LECTURE 2: CONVERGENCE

Michal Kolesár*

September 5, 2024

1. USEFUL INEQUALITIES

REFERENCE CB, Chapter 3.6 and 4.7, or HMC, Chapter 1.10

We begin with proving two fundamental inequalities. The idea of using inequalities is essentially one of approximation. Sometimes it is too hard, or too cumbersome, to compute certain properties of random variables. Then it is useful to at least bound the property by relying on an inequality. This approximation idea plays a crucial role in the study of convergence concepts and their application, such as the derivation of large-sample properties of estimators and tests, something we will study shortly.

Theorem 1 (Markov's inequality). Let X be any nonnegative random variable such that $E[X]$ exists. Then for any $t > 0$, we have $P(X \geq t) \leq E[X]/t$.

Proof. By iterated expectations,

$$E[X] = E[X \mid X \geq t]P(X \geq t) + E[X \mid X < t]P(X < t) \geq tP(X \geq t) + 0. \qquad \square$$

From Markov's inequality we can derive Chebyshev's inequality.

Corollary 2 (Chebyshev's inequality). For any random variable X with mean μ, and any $t > 0$, we have

$$P(|X - \mu| \geq t) \leq \operatorname{var}(X)/t^2.$$

Proof. Note that $|X - \mu| \geq t$ if and only if $(X - \mu)^2 \geq t^2$. Thus, $P(|X - \mu| \geq t) = P((X - \mu)^2 \geq t^2)$. Since $(X - \mu)^2$ is a nonnegative random variable, $P((X - \mu)^2 \geq t^2) \leq E[(X - \mu)^2]/t^2 = \operatorname{var}(X)/t^2$ by Markov inequality. $\square$

Example. For all distributions with finite variance, $P(|X - \mu| \geq r\sigma) \leq 1/r^2$. So X falls outside two standard deviations of its mean with probability of at most $1/4$, and outside three standard deviation with probability at most $1/9$. If X is normal, then this bound seems rather crude, and indeed it is quite crude for most distributions. However, it turns out to still be incredibly useful for large-sample approximations. $\boxtimes$

*Email: mkolesar@princeton.edu.

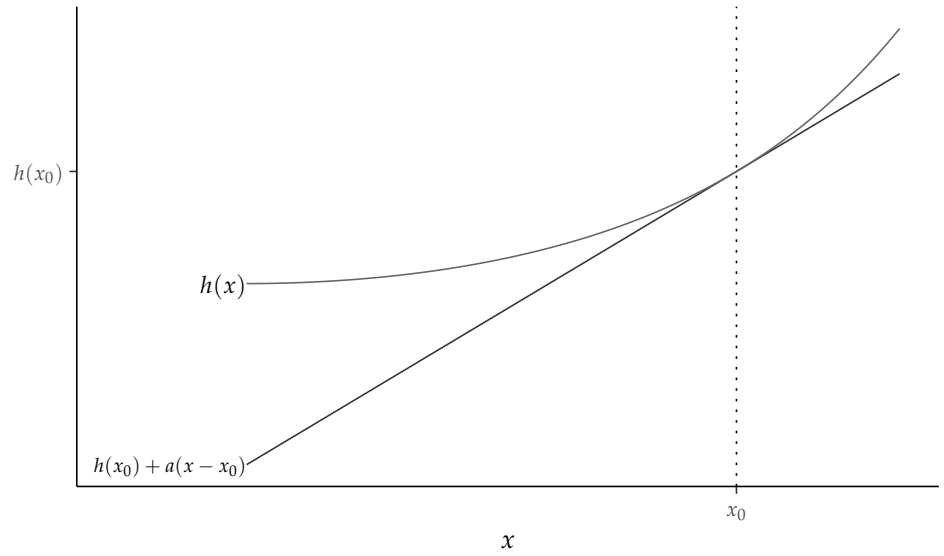

Figure 1: Graphical illustration of the proof of Jensen's inequality

Theorem 3 (Jensen's inequality, Lehmann and Casella 1998, Theorem 1.7.5). Let $h(x)$ be a convex function, and let X be a random variable such that $E[X]$ exists. Then $h(E[X]) \leq E[h(X)]$. If $h(x)$ is strictly convex, then the inequality is strict unless X is constant with probability one.

