Ma 3/103
Introduction to Probability and Statistics

KC Border
Winter 2021

Lecture 7: Concentration Inequalities and the Law of Averages

Relevant textbook passages:

Pitman [19]: Section 3.3

Larsen–Marx [17]: Section 4.3

7.1 The Law of Averages

The “Law of Averages” is an informal term used to describe a number of mathematical theorems that relate averages of sample values to expectations of random variables.

Given random variables $X_1, \dots, X_n$ on a probability space $(\Omega, \mathcal{F}, P)$, each point $\omega \in \Omega$ yields a list $X(\omega)_1, \dots, X(\omega)_n$. If we average these numbers we get $\bar{X}(\omega) = \sum_{i=1}^n X_i(\omega)/n$, the **sample average** associated with the outcome ω. The sample average, since it depends on ω, is also a random variable.

In later lectures, I’ll talk about how to determine the distribution of a sample average, but we already have a case that we can deal with. If $X_1, \dots, X_n$ are independent Bernoulli random variables, their sum has a Binomial distribution, so the distribution of the sample average is easily given. First note that the sample average can only take on the values k/n, for $k = 0, \dots, n$, and that

$$P\big(\bar{X} = k/n\big) = \binom{n}{k} p^k (1-p)^{n-k}.$$

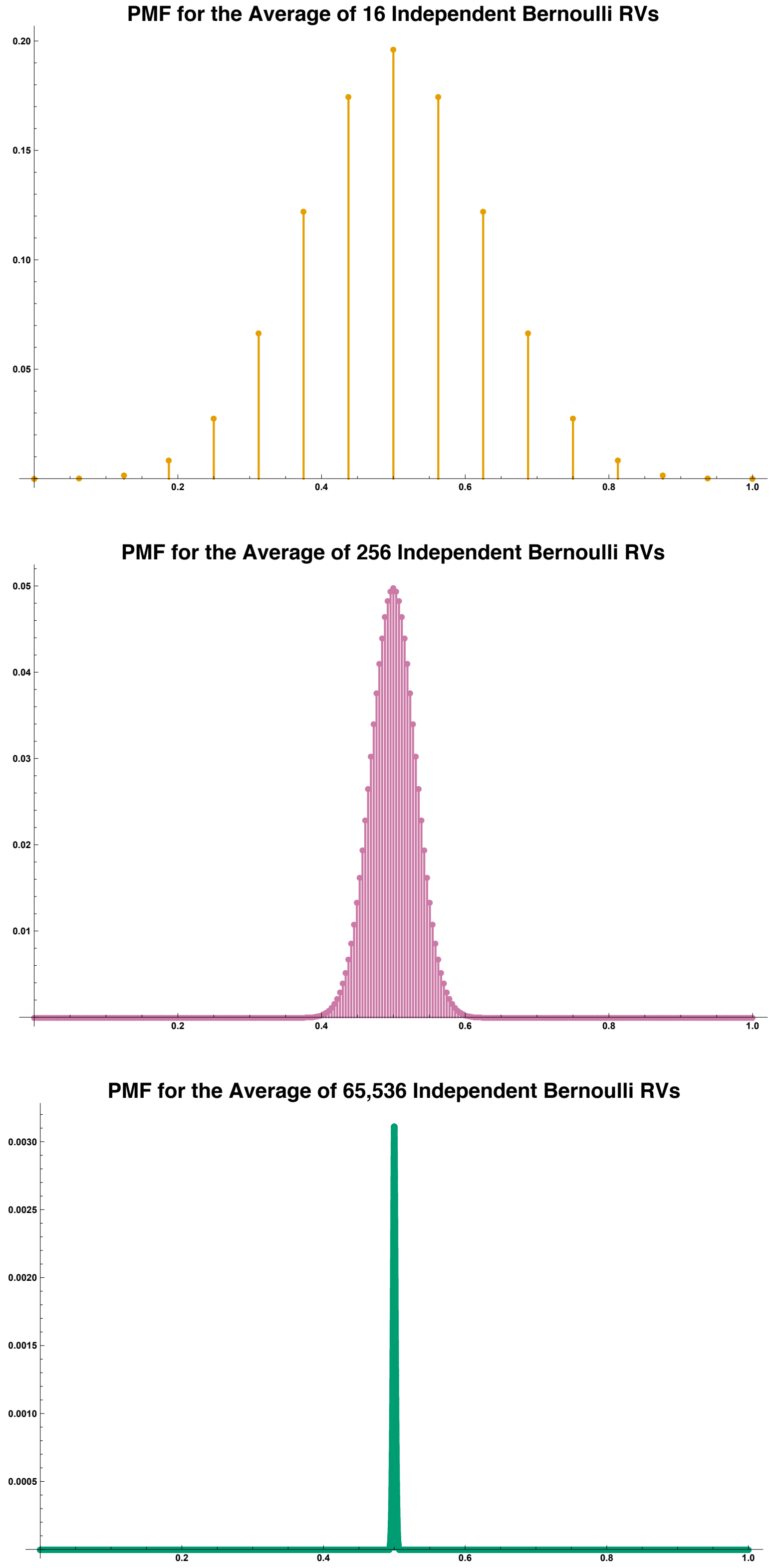
Figure 7.1 shows the probability mass function of $\bar{X}$ for the case $p = 1/2$ with various values of n. Observe the following things about the graphs.

- The sample average $\bar{X}$ is always between 0 and 1, and it is simply the fraction of successes in sequence of trials.

- If the frequency interpretation of probability is to make sense, then as the sample size grows, it should converge to the probability of success, which in this case is 1/2.

- What can we conclude about the probability that $\bar{X}$ is near 1/2? As the sample size becomes larger, the heights (which measure probability) of the dots shrink, but there are more and more of them close to 1/2. Which effect wins?

What happens for other kinds of random variables? Fortunately we do not need to know the details of the distribution to prove a Law of Averages. But we start with some preliminaries.



PMF for the Average of 16 Independent Bernoulli RVs

PMF for the Average of 256 Independent Bernoulli RVs

PMF for the Average of 65,536 Independent Bernoulli RVs

Figure 7.1.

7.2 Standardized random variables

Pitman [19]: p. 190

7.2.1 Definition *Given a random variable X with finite mean μ and finite variance σ^2, the* ***standardization*** *of X is the random variable X^* defined by*

$$X^* = \frac{X-\mu}{\sigma}.$$

Note that

$$\boldsymbol{E}\,X^* = 0, \quad \text{and} \quad \boldsymbol{Var}\,X^* = 1,$$

and

$$X = \sigma X^* + \mu,$$

so that X^* is just X measured in different units, called ***standard units****.*

[Note: Pitman uses both X^* and later X_* to denote the standardization of X.]

A convenient feature of standardized random variables is that they are invariant under change of scale and location.

7.2.2 Proposition *Let X be a random variable with mean μ and standard deviation σ, and let $Y = aX + b$, where $a > 0$. Then*

$$X^* = Y^*.$$

Proof: The proof follows from Propositions 6.7.1 and 6.10.2, which assert that $\boldsymbol{E}\,Y = a\mu + b$ and $\operatorname{SD} Y = a\sigma$. So

$$Y^* = \frac{Y - a\mu - b}{a\sigma} = \frac{\overbrace{aX+b}^{=Y} - a\mu - b}{a\sigma} = \frac{a(X-\mu)}{a\sigma} = \frac{X-\mu}{\sigma} = X^*.$$

∎

7.3 Concentration inequalities

The term **concentration inequality** refers to a proposition about the probability of the value of a random variable being concentrated near a particular point. Concentration inequalities are crucial to the proof of the Law of Large Numbers and have many applications in statistics. One could write a whole book on the subject. Indeed Boucheron, Lugosi, and Massart [2] have done so.

7.3.1 Markov

Markov's Inequality bounds the probability of large values of a nonnegative random variable in terms of its expectation. It is a very crude bound, but it is just what we need for the Weak Law of Large Numbers.

Pitman [19]: p. 174

7.3.1 Proposition (Markov's Inequality) *Let X be a nonnegative random variable with finite mean μ. For every $\varepsilon > 0$,*

$$P(X \geqslant \varepsilon) \leqslant \frac{\mu}{\varepsilon}.$$

Proof: For any point $\omega \in \Omega$ and any event E, since $X \geqslant 0$, we have

$$X(\omega) \geqslant X(\omega)\mathbf{1}_E(\omega).$$

Now let

$$E = (X \geqslant \varepsilon).$$

Then

$$X \geqslant X\mathbf{1}_E \geqslant \varepsilon \mathbf{1}_E.$$

Use the fact that **expectation is a positive linear operator** to get

$$\begin{aligned} \boldsymbol{E}\, X &\geqslant \boldsymbol{E}(X\mathbf{1}_E) \\ &\geqslant \boldsymbol{E}(\varepsilon\mathbf{1}_E) \\ &= \varepsilon\, \boldsymbol{E}(\mathbf{1}_E) \\ &= \varepsilon P(E) \\ &= \varepsilon P(X \geqslant \varepsilon). \end{aligned}$$

Divide by $\varepsilon > 0$ to get $P(X \geqslant \varepsilon) \leqslant \mu/\varepsilon$. ∎

Note that if ε is small enough, so that $\mu/\varepsilon > 1$, then Markov's Inequality is uninformative, since probabilities are always $\leqslant 1$.

7.3.2 Bienaymé–Chebyshev

Double check the references in this section.

Pitman [19]: p. 191

The following result is known as Chebychev's Inequality, after Pafnuty Chebyshev [5] (Пафну́тий Чебышёв), see, e.g., [13, p. 94].[1]

7.3.2 Proposition (Chebyshev Inequality, version 1) *Let X be a random variable with finite mean μ and variance σ^2 (and standard deviation σ). For every $\varepsilon > 0$,*

$$P(\,|X-\mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{\varepsilon^2}.$$

Proof:

$$\begin{aligned} P(\,|X-\mu| \geqslant \varepsilon) &= P\big((X-\boldsymbol{E}\,X)^2 \geqslant \varepsilon^2\big) && \text{square both sides} \\ &\leqslant \frac{\boldsymbol{E}(X-\boldsymbol{E}\,X)^2}{\varepsilon^2} && \text{Markov's Inequality} \\ &= \frac{\sigma^2}{\varepsilon^2} && \text{definition of variance,} \end{aligned}$$

where the inequality follows from Markov's Inequality applied to the random variable $(X-\boldsymbol{E}\,X)^2$ and the constant ε^2. ∎

Pitman [19]: p. 191

In fact, Chebyshev [5] proved this slightly stronger result.

