr and B_N, and set theory }

$$P(\bigcup_{N=1}^{\infty} B_N) = 1$$

$\Leftrightarrow$ { (13) }

$$\lim_{N\to\infty} P(B_N) = 1$$

$\Leftrightarrow$ { definition of limit, using $P(B_k) \leq 1$ }

$$\forall_{\delta>0}\, \exists_{N>0}\, \forall_{k\geq N}\, P(B_k) \geq 1 - \delta$$

$\Leftrightarrow$ { monotonicity of P, using $B_k \supseteq B_N$ for $k \geq N$ }

$$\forall_{\delta>0}\, \exists_{N>0}\, P(B_N) \geq 1 - \delta$$

$\Leftrightarrow$ { definitions of Pr and B_N }

$$\forall_{\delta>0}\, \exists_{N>0}\, \Pr\big(\forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\big) \geq 1 - \delta$$

Note 3. We derive for $\varepsilon > 0$, $\delta > 0$, and $N > 0$

$$\Pr\big(\forall_{n\geq N}\, |\bar{X}_n - \mu| \leq \varepsilon\big) \geq 1 - \delta$$

$\Leftrightarrow$ { definitions of Pr and C_r, and set theory }

$$P(\bigcap_{r\geq 0} C_r) \geq 1 - \delta$$

$\Leftrightarrow$ { (14) }

$$\lim_{r\to\infty} P(C_r) \geq 1 - \delta$$

$\Leftrightarrow$ { property of limits, using that $P(C_r)$ is descending }

$$\forall_{r\geq 0}\, P(C_r) \geq 1 - \delta$$

$\Leftrightarrow$ { definitions of Pr and C_r }

$$\forall_{r\geq 0}\, \Pr\big(\forall_{N\leq n\leq N+r}\, |\bar{X}_n - \mu| \leq \varepsilon\big) \geq 1 - \delta$$

Quotes from the literature

I have not been able to find a reference that explicitly presents the preceding chain of equivalent expressions for the Strong Law ([Chu74a, Ch. 4] comes close). Many authors take one of these expression as definition. Below are some typical quotes to illustrate the state of affairs. Note that each of these quotes contains a partly verbal expression, which in some cases is even ambiguous as to the order of the quantifiers.

The paraphrasing of the Strong Law in [Ros89, p. 351] resembles Equation (8), though it is also possible to read it as Equation (9):

> "In particular, [the Strong Law] shows that, with probability 1, for any positive value ε,
>
> $$\left|\sum_{i=1}^{n} \frac{X_i}{n} - \mu\right|$$
>
> will be greater than ε only a finite number of times."

Equation (10) can be recognized in [Hea71, p. 226]:

> "Indeed for arbitrarily small $\varepsilon > 0$, $\delta > 0$, and large $N = N(\varepsilon, \delta)$, $\ldots$ the [definition] of $X_n \overset{\text{a.s.}}{\to} X \ldots$ can be restated $\ldots$ as
>
> $$P\left(\bigcap_{n=N}^{\infty} \{\omega \mid |X_n(\omega) - X(\omega)| < \varepsilon\}\right) > 1 - \delta$$
>
> $\ldots$"

Equation (11) resembles the definition in [Fel68, p. 259]:

> "We say that the sequence X_k obeys the strong law of large numbers if to every pair $\epsilon > 0$, $\delta > 0$, there corresponds an N such that there is probability $1 - \delta$ or better that for every $r > 0$ all $r + 1$ inequalities
>
> $$\frac{|\mathbf{S}_n - m_n|}{n} < \epsilon, \qquad n = N, N+1, \ldots, N+r$$
>
> will be satisfied."

Proofs of the Strong Law

Again the case with finite variance is easier than the general case. Most of the proofs that I have encountered for the Strong Law assuming finite variance are based on Kolmogorov's Inequality, which is a generalization of Chebyshev's Inequality. Even in that case there are still some technical hurdles (I will not go into these). Consequently, the proofs do not give rise to an explicit bound N in (11) in terms of ε and δ. An exception is [Eis69], where the Hajek–Rényi Inequality is used, which is a generalization of Kolmogorov's Inequality.

There is, however, a nice overview article, namely [HR80], that specifically looks at bounds N in terms of ε and δ. It shows, among other things, that

$$\Pr\left(\exists_{n \geq N}\, |\bar{X}_n - \mu| \geq \varepsilon\right) \quad \leq \quad \frac{\sigma^2}{N\varepsilon^2}\ . \tag{15}$$

Compare this result to (3): it is the same upper bound but for a much larger event. This creates the impression that the Weak Law is not that much weaker than the Strong Law.

4 Concluding Remarks

We have looked at one special case to clarify the relationship between the Weak and the Strong Law of Large Numbers. The case was special in that we have assumed X_i to be a sequence of independent and identically distributed random variables with finite expectation and that we have considered the convergence of partial sample means to the common expectation. Historically, it was preceded by more special cases, for instance, with the X_i restricted to Bernoulli variables. Nowadays these laws are treated in much more generality.

My main interest was focused on Equations (8) through (11), which are equivalent—but subtly different—formulations of the Strong Law of Large Numbers. Furthermore, I have looked at "constructive" bounds related to the rate of convergence.

Let me conclude with some sobering quotes (also to be found in [Chu74b, p. 233]). Feller writes in [Fel68, p. 152]:

> "[The weak law of large numbers] is of very limited interest and should be replaced by the more precise and more useful strong law of large numbers."

In [Wae71, p. 98], van der Waerden writes:

> "[The strong law of large numbers] scarcely plays a role in mathematical statistics."

Acknowledgment I would like to thank Fred Steutel for discussing the issues in this paper and for pointing out reference [HR80].

References

[Chu74a] Kai Lai Chung. *A Course in Probability Theory*. Academic Press, second edition, 1974.

[Chu74b] Kai Lai Chung. *Elementary Probability Theory and Stochastic Processes*. Springer, 1974.

[Eis69] Marin Eisen. *Introduction to Mathematical Probability Theory*. Prentice-Hall, 1969.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume I. Wiley, third edition, 1968.

[Fel71] William Feller. *An Introduction to Probability Theory and Its Applications*, volume II. Wiley, second edition, 1971.

[Hea71] C. R. Heathcote. *Probability: Elements of the Mathematical Theory*. Unwin, 1971.

[HR80] A. H. Hoekstra and J. T. Runnenburg. Inequalities compared. *Statistica Neerlandica*, 34(2):67–82, 1980.

[LR79] R. G. Laha and V. K. Rohatgi. *Probability Theory*. Wiley, 1979.

[Ros89] Sheldon Ross. *A First Course in Probability*. Macmillan, third edition, 1989.

[Tuc67] Howard G. Tucker. *A Graduate Course in Probability*. Academic Press, 1967.

[Wae71] B. L. Waerden, van der. *Mathematische Statistik*. Springer, third edition, 1971.