Proof. Recall a function is convex if for each x_0, there exists a line through $(x_0, h(x_0))$ such that $h(x)$ is never below this line. Call its slope a, so that the line is given by $y = h(x_0) + a(x - x_0)$ (see Figure 1). But then

$$h(x) \geq h(x_0) + a(x - x_0). \tag{1}$$

Since expectations preserve inequalities,

$$E[h(X)] \geq h(x_0) + a(E[X] - x_0).$$

This holds for all x_0, and in particular for $x_0 = E[X]$:

$$E[h(X)] \geq h(E[X]) + a(E[X] - E[X]) = h(E[X]).$$

If h is strictly convex, then the inequality (1) is strict for $x_0 \neq x$. Taking expectations preserves the strict inequality unless X is constant with probability one. □

Note that it may happen under the conditions of the theorem that $E[h(X)] = \infty$. If h is concave, then $-h$ is convex, so $E[h(X)] \leq h(E[X])$.

Example. If $E[X^4]$ exists, then so does $E[X^2]$. Also, if $X \geq 0$, then $1/E[X] < E[1/X]$, with the right-hand side (RHS) possibly infinite. ⊠

2. TYPES OF CONVERGENCE

REFERENCE CB, Chapter 5.5, HMC, Chapter 5.1–5.2

Let $S_1, \ldots, S_n, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $X_n = n^{-1} \sum_{i=1}^{n} S_i$ (say S_i is person i's shoe size). Then we may want to approximate the random variable X_n for n large (say $n = 1000$), by the "limit" of X_n, $n = 1, 2, \ldots$, provided such a limit exists—it is often more tractable than X_n itself—even if it never actually happens in real life that we do keep sampling ad infimum. So let us think about what "convergence to a limit X" of a sequence of random variables X_n could mean.

2.1. *Definitions*

Let $X_1, X_2, \ldots$ be a sequence of random variables on the probability space $(\Omega, \mathcal{A}, P)$. Then, for any realized outcome $\omega \in \Omega$, we have a sequence of real numbers, $X_1(\omega)$, $X_2(\omega), \ldots$

If X_n were just a sequence of (nonrandom) numbers, then we could define the limit in the usual way: $X_n \to X$ if $\lim_n X_n = X$. However, X_n are random, or, equivalently, we're dealing with having to define the limit of a sequence of functions.

Definition 4. The sequence $\{X_n\}_{n=1}^{\infty}$ *converges* to some random variable X *almost surely* if $P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$. In this case we write $X_n \to X$ a.s., or $X_n \overset{\text{a.s.}}{\to} X$.

This just generalizes the concept of pointwise convergence for functions: a sequence of functions $f_1, f_2, \ldots$ converges to the function f pointwise if $f_n(x) \to f(x)$ for all x. The above definition relaxes this a bit by allowing the convergence not to happen for a few ω's (that collectively have probability zero of happening).

Definition 5. We say that $\{X_n\}_{n=1}^{\infty}$ *converges* to X *in probability* if $P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) \to 0$ as $n \to \infty$ for any $\varepsilon > 0$. In this case we write $X_n \overset{p}{\to} X$, or $\operatorname{plim}_{n\to\infty} X_n = X$.

This converts the problem by defining the probability that X_n and X are more than ϵ apart, $p_n(\epsilon) = P(|X_n - X| > \epsilon)$. Since $p_n(\epsilon)$ is just a sequence of real numbers, we define $X_n \overset{p}{\to} X$ if $p_n(\epsilon) \to 0$ for all $\epsilon > 0$. That is, the probability that X_n and X are more than ϵ apart vanishes as $n \to \infty$.

Definition 6. We say that $\{X_n\}_{n=1}^{\infty}$ *converges* to X *in quadratic mean* (or in the second moment) if $E[(X_n - X)^2] \to 0$ as $n \to \infty$. In this case we write $X_n \to_{L_2} X$.

It's called convergence in L_2 because it applies the concept of L_2 convergence of functions: a sequence of functions $f_1, f_2, \ldots$ converges to f in L_2 if $\int (f_n(x) - f(x))^2 \, \mathrm{d}x \to 0$ (the quantity $\|f\|_2 = \sqrt{\int f^2}$ is called the L_2 norm of a function, it generalizes the concept of Euclidean distance, of Euclidean norm, to functions). Again, this concept defines convergence by turning the problem into a statement about a sequence of numbers $\Delta_n = E[|X_n - X|^2]$.