7.3.3 Proposition (Chebyshev Inequality, version 2) *Let X_i have expectation μ_i and variance σ_i^2, $i = 1, \dots, n$. Then for every $\varepsilon > 0$,*

$$P\Big(\,|\textstyle\sum_i X_i - \mu_i| > \varepsilon\sqrt{\sum_i \sigma_i^2}\Big) < 1/\varepsilon^2.$$

By combining the Bienaymé Equality 6.11.1 with Chebyshev's Inequality, we get the following result on sums of *independent* random variables. It is now generally known as the Bienaymé–Chebyshev Inequality, see, e.g., Loève [18, p. 246].

Pitman [19]: p. 191

7.3.4 Proposition (Bienaymé–Chebyshev Inequality, version 1) *Let X_i be independent random variables with expectation μ_i and variance σ_i^2, $i = 1, \dots, n$. Then for every $\varepsilon > 0$,*

$$P\Big(|\textstyle\sum_i X_i - \mu_i| > \varepsilon\sqrt{\sum_i \sigma_i^2}\Big) < 1/\varepsilon^2.$$

We may rewrite the Chebyshev Inequality in terms of standard units as:

7.3.5 Proposition (Chebyshev Inequality, version 3) *Let X be a random variable with finite mean μ and variance σ^2 (and standard deviation σ), and standardization X^*. For every $k > 0$,*

Pitman [19]: p. 191

$$P(\,|X^*| \geqslant k) \leqslant \frac{1}{k^2}.$$

For many distributions, the Bienaymé–Chebyshev bound is very generous. The next set of inequalities tightens the bound considerably for many distributions.

7.3.3 Hoeffding

Hoeffding's Inequality [14] in its second form is a favorite among the computer scientists I know. These versions are taken from Wasserman [30, Theorems 4.4, 4.5, p. 64–65].

We start with the following lemma on moment generating functions (see Section 8.7) for bounded random variables. The following proof is taken from Wasserman [30, Appendix 4.4, p. 67].

7.3.6 Lemma (Hoeffding's Lemma) *Let X be a random variable satisfying $\boldsymbol{E}\,X = 0$, $a \leqslant X \leqslant b$ $(a < 0 < b)$. Then the moment generating function of X satisfies for $t > 0$,*

$$\boldsymbol{E}\,e^{tX} \leqslant e^{t^2(b-a)^2/8}.$$

Proof: Since

$$a \leqslant X \leqslant b,$$

each value of X is a (random) convex combination a and b. Indeed

$$X = (1-\lambda)a + \lambda b, \quad \text{where } \lambda \text{ is the random variable } \frac{X-a}{b-a}.$$

Now the exponential function is convex, so for each value of X,

$$e^{tX} \leqslant (1-\lambda)e^{ta} + \lambda e^{tb}.$$

Since **expectation is a positive linear operator**, taking expectations of both sides gives:

$$\boldsymbol{E}\,e^{tX} \leqslant \boldsymbol{E}(1-\lambda)e^{ta} + \lambda e^{tb} = e^{ta}(1 - \boldsymbol{E}\,\lambda) + e^{tb}\,\boldsymbol{E}\,\lambda. \tag{1}$$

Remember, it is λ that is random, in fact,

$$\lambda = \frac{X-a}{b-a} \text{ so } \boldsymbol{E}\,\lambda = \frac{\boldsymbol{E}\,X - a}{b-a}.$$

[1] Chebyshev seems to be the currently preferred transliteration. The published result [5] lists the author as de Tchébychef. Loève [18] refers to hims as Tchebichev. See wp for more variations.

By hypothesis, $\boldsymbol{E}\,X = 0$, so $\boldsymbol{E}\,\lambda = -a/(b-a)$, also $1 - \boldsymbol{E}\,\lambda = b/(b-a)$. Substituting this into (1) gives

$$\boldsymbol{E}\,e^{tX} \leqslant \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}. \tag{2}$$

It is not immediately obvious, but we can write the right-hand side of (2) as a function of

$$u = t(b-a)$$

alone. Set

$$\gamma = \frac{-a}{b-a}, \text{ so } 1-\gamma = \frac{b}{b-a} \text{ and } ta = -\gamma u.$$

Then, factoring out $e^{ta} = e^{-\gamma u}$ gives

$$\begin{aligned} \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} &= (1-\gamma)e^{ta} + \gamma e^{tb} \\ &= e^{ta}\big(1-\gamma+\gamma e^{t(b-a)}\big) \\ &= e^{-\gamma u}\left(1-\gamma+\gamma e^{t(b-a)}\right) \\ &= e^{-\gamma u}\left(1-\gamma+\gamma e^{u}\right). \end{aligned}$$

Taking the logarithm, define

$$g(u) = -\gamma u + \ln\left(1-\gamma+\gamma e^{u}\right),$$

so that

$$\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} = e^{g(u)}. \tag{3}$$

It follows form the exact form of Taylor's Theorem that for each $u > 0$, there is a point $\tilde{u} \in (0,u)$ satisfying

$$g(u) = g(0) = g'(0)u + \frac{1}{2}g''(\tilde{u})u^2.$$

Now observe that $g(0) = 0$, and

$$g'(u) = -\gamma + \frac{1}{1-\gamma+\gamma e^{u}}\gamma e^{u}$$

so $g'(0) = 0$, and

$$g''(u) = \frac{\gamma e^{u}(1-\gamma+\gamma e^{u}) - \big(\gamma e^{u}\big)^2}{(1-\gamma+\gamma e^{u})^2} = \frac{\gamma(1-\gamma)e^{u}}{(1-\gamma+\gamma e^{u})^2} = \frac{\gamma e^{u}}{1-\gamma+\gamma e^{u}}\,\frac{1-\gamma}{1-\gamma+\gamma e^{u}},$$

which is of the form $c(1-c)$ where $0 < c < 1$, so it is bounded above by $1/4$. Thus for $u > 0$, Taylor's Theorem implies

$$g\big(t(b-a)\big) = g(u) = \frac{1}{2}g''(\tilde{u})u^2 \leqslant \frac{1}{8}u^2 = \frac{1}{8}t^2(b-a)^2.$$

Returning now to (2) and (3) we (finally) conclude

$$e^{tX} \leqslant e^{g\left(t(b-a)\right)} \leqslant e^{t^2(b-a)^2/8}.$$

Whew! ∎

7.3.7 Proposition (Hoeffding's Inequality) *Let $X_1, \dots, X_n$ be a sequence of independent bounded random variables (not necessarily identically distributed), satisfying $\boldsymbol{E}\, X_i = 0$, and $a_i \leqslant X_i \leqslant b_i$, $i = 1, \dots, n$, and set*

$$S = \sum_{i=1}^{n} X_i.$$

Then for any $\varepsilon > 0$ and $t > 0$,

$$P(S \geqslant \varepsilon) \leqslant e^{-t\varepsilon} \prod_{i=1}^{n} e^{t^2(b_i - a_i)^2/8}.$$

Proof: Without loss of generality, we may assume each $a_i < 0 < b_i$. (Why?) The proof will use the moment generating functions of the X_i's so multiply both sides of $S \geqslant \varepsilon$ by t and exponentiating to get

$$P(S \geqslant \varepsilon) = P\big(e^{tS} \geqslant e^{t\varepsilon}\big).$$

Now we use Markov's Inequality 7.3.1 to get

$$P\big(e^{tS} \geqslant e^{t\varepsilon}\big) \leqslant \frac{\boldsymbol{E}\, e^{tS}}{e^{t\varepsilon}}. \tag{4}$$

Now we use the fact that the mgf of an independent sum is the product of the summands' mgfs, that is,

$$\boldsymbol{E}\, e^{tS} = \boldsymbol{E}\, e^{t(X_1 + \cdots + X_n)} = \boldsymbol{E} \prod_{i=1}^{n} e^{tX_i} = \prod_{i=1}^{n} \boldsymbol{E}\, e^{tX_i}, \tag{5}$$

where the last equality holds because the X_i's are independent.

The result now follows from Lemma 7.3.6. ∎

7.3.8 Corollary (Hoeffding's Inequality for the Bernoulli case) *Let $X_1, \dots, X_n$ be independent Bernoulli(p) random variables. Then for every $\varepsilon > 0$,*

$$P\big(\big|\bar{X} - p\big| \geqslant \varepsilon\big) \leqslant 2e^{-n\varepsilon^2},$$

where $\bar{X} = (X_1 + \cdots + X_n)/n$ is the sample fraction of successes.

The corollary is proved by applying Proposition 7.3.7 with $t = 4n\varepsilon$, first to $(X_i - p)/n$ (each $a_i = -p/n$ and $b_i = (1-p)/n$), and then to $(p - X_i)/n$.

7.3.4 Vysochanskiĭ–Petunin

The following inequalities are discussed in Pukelsheim [20], who supplies elementary proofs. They improve on the Bienaymé–Chebyshev Inequality for the common case of unimodal densities. I learned about them from Casella and Berger [4, p. 137].[2]

7.3.9 Definition *A distribution is **unimodal**, or has a **unimodal density**, if there is a density f and a **mode** m such that f is nondecreasing to the left of m and nonincreasing to the right of m.*

The mode m need not be unique. For instance, every $x \in [0, 1]$ is a mode of the Uniform$[0, 1]$ distribution.

Gauss [12] proved the following in 1823.

[2] Note: Casella and Berger [4, p. 137] contains an unfortunate typo. Their second case right-hand side in Proposition 7.3.11 is $\frac{4\xi^2}{9\varepsilon^2} - \frac{1}{3}$ instead of the correct $\frac{4\xi^2}{3\varepsilon^2} - \frac{1}{3}$.

7.3.10 Proposition (Gauss's Inequality) *Let X have a unimodal density with mode m and let $\tau^2 = \boldsymbol{E}(X-m)^2$. Then*

$$P(\,|X-m| > \varepsilon) \leqslant \begin{cases} \frac{4\tau^2}{9\varepsilon^2} & \text{for } \varepsilon \geqslant \frac{2}{\sqrt{3}}\tau \\ 1 - \frac{\varepsilon}{\sqrt{3}\tau} & \text{for } \varepsilon \leqslant \frac{2}{\sqrt{3}}\tau. \end{cases}$$

The following was proven by Vysochanskiĭ and Petunin [27, 28]. It replaces the mode m by an arbitrary point a.

7.3.11 Proposition (Vysochanskiĭ–Petunin Inequality) *Let X have a unimodal density. Given a, define $\xi^2 = \boldsymbol{E}(X-a)^2$. Then*

$$P(\,|X-a| > \varepsilon) \leqslant \begin{cases} \frac{4\xi^2}{9\varepsilon^2} & \text{for } \varepsilon \geqslant \frac{2\sqrt{2}}{\sqrt{3}}\xi \\ \frac{4\xi^2}{3\varepsilon^2} - \frac{1}{3} & \text{for } \varepsilon \leqslant \frac{2\sqrt{2}}{\sqrt{3}}\xi. \end{cases}$$

The following is a special case of the Vysochanskiĭ–Petunin Inequality, for $a = \boldsymbol{E}\,X$.

7.3.12 Corollary (The three-sigma rule) *If X has a unimodal density, and $\boldsymbol{E}\,X = \mu$ (μ is not necessarily equal to the mode) and $\boldsymbol{Var}\,X = \sigma^2$, then*

$$P(|X-\mu| > 3\sigma) \leqslant \frac{4}{81} < 0.05.$$

7.4 Sums and averages of i.i.d. random variables

Pitman [19]: pp. 193–195

Let $X_1, \dots, X_i, \dots$ be a sequence of independent and identically distributed random variables. For each n define a new random variable S_n by

$$S_n = \sum_{i=1}^{n} X_i.$$

These are the **partial sums** of the sequence. Assume $\mu = \boldsymbol{E}\,X_1$ ($= \boldsymbol{E}\,X_i$ for any i since the X_i are identically distributed). Since expectation is a linear operator, we have

$$\boldsymbol{E}\,S_n = n\mu, \qquad \text{so } \boldsymbol{E}\left(\frac{S_n}{n}\right) = \mu.$$

Let $\sigma^2 = \boldsymbol{Var}\,X_1$ ($= \boldsymbol{Var}\,X_i$ for any i since the X_i are identically distributed). Since the random variables X_i are independent, we have

$$\boldsymbol{Var}\,S_n = n\sigma^2 \qquad \text{and} \qquad \text{SD}\,S_n = \sqrt{n}\sigma$$

Recall that for any random variable Y, we have $\boldsymbol{Var}(aY) = a^2\ \boldsymbol{Var}\,Y$. Thus

$$\boldsymbol{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\ \boldsymbol{Var}\,S_n = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Thus the standard deviation (the square root of the variance) of the average S_n/n is given by

$$\text{standard deviation } \frac{S_n}{n} = \frac{\sigma}{\sqrt{n}}.$$

(Pitman [19, p. 194] calls this the **Square Root Law**.)

Note that the variance of the **sum** of n independent and identically distributedrandom variables grows linearly with n, so the standard deviation grows like $\sqrt{n}$; but the variance of the **average** is proportional to $1/n$, so the standard deviation is proportional to $1/\sqrt{n}$.

Aside: This is confusing to many people, and can lead to bad decision making. The Nobel prize-winning economist Paul Samuelson [21] reports circa 1963 that he offered "some lunch colleagues to bet each \$200 to \$100 that the side of coin *they* specified would not appear at the first toss."

A "distinguished colleague" declined the bet, but said, "I'll take you on if you promise to let me make 100 such bets." Samuelson explains that, "He and many others give something like the following explanation. 'One toss is not enough to make it reasonably sure that the law of averages will turn out in my favor. But in a hundred tosses of a coin, the law of large numbers will make it a darn good bet. I am so to speak, virtually sure to come out ahead in such a sequence ..."

Samuelson points out that this is *not* what the Law of Large Numbers guarantees. He says that his colleague "should have asked for to subdivide the risk and asked for a sequence of 100 bets each of which was \$1 against \$2."

Let's compare means, standard deviations and the probability of losing money,

Bet	Expected Value	Std. Dev.	Probability of being a net Loser
\$200 v. \$100 on one coin flip	\$50	\$111.80	0.5
100 bets of \$200 v. \$100	\$5000	\$1, 118.03	0.0003
100 bets of \$2 v. \$1	\$50	\$11.10	0.0003

Betting \$2 to \$1 has expectation \$0.50 and standard deviation \$1.11, so 100 such bets has mean \$50 and standard deviation \$11.10, which is a lot less risky than the first case.

7.5 The Weak Law of Large Numbers

It follows that as n gets large the variance of the partial sums S_n grows unboundedly, while the variance of the average S_n/n is shrinking to zero. That means that the averages are getting more and more concentrated around their mean μ. This is what is commonly referred to as the **Law of Averages**. There are a few versions. Here is one.

7.5.1 The Weak Law of Large Numbers, version 1 *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables, with common expectation μ and variance $\sigma^2 < \infty$. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_i, \qquad n = 1, 2, 3, \dots .$$

Let $\varepsilon > 0$ be given. Then

$$P\left(\left| \frac{S_n}{n} - \mu \right| \geqslant \varepsilon \right) \leqslant \frac{1}{n} \frac{\sigma^2}{\varepsilon^2}.$$

Proof: Note that if S is the sample space for a single experiment, this statement never requires consideration of a sample space more complicated than S^n. The result is simply the Bienaymé–Chebyshev Inequality (version 1) applied to the random variable S_n/n, which has mean μ and variance σ^2/n:

$$P\left(\left| \frac{S_n}{n} - \mu \right| \geqslant \varepsilon \right) \leqslant \frac{\sigma^2/n}{\varepsilon^2}.$$

■

An important fact about this theorem is it tells us how fast in n the probability of the deviation of the average from the expected value goes to zero: it's bounded by a constant times $1/n$.

7.6 ⋆ Convergence in probability

Not in the textbooks.

7.6.1 Definition *A sequence $Y_1, Y_2, \dots$ of random variables (not necessarily independent) on a common probability space $(\Omega, \mathcal{F}, P)$* ***converges in probability*** *to a random variable X if*

$$(\forall \varepsilon > 0) \left[\lim_{n\to\infty} P(|Y_n - Y| \geqslant \varepsilon) = 0; \right]$$

or equivalently

$$(\forall \varepsilon > 0) \left[\lim_{n\to\infty} P(|Y_n - Y| \leqslant \varepsilon) = 1; \right]$$

or equivalently

$$(\forall \varepsilon > 0)\ (\forall \delta > 0)\ (\exists N)\ (\forall n \geqslant N)\ [P(|Y_n - Y| > \varepsilon) < \delta],$$

in which case we may write

$$Y_n \xrightarrow{P} Y,$$

or

$$\operatorname*{plim}_{n\to\infty} Y_n = Y.$$

(The symbol plim *is pronounced "p-lim.")*

This allows us to rewrite the WLLN:

7.6.2 The Weak Law of Large Numbers, version 2 *Let $X_1, X_2, \dots$ be a sequence of independent an identically distributed random variables, with common expectation μ and variance $\sigma^2 < \infty$. Define the partial sums*

$$S_n = \sum_{i=1}^{n} X_i, \qquad n = 1, 2, 3, \dots .$$

Then

$$\operatorname*{plim}_{n\to\infty} \frac{S_n}{n} = \mu.$$

Note that this formulation throws away the information on the rate of convergence.

7.7 ⋆ "Infinitely often"

Let E_1, E_2, … be an infinite sequence of events in Ω. Which points belong to infinitely many of the E_n's?

For ω to belong to infinitely many of the E_n's, for each n there must be some $m \geqslant n$ for which $\omega \in E_m$. (Otherwise ω belongs to at most n of the sets.) That is,

$$\omega \in \bigcup_{m=n}^{\infty} E_m.$$

But this must be true for each n, so we must have

$$\omega \in \bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big).$$

This set is the event where elements in the sequence E_n occur "infinitely often," abbreviated

$$(E_n \text{ i.o.}) = \bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big).$$

This set is also known as $\limsup_n E_n$

The complementary event, that only finitely many of the events occur is by DeMorgan's laws

$$(E_n \text{ finitely often}) = \bigcup_{n=1}^{\infty}\big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\big).$$

Thus $\bigcup_{n=1}^{\infty}\big(\bigcap_{m=n}^{\infty} E_m\big)$ is the event that *all but finitely many* of the events E_1, E_2, ... occur. This set is also called the $\liminf_n E_n$.

Aside: The indicator functions of $\limsup_n E_n$ and $\liminf_n E_n$ are $\limsup_n \mathbf{1}_{E_n}$ and $\liminf_n \mathbf{1}_{E_n}$, respectively, if you know what those are.

7.8 ⋆ The Borel–Cantelli Lemma

The Borel–Cantelli Lemma is actually a pair of useful result concerning the general problem of events occurring infinitely often.

7.8.1 First Borel–Cantelli Lemma *Let E_1, E_2, ... be an infinite sequence of events in the probability space $(\Omega, \mathcal{F}, P)$.*

If $\sum_{n=1}^{\infty} P(E_n) < \infty$, then $P(E_n$ i.o.$) = 0$.

7.8.2 Second Borel–Cantelli Lemma *Let E_1, E_2, ... be an infinite sequence of events in the probability space $(\Omega, \mathcal{F}, P)$.*

If the events are mutually independent, and if $\sum_{n=1}^{\infty} P(E_n) = \infty$, then $P(E_n$ i.o.$) = 1$.

7.8.3 Example Before proceeding with the proof, here is a simple example that shows that the independence hypothesis is necessary for the Second Borel–Cantelli Lemma.

Let $\Omega = [0,1]$ with the uniform probability. Let $E_n = (0, 1/n)$. Then $\sum_{n=1}^{\infty} P(E_n) = \infty$, but $(E_n \text{ i.o.}) = \varnothing$. □

These results are standard, and proofs may be fond for instance in Jacod and Protter [15, pp. 71–72], or Breiman [3, Lemma 3.14, pp. 41–42].

Proof of the First Borel–Cantelli Lemma: Assume $\sum_{m=1}^{\infty} P(E_m) < \infty$. Then

$$\lim_{n\to\infty} \sum_{m=n}^{\infty} P(E_m) = 0.$$

Now

$$\begin{aligned}
P(E_n \text{ i.o.}) &= P\left(\bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big)\right) \\
&= \lim_{n\to\infty} P\Big(\bigcup_{m=n}^{\infty} E_m\Big) && \text{Proposition 2.4.3} \\
&\leqslant \lim_{n\to\infty} \sum_{m=n}^{\infty} P(E_m) && \text{Boole's Inequality 2.2.2} \\
&= 0.
\end{aligned}$$

∎

Proof of the Second Borel–Cantelli Lemma: Assume the events are mutually independent, and that $\sum_{n=1}^{\infty} P(E_n) = \infty$. Now

$$\begin{aligned}
P(E_n \text{ i.o.}) &= P\left(\bigcap_{n=1}^{\infty}\Big(\bigcup_{m=n}^{\infty} E_m\Big)\right) \\
&= \lim_{n\to\infty} P\Big(\bigcup_{m=n}^{\infty} E_m\Big) && \text{Proposition 2.4.3} \\
&= \lim_{n\to\infty} 1 - P\Big(\bigcup_{m=n}^{\infty} E_m\Big)^{\mathrm{c}} && P(E^{\mathrm{c}}) = 1 - P(E) \\
&= 1 - \lim_{n\to\infty} P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) && \text{de Morgan's Laws.}
\end{aligned}$$

So to prove $P(E_n \text{ i.o.})$, it suffices to show that for any n,

$$P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) = 0. \tag{6}$$

Since the events E_n are mutually independent, so are the events E_n^{c} (Lemma 2.6.3), so

$$P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} P\Big(\bigcap_{m=n}^{k} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} \prod_{m=n}^{k} P(E_m^{\mathrm{c}}) = \lim_{k\to\infty} \prod_{m=n}^{k} \big(1 - P(E_m)\big).$$

Taking logarithms implies

$$\lim_{k\to\infty} \ln P\Big(\bigcap_{m=n}^{k} E_m^{\mathrm{c}}\Big) = \lim_{k\to\infty} \sum_{m=n}^{k} \ln P(E_m^{\mathrm{c}}) = \lim_{k\to\infty} \sum_{m=n}^{k} \ln\big(1 - P(E_m)\big).$$

We now use the fact that the logarithm is a concave function so by the Subgradient Inequality (see Section 6.9 ⋆) the first-order Taylor series overestimates the logarithm. That is,

$$\ln(1-x) \leqslant \ln(1) + \ln'(1)(-x) = -x.$$

Thus

$$\sum_{m=n}^{k} \ln\big(1 - P(E_m)\big) \leqslant -\sum_{m=n}^{k} P(E_m).$$

Letting $k \to \infty$ we have

$$\ln P\Big(\bigcap_{m=n}^{\infty} E_m^{\mathrm{c}}\Big) \leqslant -\sum_{m=n}^{\infty} P(E_m) = -\infty,$$

which proves (6). ∎

7.8.4 Remark We can prove the First Borel–Cantelli Lemma by the "method of indicators." Let

$$X(\omega) = \#\left\{i : \omega \in E_i\right\} = \sum_{i=1}^{\infty} \mathbf{1}_{E_i}.$$

Proposition 8.2.1 says that for each n,

$$\boldsymbol{E} \sum_{i=1}^{n} \mathbf{1}_{E_i} = \sum_{i=1}^{n} P(E_i),$$

so by the Monotone Convergence Theorem 5.13.2,

$$\boldsymbol{E}\, X = \boldsymbol{E} \sum_{i=1}^{\infty} \mathbf{1}_{E_i} = \sum_{i=1}^{\infty} P(E_i).$$

So if $\boldsymbol{E}\, X < \infty$, it must be that $P(X = \infty) = 0$, but the event $(X = \infty) = (E_n \text{ i.o.})$.

7.9 ⋆ Tail events and Zero-One laws

In a similar vein we have the following result, which I shall not prove here.

Let $X_1, \dots, X_n, \dots$ is a sequence of random variables. An event E is a **tail event** if it belongs to the σ-algebra generated by the tail sequence $X_n, X_{n+1}, \dots$ for every n. Letting $\sigma(X_n, X_{n+1}, \dots)$ denote the σ-algebra generated by the tail sequence, a tail event is an element of the **tail $\boldsymbol{\sigma}$-algebra**

$$\bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \dots).$$

7.9.1 Kolmogorov Zero-One Law *If $X_1, \dots, X_n, \dots$ is a sequence of mutually independent random variables, and E is a tail event, then either $P(E) = 0$ or $P(E) = 1$.*

For a proof see, e.g., Jacod and Protter [15, Theorem 10.6, p. 72], or Breiman [3, Theorem 3.12, p. 40].

7.10 ⋆ The Strong Law of Large Numbers

There is another version of the Law of Averages that in some ways strengthens the Weak Law, and is called the **Strong Law of Large Numbers** or sometimes **Kolmogorov's Strong Law of Large Numbers**. When I first encountered it in 1975, the proof of the Strong Law was a *tour de force* of "hard analysis," chock full ε's and δ's and clever approximations and intermediate lemmas. Instead, I will discuss the "elementary" proof by Etemadi [9]. His proof is rather terse, and my exposition benefits from that of Durrett [8, Section 2.4, pp. 65–67]

One of the ways that the Strong Law strengthens the Weak Law is that it drops the restriction that the variables have finite variance. Etemadi's version also uses pairwise independence, not mutual independence.

Not in the textbooks.

7.10.1 Strong Law of Large Numbers *Let $X_1, X_2, \dots$ be a sequence of pairwise independent and identically distributed random variables on the probability space $(\Omega, \mathcal{F}, P)$, with common finite expectation μ. Define the partial sums*

$$S_n = X_1 + \cdots + X_n, \qquad n = 1, 2, 3, \dots .$$

Then

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{} \mu \ a.s.$$

So you can appreciate how such things can be proven, and the amount of hard work and cleverness that went into the proof, I give a proof in Section 7.14⋆.

7.11 ⋆ Convergence of Empirical CDFs

A corollary of the Law of Large Numbers is that the empirical distribution function of a sample of independent and identically distributed random variables converges to the ideal (theoretical) cumulative distribution function. In fact, this may be the best way to judge how well your data conforms to your theoretical model.

Not in the textbooks.

7.11.1 Definition *Given random variables $X_1, X_2, \dots, X_n, \dots$, for each n, the* ***empirical cumulative distribution function*** *F_n evaluated at x is defined to be the fraction of the first n random variables that have a value $\leqslant x$. That is,*

$$F_n(x) = \frac{\#\{i : i \leqslant n \ \&\ X_i \leqslant x\}}{n}.$$

This makes each $F_n(x)$ a random variable, or each F_n a ***random function****, so more pedantically, letting Ω denote the sample space on which the random variables are defined, what I should have written is*

$$F_n(x)(\omega) = \frac{\#\{i : i \leqslant n \ \&\ X_i(\omega) \leqslant x\}}{n}.$$

Recalling that the indicator function $\mathbf{1}_{(-\infty,x]}(y)$ is equal to one if $y \leqslant x$ and zero otherwise, we may rewrite this as

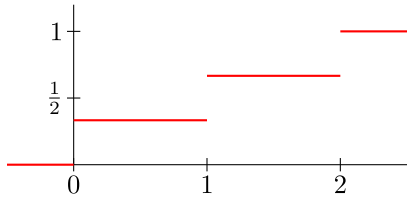
$$F_n(x) = \frac{\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)}{n}$$

7.11.2 Example Let X_i, $i = 1, 2$, be independent uniform random variables on the finite set $\{0, 1, 2\}$. (One way to get such a random variable is to roll a die, divide the result by 3, and take the remainder. Then 1 or 4 gives $X_i = 1$, 2 or 5 gives $X_i = 2$, and 3 or 6 gives $X_i = 0$.)

For the repeated experiments there are 9 possible outcomes:

$$(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2).$$

Each of these outcomes has an empirical cumulative distribution function F_2 associated with it. There however only 6 distinct functions, since different points in the sample space may give rise to the same empirical cdf. Table 7.1 shows the empirical cdf associated to each point in the sample space. You should check that for each x if you look at the value of the empirical cdf at x and weight them by the probabilities associated with each empirical cdf, the resulting function is the cdf of the distribution of each X_i:



□

Now assume that $X_1, \dots, X_n$ are independent and identically distributed, with common cumulative distribution function F. Since

$$\boldsymbol{E}\,\mathbf{1}_{(-\infty,x]}(X_i) = \operatorname{Prob}(X_i \leqslant x) = F(x),$$

and since expectation is a linear operator, for each x we have

$$\boldsymbol{E}\,F_n(x) = \boldsymbol{E}\,\frac{\sum_{i=1}^n \mathbf{1}_{(-\infty,x]}(X_i)}{n} = F(x).$$

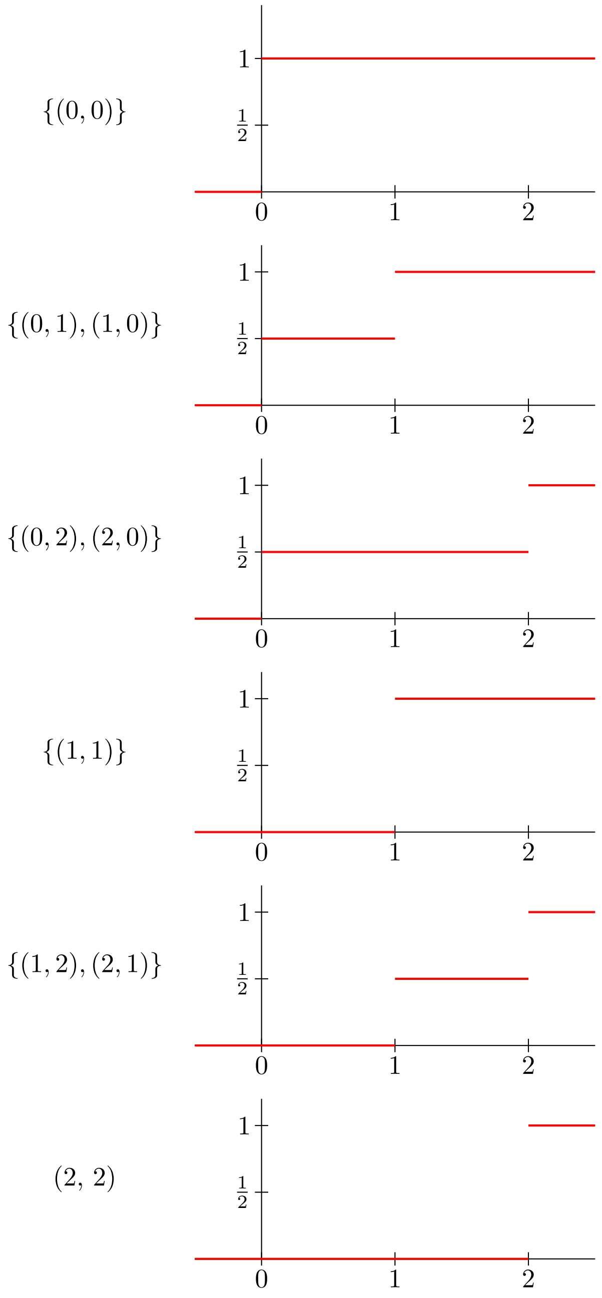


Table 7.1. Empirical cdfs at different points in the sample space for Example 7.11.2.

Now what is the variance of an indicator function? An indicator $\mathbf{1}_A$ is just a Bernoulli random variable with probability of success $P(A)$. Thus its variance is $P(A) - P(A)^2$, which is certainly finite. It follows from the Weak Law of Large Numbers that for each n we have the following

$$\operatorname{Prob}\left(|F_n(x) - F(x)| \geqslant \varepsilon\right) \leqslant \frac{F(x) - F(x)^2}{n\varepsilon^2},$$

and from the Strong Law that

$$\operatorname{Prob}\left(F_n(x) \to F(x)\right) = 1.$$

This result is a weaker version of the following theorem on uniform convergence, which we shall not prove. You can find it, for example, in Kai Lai Chung [6, Theorem 5.5.1, p. 133], or Durrett [8, Theorem 2.4.9, p. 68].

7.11.3 Glivenko–Cantelli Theorem *Assume $X_1, \dots, X_n, \dots$ are independent and identically distributed, with common cumulative distribution function F. Then*

$$\operatorname{Prob}\left(\sup_{x \in \boldsymbol{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{} 0\right) = 1.$$

7.12 ⋆ Resampling

The practical implication of the Glivenko–Cantelli Theorem is this:

Since with probability one, the empirical cdf converges to the true cdf, we can estimate the cdf **nonparametrically**. That is, we don not have to choose a parametrized family of distributions in order to identify the cdf.

Moreover, since the empirical cdf is also the cdf of a uniform random variable on the sample values, with probability one, as the sample size n becomes large, **resampling** by drawing sample points independently at random with replacement is (almost) the same as getting new independent sample points. This resampling scheme is called the **bootstrap**, and is the basis for many nonparametric statistical procedures.

7.12.1 Remark One of the philosophical questions regarding sampling that statisticians have dealt with is how to treat a population such as the U.S. population. At any given instant, the population and its characteristics are fixed, but the population is constantly changing. One way to think about this is that the U.S. population is an independent sample of approximately 328 million (as of 2019) draws from some underlying ideal U.S. population. So what does a sample from the realized population tell us about the the ideal population? According to the Glivenko–Cantelli Theorem, the difference is for almost all purposes, negligible.

The Glivenko–Cantelli Theorem also guarantees that we can find rules for creating **histograms** of our data that converge to the underlying density from which it is drawn.

7.13 ⋆ Histograms and densities

A **histogram** of a set of numbers is a bar graph that helps to visualize their distribution. The numbers are placed in **bins** or **classes** (that is, nonoverlapping intervals), and the height of the bar indicates the number of data in that bin. The histogram was named and popularized by the statistician Karl Pearson. Scott [22] reports that the use of histograms goes back at least to the mortality studies of John Graunt in 1662, and cites Westergaard [31].

If the data are the results of independent and identically distributed draws from a density f, and if the histogram is normalized so that the total area of the bars is one, then the histogram can be used to approximate the density. As the sample size gets larger, we would expect that the normalized histogram would get closer to the density. In order for this to happen, the width of bins must shrink with the sample size.

This raises the question of how to choose the width and number of the bins. There is also the issue of making sure that data do not fall on bin boundaries, but that is relatively straightforward.

- Herbert Sturges [25] argued that if the sample size n is of the form 2^k, and the data are approximately Normal, we can choose $k+1$ bins so that the number in bin j should be close to the binomial coefficient $\binom{n}{j}$. "For example, 16 items would be divided normally into 5 classes, with class frequencies, 1, 4, 6, 4, 1." For sample sizes n that are not powers of two, he argued that the range of the data should be divided into the number of bins that is the largest "convenient" integer (e.g., a multiple of 5) that is less than or equal to

$$1 + \log_2 n.$$

- David Freedman and Persi Diaconis [11] argue that one should try to minimize the L_2-norm of the difference between the normalized histogram and the density. This depends on unknown parameters, but they argued that the following rule is simple and approximately correct:

> For the cell width, take twice the interquartile range of the data, and divide by the cube root of the sample size.

- David Scott [22] argues that the mean-square error minimizing bin width for large n is given by

$$\left(\frac{6}{n\int_{-\infty}^{\infty} f'(x)^2\,dx}\right)^{1/3}.$$

- Matt Wand [29] derived more sophisticated methods for binning data based on kernel estimation theory.

- Kevin Knuth [16] has recently suggested an algorithmic method for binning based on a Bayesian procedure. It reportedly works better for densities that are not unimodal.

Each of these methods will create histograms that approximate the density better and better as the sample size increases. Both R and Mathematica have options for using the Sturges, Freedman–Diaconis, and Scott rules in plotting histograms. Mathematica implements Wand's method and it is available as an R package (`dpih`). Mathematica also implements Knuth's rule.

7.14★ Proof of the Strong Law of Large Numbers

The proof here is based on Durrett's [8, Section 2.4, pp. 65–67] exposition of Etemadi [9].

7.14.1 Remark Here is the strategy for the proof.

In order to prove that $S_n/n \to \mu$ a.s. we need to show that for every $\varepsilon > 0$, for every $\omega \in \Omega$ (or more precisely, for all ω in some subset $A \subset \Omega$, with $P(A) = 1$), there is some $N(\omega)$ such for all $n \geqslant N(\omega)$, $\bigl|\bigl(S_n(\omega)/n\bigr) - \mu\bigr| \leqslant \varepsilon$. In other words, we need to show that with probability one the events $E_n = \bigl(\bigl(S_n/n\bigr) - \mu > \varepsilon\bigr)$ do not occur infinitely often. The First Borel–Cantelli Lemma says that it suffices to show that for $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\Bigl(\bigl|\bigl(S_n(\omega)/n\bigr) - \mu\bigr| > \varepsilon\Bigr) < \infty.$$

This is the ultimate goal, but we take a roundabout path to get there.

We start by introducing “truncated” versions of the X_k’s. (Cf. Loève [18, p. 245].) Define

$$Y_k = X_k \mathbf{1}_{(\,|X_k| \leqslant k)} = \begin{cases} |X_k| & |X_k| \leqslant k \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and define the totals

$$T_n = Y_1 + \cdots + Y_n.$$

We show in Lemma 7.14.3 that

$$\boldsymbol{E}\, Y_k \to \boldsymbol{E}\, X_1 = \mu \text{ as } k \to \infty.$$

and we show in Lemma 7.14.2 that

$$P(Y_k \neq X_k \text{ i.o.}) = 0.$$

This in turn implies that T_n/n and S_n/n have the same limit (Lemma 7.14.4).

So we will prove instead that $T_n/n \to \mu$ a.s., which we shall show by proving that

$$\sum_{n=1}^{\infty} P(\,|(T_n/n) - \boldsymbol{E}(T_n/n)| > \varepsilon) < \infty.$$

Since each Y_k is bounded, it has a finite variance, so we can use the Bienaymé–Chebyshev Inequality to get a handle on each term $P(\,|(T_n/n) - \boldsymbol{E}(T_n/n)| > \varepsilon)$ of the above series. But that is not enough to easily sum the series, so we sum a subsequence instead. The subsequence is governed by a parameter $\alpha > 1$ via $k_n = \lfloor \alpha^n \rfloor$. We shall use Lemma 7.14.6 below to show that

$$\sum_{n=1}^{\infty} P(\,|(T_{k_n}/k_n) - \boldsymbol{E}(T_{k_n}/k_n)| > \varepsilon) < \infty,$$

so that

$$T_{k_n}/k_n \to \mu \text{ as } n \to \infty. \tag{8}$$

Durrett claims that so far this is standard stuff, and can be found for instance in Loève [18, Equivalence Lemma, p. 245–246], but Etemadi’s insight is that it suffices to prove the result for nonnegative X_k’s. This is because each $X_k = X_k^+ - X_k^-$, and each of X_k^+ and X_k^- have finite expectations. Also, if the X_k’s are pairwise independent, then the X_k^+’s are also pairwise independent. So assuming that the X_k’s are nonnegative, we show that (8) implies that

$$\frac{1}{\alpha}\mu \leqslant \liminf T_n/n \leqslant \limsup T_n/n \leqslant \alpha\mu,$$

where $\alpha > 1$ is the parameter used to generate the subsequence k_n. (For those of you have not yet had Ma 108, lim inf and lim sup give the worst case subsequences. If they are equal, then the limit exists and agrees with both.) But since α is arbitrary, letting $\alpha \searrow 1$, gives the final result.

Keep this guide in mind as we wade into the morass of details.

Proof of the Strong Law of Large Numbers 7.10.1: We start the following.

7.14.2 Lemma *Define Y_k as in equation* (7). *Then*

$$P(Y_k \neq X_k \text{ i.o.}) = 0.$$

Proof: Let E_k be the event that $(X_k \neq Y_k) = (\,|X_k| > k)$. Now

$$\begin{aligned}
\sum_{k=1}^{\infty} P(E_k) &= \sum_{k=1}^{\infty} P(\,|X_k| > k) && \text{definition of } Y_k \\
&= \sum_{k=1}^{\infty} P(\,|X_1| > k) && X_k\text{'s are identically distributed}
\end{aligned}$$

but $P(\,|X_1| > k) \geqslant P(\,|X_1| > t)$ for each $k-1 \leqslant t \leqslant k$, so

$$\begin{aligned}
&\leqslant \int_0^{\infty} P(\,|X_1| > t)\,dt \\
&= \boldsymbol{E}\,|X_1| && \text{integration by parts, see Proposition 6.4.3} \\
&< \infty.
\end{aligned}$$

Therefore from the First Borel–Cantelli Lemma 7.8.1 it follows that $P(E_n \text{ i.o.}) = 0$. This proves the lemma. ∎

7.14.3 Lemma

$$\boldsymbol{E}\,Y_n \to \boldsymbol{E}\,X_1 = \mu \textit{ as } n \to \infty.$$

Proof: Since each X_k has the same distribution F, for every n, $X_k\mathbf{1}_{(X_k \leqslant n)}$ has a distribution independent of k so the Dominated Convergence Theorem (or assuming the nonnegative case, the Monotone Convergence Theorem) yields

$$\boldsymbol{E}\,X_1 = \lim_{n\to\infty} \boldsymbol{E}\,X_1\mathbf{1}_{(X_1 \leqslant n)} = \lim_{n\to\infty} \boldsymbol{E}\,X_n\mathbf{1}_{(X_n \leqslant n)} = \lim_{n\to\infty} \boldsymbol{E}\,Y_n.$$

∎

7.14.4 Lemma *It suffices to prove that $T_n/n \to \mu$ a.s..*

Proof: By Lemma 7.14.2, for each ω (more precisely for each $\omega \in A$, where $P(A) = 1$) there is a number $N(\omega) < \infty$ such that

$$(\,\forall k \geqslant N(\omega)\,)\ [\,Y_k(\omega) = X_k(\omega)\,].$$

Now for $n > N(\omega)$,

$$\begin{aligned}
\frac{S_n(\omega)}{n} &= \underbrace{\frac{\sum_{k=1}^{N(\omega)} X_k(\omega)}{n}}_{\to 0} + \frac{\sum_{k=N(\omega)+1}^{n} X_k(\omega)}{n} \\
\frac{T_n(\omega)}{n} &= \underbrace{\frac{\sum_{k=1}^{N(\omega)} Y_k(\omega)}{n}}_{\to 0} + \frac{\sum_{k=N(\omega)+1}^{n} X_k(\omega)}{n},
\end{aligned}$$

which shows that S_n/n and T_n/n have the same limit. This ends the proof of Lemma 7.14.4. ∎

Since each Y_k is a bounded random variable, it will have a finite variance. The next step is to show:

7.14.5 Claim

$$\sum_{k=1}^{\infty} \frac{\boldsymbol{Var}\,Y_k}{k^2} \leqslant 4\,\boldsymbol{E}\,|X_1| < \infty. \tag{9}$$

Proof: We use Proposition 6.4.3, which implies that

$$\boldsymbol{E}\,Y_k^2 = \int_0^\infty 2yP(\,|Y_k| > y)\,dy$$

to get

$$\boldsymbol{Var}\,Y_k \leqslant \boldsymbol{E}\,Y_k^2 = \int_0^\infty 2yP(\,|Y_k| > y)\,dy \leqslant \int_0^k 2yP(\,|X_1| > y)\,dy.$$

(Remember $Y_k = 0$ if $X_k > k$, and the X_i's are identically distributed). Thus

$$\begin{aligned}\sum_{k=1}^\infty \frac{\boldsymbol{Var}\,Y_k}{k^2} &\leqslant \sum_{k=1}^\infty \frac{1}{k^2}\int_0^\infty \mathbf{1}_{(y<k)}(y)\,2yP(\,|X_1| > y)\,dy \\ &= \int_0^\infty \left(\sum_{k=1}^\infty \frac{1}{k^2}\mathbf{1}_{(y<k)}(y)\right) 2yP(\,|X_1| > y)\,dy\end{aligned} \tag{10}$$

where $\mathbf{1}_{(y<k)}(y) = 1$ if $y < k$, and $= 0$ otherwise; and the final equality comes from interchanging summation and integration (Fubini's Theorem—I really should be more explicit about it somewhere).

We now use the following algebraic fact (derived in Lemma 7.14.6 below): If $y > 0$, then

$$2y\sum_{k>y}\frac{1}{k^2} \leqslant 4.$$

Substituting this into (10) and recalling Proposition 6.4.3 gives

$$\sum_{k=1}^\infty \frac{\boldsymbol{Var}\,Y_k}{k^2} \leqslant \int_0^\infty \left(\sum_{k=1}^\infty \frac{1}{k^2}\mathbf{1}_{(y<k)}(y)\right) 2yP(\,|X_1| > y)\,dy \leqslant \int_0^\infty 4P(\,|X_1| > y)\,dy = 4\,\boldsymbol{E}\,|X_1|\,.$$

This proves Claim 7.14.5. ∎

Pick $\varepsilon > 0$. By the Bienaymé–Chebyshev Inequality 7.3.3 for any k we have

$$P(\,|(T_k/k) - \boldsymbol{E}(T_k/k)| > \varepsilon) = P(\,|T_k - \boldsymbol{E}\,T_k| > \varepsilon k) \leqslant \boldsymbol{Var}\,T_k/(\varepsilon k)^2 = \frac{1}{(\varepsilon k)^2}\sum_{m=1}^k \boldsymbol{Var}\,Y_m$$

since the variance of a sum of pairwise independent random variables is the sum of their variances.[3]

We really want to show that $\sum_{k=1}^\infty P(\,|T_k - \boldsymbol{E}\,T_k| > \varepsilon k) < \infty$, so we could use the Borel–Cantelli Lemma I to show that $|T_k - \boldsymbol{E}\,T_k| \to 0$, but we (at least I) cannot do that directly. Instead we have to take a roundabout approach and show that the sum over a certain kind of subsequence is finite. The we let the subsequences become more inclusive to show that the entire sum is finite. So get comfortable and prepare for an excursion.

Pick an arbitrary $\alpha > 1$, and consider a subsequence $k_n = \lfloor \alpha^n \rfloor$, where $\lfloor x \rfloor$ denotes the integer part of x, the largest integer $\leqslant x$. Then (i) it is possible that $k_{n+1} = k_n$, but for n large enough, $k_{n+1} > k_n$; and (ii) $\lim_{n\to\infty} k_{n+1}/k_n = \alpha$. Summing over this subsequence gives

$$\sum_{n=1}^\infty P(\,|T_{k_n} - \boldsymbol{E}\,T_{k_n}| > \varepsilon k_n) \leqslant \sum_{n=1}^\infty \boldsymbol{Var}\,T_{k_n}/(\varepsilon k_n)^2 = \sum_{n=1}^\infty \frac{1}{(\varepsilon k_n)^2}\sum_{m=1}^{k_n} \boldsymbol{Var}\,Y_m. \tag{11}$$

[3] Is this true?

We can use Fubini's Theorem to interchange the order of summation to get

$$\sum_{m=1}^{k_n} \frac{1}{(\varepsilon k_n)^2} \sum_{n=1}^{\infty} \boldsymbol{Var}\, Y_m = \frac{1}{\varepsilon^2} \sum_{(n,m):k_n \geqslant m} \frac{1}{k_n^2}\, \boldsymbol{Var}\, Y_m = \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \boldsymbol{Var}\, Y_m \underbrace{\sum_{n:k_n \geqslant m} \frac{1}{k_n^2}}_{a}. \tag{12}$$

Let's examine the last term a above. We would like to get a nice upper bound on $\frac{1}{k_n^2} = \frac{1}{\lfloor \alpha^n \rfloor^2}$. I claim that for $\alpha > 1$ and $n \geqslant 1$ we have $\lfloor \alpha^n \rfloor > \alpha^n/2$.[4] Inverting this and squaring gives

$$\frac{1}{\lfloor \alpha^n \rfloor^2} < \frac{4}{\alpha^{2n}}.$$

Recalling that $k_n = \lfloor \alpha^n \rfloor$ we can rewrite term a as

$$a = \sum_{n:k_n \geqslant m} \frac{1}{k_n^2} = \sum_{n:\alpha^n \geqslant m} \frac{1}{\lfloor \alpha^n \rfloor^2} < 4 \underbrace{\sum_{n:\alpha^n \geqslant m} \frac{1}{(\alpha^2)^n}}_{\text{partial series } b}. \tag{13}$$

We now find a bound for the partial geometric series b. For any geometric series $\sum_{n=k}^{\infty} \beta^n$, with $0 < \beta < 1$, we have $\sum_{n=k}^{\infty} \beta^n = \beta^k \sum_{n=0}^{\infty} \beta^n = \beta^k/(1-\beta)$. Note that β^k is the first term in partial series. Now the first index in the series b is some n with $\alpha^n \geqslant m$, so $1/\alpha^n \leqslant 1/m$, which implies

$$\left(\frac{1}{\alpha^2}\right)^n \leqslant \left(\frac{1}{m}\right)^2.$$

The left-hand side is the first term in the partial series b, so we may use this in (13) to get

$$a = \sum_{n:k_n \geqslant m} \frac{1}{k_n^2} < 4 \underbrace{\sum_{n:\alpha^n \geqslant m} \frac{1}{(\alpha^2)^n}}_{b} \leqslant 4 \frac{1}{m^2 \left(1 - \frac{1}{\alpha^2}\right)}.$$

Combining this with (11) gives

$$\begin{aligned} P\left(|(T_k/k) - \boldsymbol{E}(T_k/k)| > \varepsilon \right) &= \sum_{n=1}^{\infty} P\left(|T_{k_n} - \boldsymbol{E}\, T_{k_n}| > \varepsilon k_n \right) \\ &\leqslant \frac{1}{\varepsilon^2} \sum_{m=1}^{\infty} \boldsymbol{Var}\, Y_m \sum_{n:k_n \geqslant m} \frac{1}{k_n^2} \\ &\leqslant 4 \frac{1}{\varepsilon^2 \left(1 - \frac{1}{\alpha^2}\right)} \sum_{m=1}^{\infty} \frac{\boldsymbol{Var}\, Y_m}{m^2} \\ &< \infty, \end{aligned}$$

where the final inequality is a result of (9). Thus as in the proof of Lemma 7.14.4,

$$\frac{|T_{k_n} - \boldsymbol{E}\, T_{k_n}|}{k_n} \to 0 \text{ as } n \to \infty. \tag{14}$$

Lemma 7.14.3 combined with equation (14) yields

$$\frac{T_{k_n}}{k_n} \to \mu \text{ as } k \to \infty.$$

[4] To see this, let $x > 1$. Then $\lfloor x \rfloor \geqslant 1$. Moreover $1 > x - \lfloor x \rfloor$, so $\lfloor x \rfloor > x - \lfloor x \rfloor$. Add $\lfloor x \rfloor$ to both sides to get $2\lfloor x \rfloor > x$, divide by 2, and evaluate at $x = \alpha^n$.

Thus we have a convergent subsequence, but we need to fill in the intermediate terms. Now we are about to assume that the X_k's (and hence Y_k's) are nonnegative. In this case, for $k_n \leqslant m \leqslant k_{n+1}$, we have

$$T_{k_n} \leqslant T_m \leqslant T_{k_{n+1}},$$

and then by dividing the smaller terms by the larger ones yields

$$\frac{T_{k_n}}{k_{n+1}} \leqslant \frac{T_m}{m} \leqslant \frac{T_{k_{n+1}}}{k_n}$$

so

$$\underbrace{\frac{k_n}{k_{n+1}}}_{\to 1/\alpha} \frac{T_{k_n}}{k_n} \leqslant \frac{T_m}{m} \leqslant \underbrace{\frac{k_{n+1}}{k_n}}_{\to \alpha} \frac{T_{k_{n+1}}}{k_{n+1}}$$

Recalling that $k_n = \lfloor \alpha^n \rfloor$, we see that $1 \leqslant k_{n+1}/k_n \nearrow \alpha$, we have

$$\frac{1}{\alpha} \underbrace{\frac{T_{k_n}}{k_n}}_{\to \mu} \leqslant \frac{T_m}{m} \leqslant \alpha \underbrace{\frac{T_{k_{n+1}}}{k_{n+1}}}_{\to \mu}.$$

Letting $n \to \infty$, we have

$$\frac{1}{\alpha}\mu \leqslant \liminf_m T_m/m \leqslant \limsup_m T_m/m \leqslant \alpha\mu.$$

Now letting $\alpha \searrow 1$, gives

$$\lim_{n\to\infty} \frac{T_n}{n} = \mu,$$

so by Lemma 7.14.4 the proof of the strong law is complete. ∎

The above proof relied on the following fact.

7.14.6 Lemma *If $y \geqslant 0$, then*

$$2y \sum_{k>y} \frac{1}{k^2} \leqslant 4.$$

Proof: Following Durrett [8, Lemma 2.4.4, p. 66], break the result into two cases: (i) $y \geqslant 1$, and (ii) $0 \leqslant y < 1$.

First observe that for an integer $m \geqslant 2$, we have

$$\sum_{k>m} \frac{1}{k^2} < \int_{m-1}^{\infty} \frac{1}{x^2}\,dx = \frac{1}{m-1}.$$

Case (i): Assume $y \geqslant 1$, then the first integer $k > y$ is $\lfloor y \rfloor + 1 \geqslant 2$, so by the observation just noted,

$$2y \sum_{k>y} \frac{1}{k^2} < \frac{2y}{\lfloor y \rfloor} \leqslant 4.$$

Case (ii): Let $0 \leqslant y < 1$, then observe that

$$2y \sum_{k>y} \frac{1}{k^2} \leqslant 2\left(1 + \sum_{k=2}^{\infty} \frac{1}{k^2}\right) < 4.$$

∎

Aside: Euler proved that $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \approx 1.6$, but the proof is beyond the scope of this course.

The next result is a converse to the strong law that shows that finiteness of the expectation cannot be dispensed with. It may be found in Feller [10, Theorem 4, p. 241].

7.14.7 Theorem *Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables. If $\boldsymbol{E}\,|X_i| = \infty$, then for any numerical sequence c_n,*

$$\limsup_{n\to\infty} \left| \frac{S_n}{n} - c_n \right| = \infty \text{ with probability one.}$$

One implication of this is that if the expectation is not finite, the averages will become arbitrarily large infinitely often with probability one.

7.15 ⋆ Sample spaces for independent and identically distributed random variables

Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables.

WHOA! What kind of sample space are we talking about here? Well let's start out with a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X on Ω. To get n repeated trials of the experiment, we take as our sample space Ω^n with the probability P^n.[5] When Ω is finite (or discrete) P^n is defined by

Not in the textbooks.

$$P^n\big((\omega_1, \dots, \omega_n)\big) = P(\omega_1)P(\omega_2)\cdots P(\omega_n).$$

One way to get a finite sequence $X_1, \dots, X_n$ of independent and identically distributed random variables is to take as your sample space Ω^n with probability P^n and for each point $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$ in Ω^n define

$$X_i(\boldsymbol{\omega}) = X(\omega_i).$$

This is fine as long as we always get to work with some finite n. This is adequate for the Weak Law of Large Numbers, but it won't do for the Strong Law. For that we want a probability measure on the set of infinite sequences, $\Omega^\infty = \Omega \times \Omega \times \cdots$. "What's the big deal?" you may ask. For a sequence $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ why don't we just set the probability of $\boldsymbol{\omega}$ to $P(\omega_1)P(\omega_2)P(\omega_3)\cdots$? Well, unless X is degenerate this will always be zero (assuming that we mean the limit of the finite products), which is not very helpful. Moreover, even if Ω has only two elements (as in the Bernoulli trials case), Ω^∞ is uncountably infinite, which we know means measure-theoretic trouble. Nevertheless, there is a meaningful way (provided Ω is not too weird) to put a probability measure P^∞ on Ω^∞, so that defining $X_i(\boldsymbol{\omega}) = X(\omega_i)$ makes $X_1, X_2, \dots$ a sequence of independent and identically distributed random variables on Ω^∞, which all have the same distribution as X. This result is known as the **Kolmogorov Extension Theorem**, after Andrey Kolmogorov. Proving it requires a course in measure theory and topology, but you can find an excellent proof in Roko Aliprantis and KC Border [1, Theorem 15.6, p. 522].

7.16 ⋆ Comparing the Weak and Strong Laws

In order to understand the difference between the Weak and Strong Laws, let's confine our attention to the case where the expectation is zero. In order to understand the Weak Law, we only needed to compute probabilities on the finite product space Ω^n, which given independence we get my multiplying probabilities defined by P. In order to frame the Strong Law we take the sample space to be $\boldsymbol{\Omega} = \Omega^\infty$, the space of all infinite sequences sample outcomes. A typical element in $\boldsymbol{\Omega}$ is an infinite sequence $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ of outcomes in Ω. Given random variables

[5] The set of events is what is called the product σ-algebra $\mathcal{F}^n$, and I won't go into details about that now.

$X_1, X_2, \dots$ defined on the common sample space Ω, by abusing notation, we may identify them with random variables X_i on the big sample space $\boldsymbol{\Omega} = \Omega^\infty$ by $X_i\big((\omega_1, \omega_2, \dots)\big) = X_i(\omega_i)$. This space has a probability measure P^∞ that is consistent with P, but we typically cannot compute probabilities as "infinite products."

Not in the textbooks.

The Strong Law says that the set of infinite sequences $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots)$ in $\boldsymbol{\Omega}$ for which $S_n(\boldsymbol{\omega})/n$ becomes small (recall we are in the mean = 0 case) at some index N in the sequence and stays small for all $n \geqslant N$ has probability one under the measure P^∞. The catch is that the Strong Law does not give a hint as to how large N is. The Weak Law says that with probability $\big((n-1)/n\big)(\sigma^2/\varepsilon^2)$, the average S_n/n is within ε of the mean zero. By taking n large enough we can make this probability as close to one as we want, but we can't be sure. In fact, there can be (infinitely many) sequences $\boldsymbol{\omega}$ for which S_n/n becomes arbitrarily large for infinitely many n. (It is just that the set of these sequences has probability zero under P^∞.)

So one difference is that the Weak Law tells us how big n has to be to get the degree of (still positive) uncertainty we wish, but the Strong Law assures us that S_n/n will surely become small and stay small, but we just don't know when.

Kai Lai Chung [7, p. 233] points out that the relevance of the two versions is matter of dispute among probabilists. For instance, William Feller [10, p. 237] argues that,

> "In practice one is rarely interested in the probability $P\big(n^{-1}\,|S_n| > \varepsilon\big)$ for any particular large value of n. A more interesting question is whether $n^{-1}\,|S_n|$ will ultimately become and remain small, that is, whether $n^{-1}\,|S_n| < \varepsilon$ simultaneously for all $n \geqslant N$. Accordingly we ask for the probability of the event that $n^{-1}\,|S_n| \to 0$."

On the other hand, B. L. van der Waerden [26, p. 100] claims that the Strong Law of Large Numbers "hardly plays any role in mathematical statistics."

One thing I do know is that Claude Shannon's [23, 24] theory of information and coding relies only on the Weak Law. If you want to be well-educated in the 21$^{\text{st}}$ century, I strongly recommend you take Professor Michelle Effros's course **EE/Ma 126: Information Theory**.

7.17 ⋆ Convergence in probability vs. almost-sure convergence

Recall Definition 2.1.5.

2.1.5 Definition *A property is said to hold* ***almost surely****, abbreviated* ***a.s.****, if the set of outcomes for which it holds has probability one.*

We can rephrase the SLLN as $S_n/n \to \boldsymbol{E}\,X$ a.s..

More generally,

for any sequence $Y_1, Y_2, \dots$ on the common probability space $(\Omega, \mathcal{F}, P)$, we say that Y_n **converges almost surely** to Y, written $Y_n \to Y$a.s. if

$$P\{\omega \in \Omega : Y_n(\omega) \to Y(\omega)\} = 1.$$

I am including this discussion of the difference between convergence in probability and almost-sure convergence in theses notes, but **I do not plan to go over this material in class or examine you on it**. It is here for the benefit of those who want to understand the material more fully.

Let

$$\begin{aligned} A_n(\varepsilon) &= (\,|Y_n - Y| \leqslant \varepsilon) \\ &= \{\omega \in \Omega : |Y_n(\omega) - Y(\omega)| \leqslant \varepsilon\}. \end{aligned}$$

We can rewrite convergence in probability as

$$\begin{aligned}&\operatorname*{plim}_{n\to\infty} Y_n = Y\\ &\iff (\forall k\,)\ (\forall m\,)\ (\exists M\,)\ (\forall n \geqslant M\,)\ \big[\,P\big(A_n(1/k)\big) > 1-(1/m)\,\big]\end{aligned} \tag{15}$$

Now observe that the event

$$\begin{aligned}(Y_n \to Y) &= \{\omega \in \Omega : Y_n(\omega) \to Y(\omega)\}\\ &= \big\{\omega \in \Omega :\ (\forall \varepsilon > 0\,)\ (\exists N\,)\ (\forall n \geqslant N\,)\ [\,|Y_n(\omega) - Y(\omega)| < \varepsilon\,]\big\}\\ &= \big\{\omega \in \Omega :\ (\forall k\,)\ (\exists N\,)\ (\forall n \geqslant N\,)\ [\,|Y_n(\omega) - Y(\omega)| < 1/k\,]\big\}\end{aligned}$$

Now we use the cool trick of rewriting this as

$$= \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k).$$

In other words,

$$Y_n \to Y \text{a.s.} \iff P\Big(\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1. \tag{16}$$

The following argument uses the simple facts that for any countably additive probability measure P, if $P(\bigcap_j A_j) = 1$, then $P(A_j) = 1$ for all j; and if $P(\bigcup_{j=1}^{\infty} A_n) = 1$, then for each $\delta > 0$, there exists N such that $P(\bigcup_{j=1}^{N} A_j) > 1 - \delta$.

So observe that

$$\begin{aligned}&P\Big(\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1\\ &\implies (\forall k\,) \left[P\Big(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) = 1\right]\\ &\implies (\forall k\,)\ (\forall m\,)\ (\exists M\,) \left[P\Big(\bigcup_{N=1}^{M} \bigcap_{n=N}^{\infty} A_n(1/k)\Big) > 1-(1/m)\right]\end{aligned}$$

But $\bigcup_{N=1}^{M} \bigcap_{n=N}^{\infty} A_n(1/k) = \bigcap_{n=M}^{\infty} A_n$ (because letting $E_N = \bigcap_{n=N}^{\infty} A_n(1/k)$, we have $E_1 \subset E_2 \subset \cdots \subset E_M$), so

$$\begin{aligned}&\implies (\forall k\,)\ (\forall m\,)\ (\exists M\,) \left[P\Big(\bigcap_{n=M}^{\infty} A_n(1/k)\Big) > 1-(1/m)\right]\\ &\implies (\forall k\,)\ (\forall m\,)\ (\exists M\,)\ (\forall n \geqslant M\,)\ \big[\,P\big(A_n(1/k)\big) > 1-(1/m)\,\big]\end{aligned}$$

so by (15)

$$\implies \operatorname{plim} Y_n = Y.$$

So we have proven the following proposition.

7.17.1 Proposition

$$Y_n \xrightarrow{\text{a.s.}} Y \implies Y_n \xrightarrow{P} Y.$$

7.17.1 Converse Not True

Consider the sequence defined by

$$\begin{aligned} &Y_1 = \mathbf{1}_{[0,\frac{1}{2})},\ Y_2 = \mathbf{1}_{[\frac{1}{2},1)} \\ &Y_3 = \mathbf{1}_{[0,\frac{1}{4})},\ Y_4 = \mathbf{1}_{[\frac{1}{4},\frac{1}{2})},\ Y_5 = \mathbf{1}_{[\frac{1}{2},\frac{3}{4})},\ Y_6 = \mathbf{1}_{[\frac{3}{4},1)} \\ &Y_7 = \mathbf{1}_{[0,\frac{1}{8})},\ \text{etc.} \end{aligned}$$

Then $Y_n \xrightarrow{P} 0$, but $Y_n \not\xrightarrow{\text{a.s.}} 0$.

However we do have the following result.

7.17.2 Proposition *If $Y_n \xrightarrow{P} Y$, then for some subsequence, $Y_{n_k} \xrightarrow{\text{a.s.}} Y$.*

Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] S. Boucheron, G. Lugosi, and P. Massart. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford: Oxford University Press.

[3] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[4] G. Casella and R. L. Berger. 2002. *Statistical inference*, 2d. ed. Belmont, California: Brooks/Cole Cengage Learning.

[5] P. L. Chebyshev. 1867. Des valeurs moyennes. *Journal de Mathématiques Pures et Appliquées. Deuxième Série* 12:177–184. The author is given as P.-L. de Tchébychef. Translated from the Russian by N. de Khanikof.
https://gallica.bnf.fr/ark:/12148/bpt6k16411c/f185.image

[6] K. L. Chung. 1974. *A course in probability theory*, 2d. ed. Number 21 in Probability and Mathematical Statistics. Orlando, Florida: Academic Press.

[7] ——— . 1979. *Elementary probability theory with stochastic processes*. Undergraduate Texts in Mathematics. New York, Heidelberg, and Berlin: Springer–Verlag.

[8] R. Durret. 2019. *Probability: Theory and examples*, 5th. ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.
DOI: 10.1017/9781108591034

[9] N. Etemadi. 1981. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 55(1):119–122.
DOI: 10.1007/BF01013465

[10] W. Feller. 1971. *An introduction to probability theory and its applications*, 2d. ed., volume 2. New York: Wiley.

[11] D. Freedman and P. Diaconis. 1981. On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 57(4):453–476.
DOI: 10.1007/BF01025868

[12] C. F. Gauss. 1823. Theoria combinationis observationum erroribus minimis obnoxiae,pars prior. *Commentationes Societatis Regiae Scientiarum Gottingensis Recentiores* 5.

[13] I. Guttman, S. S. Wilks, and J. S. Hunter. 1971. *Introductory engineering statistics*, second ed. New York: John Wiley & Sons.

[14] W. Hoeffding. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30. http://www.jstor.org/stable/2282952

[15] J. Jacod and P. Protter. 2004. *Probability essentials*, 2d. ed. Berlin and Heidelberg: Springer.

[16] K. H. Knuth. 2013. Optimal data-based binning for histograms. arXiv:physics/0605197v2 [physics.data-an] http://arxiv.org/pdf/physics/0605197v2.pdf

[17] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[18] M. Loève. 1977. *Probability theory*, 4th. ed. Number 1 in Graduate Texts in Mathematics. Berlin: Springer–Verlag.

[19] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[20] F. Pukelsheim. 1994. The three sigma rule. *The American Statistician* 48(2):88–91. DOI: 10.1080/00031305.1994.10476030

[21] P. A. Samuelson. 1963. Risk and uncertainty: A fallacy of large numbers. *Scientia* 98:108–113. https://www.casact.org/pubs/forum/94sforum/94sf049.pdf

[22] D. W. Scott. 1979. On optimal and data-based histograms. *Biometrika* 66(3):605–610. http://www.jstor.org/stable/2335182

[23] C. E. Shannon. 1948. A mathematical theory of communication [Introduction, Parts I and II]. *Bell System Technical Journal* 27(3):379–423. This issue includes the Introduction, Part I: Discrete Noiseless Systems, and Part II: The Discrete Channel with Noise. Part III is in [24]. http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-3-379.pdf

[24] ——— . 1948. A mathematical theory of communication: Part III: Mathematical preliminaries. *Bell System Technical Journal* 27(4):623–656. http://www.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-4-623.pdf

[25] H. A. Sturges. 1926. The choice of a class interval. *Journal of the American Statistical Association* 21(153):65–66. http://www.jstor.org/stable/2965501

[26] B. L. van der Waerden. 1969. *Mathematical statistics*. Number 156 in Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete. New York, Berlin, and Heidelberg: Springer–Verlag. Translated by Virginia Thompson and Ellen Sherman from *Mathematische Statistik*, published by Springer-Verlag in 1965, as volume 87 in the series Grundlerhen der mathematischen Wissenschaften.

[27] D. F. Vysochanskiĭ and Y. I. Petunin. 1980. Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics* 21:25–36.

[28] ——— . 1983. A remark on the paper 'Justification of the 3σ rule for unimodal distributions'. *Theory of Probability and Mathematical Statistics* 27:27–29.

[29] M. P. Wand. 1997. Data-based choice of histogram binwidth. *The American Statistician* 51(1):59–64. DOI: 10.1080/00031305.1997.10473591

[30] L. Wasserman. 2010. *All of statistics: A concise course in statistical inference*. Springer Texts in Statistics. New York: Springer Science+Business Media.

[31] H. L. Westergaard. 1969. *Contributions to the history of statistics*. New York: A. M. Kelley. Reprint of the 1932 edition.