Definition 7. We say that $\{X_n\}_{n=1}^{\infty}$ *converges* to X *in distribution* if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}$ where $F_X(x)$ is continuous. In this case we write $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$. Sometimes this is called convergence in law, or weak convergence.

Why the restriction to the continuity points? Let $F(x) = \mathbb{1}\{x \geq 0\}$, the cumulative distribution function (CDF) of a random variable that is zero with probability one. Let $F_n(x) = \mathbb{1}\{x \geq 1/n\}$, the CDF of a random variable that is equal to $1/n$ with probability one. It really makes sense to say that the distribution F_n converges to F even if $F_n(0) \to 0 \neq 1 = F(0)$. More generally, let $X_n = X + 1/n$, and let $X \sim F$. Then $F_n(x) = F(x - 1/n) \to F(x_-) = \lim_{y\uparrow x} F(y)$, so convergence occurs only at continuity points.

Digression. Another worry is that if we only restrict attention to continuity points, this may not be enough to identify the limit $F(x)$. However, since F is increasing and right-continuous, one can show that it has at most countably many discontinuity points.

2.2. *Relations between different types of convergence*

We can use the Markov inequality to prove that convergence in quadratic mean implies convergence in probability:

Theorem 8. If $X_n \to_{L_2} X$, then $X_n \to_p X$.

Proof. By Markov inequality,

$$P(|X_n - X| > \varepsilon) = P(|X_n - X|^2 > \varepsilon^2) \leq E[|X_n - X|^2]/\varepsilon^2 \to 0.$$

for any $\varepsilon > 0$. □

Convergence in probability implies convergence in distribution:

Theorem 9. If $X_n \to_p X$, then $X_n \Rightarrow X$.

Proof. Note that $X_n \leq x$ and $X > x + \varepsilon$ implies $|X_n - X| > \varepsilon$. Thus,

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) \\ &= P(X_n \leq x,\, X \leq x + \varepsilon) + P(X_n \leq x,\, X > x + \varepsilon) \\ &\leq P(X \leq x + \varepsilon) + P(|X_n - X| > \varepsilon) \\ &= F_X(x + \varepsilon) + P(|X_n - X| > \varepsilon). \end{aligned}$$

for any $x \in \mathbb{R}$ and $\varepsilon > 0$. Similarly,

$$F_X(x - \varepsilon) \leq F_{X_n}(x) + P(|X_n - X| > \varepsilon).$$

Combining the two displays above yields

$$F_X(x - \varepsilon) - P(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x + \varepsilon) + P(|X_n - X| > \varepsilon). \tag{2}$$

Next, if x is a point of continuity of F_X, for any $\delta > 0$, there exists $\varepsilon > 0$ such that

$$F_X(x+\varepsilon) - \delta \leq F_X(x) \leq F_X(x-\varepsilon) + \delta. \tag{3}$$

Combining Equations (2) and (3) yields

$$F_X(x) - \delta - P(|X_n - X| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x) + \delta + P(|X_n - X| > \varepsilon),$$

which we can write as

$$|F_X(x) - F_{X_n}(x)| \leq \delta + P(|X_n - X| > \varepsilon).$$

Now, by definition of convergence in probability, $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$, so that

$$\lim_{n\to\infty} |F_{X_n}(x) - F_X(x)| \leq \delta.$$

Since this holds for any δ and any $x \in \mathbb{R}$ where $F_X(x)$ is continuous, the result follows. □

As a homework exercise, you'll be asked to prove the following two theorems

Theorem 10. If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{p} X$.

As a step towards proving this result, you'll show that almost sure convergence is equivalent to saying that the entire tail $\{X_k\}_{k\geq n}$ of a sequence of random variables is close to its limit *at once* in that it's equivalent to $\lim_{n\to\infty} P(\sup_{k\geq n}|X_k - X| > \epsilon) = 0$. On the other hand, convergence in probability controls how close a single random variable X_n is to its limit, in the sense that $\lim_{n\to\infty} P(|X_n - X| > \epsilon) \to 0$.

Theorem 11. If c is some constant and $X_n \Rightarrow c$, then $X_n \to_p c$.

Theorems 8–11 give all correct implications. Any other implication is, in general, incorrect. As an exercise, you can think of counterexamples. One such counterexample:

Example (Convergence in L_2 does not follow from convergence in probability). Let $\Omega = [0,1]$ be a sample space. Let $X_n(\omega) = n$ if $\omega \in [0, 1/n]$ and 0 otherwise, and suppose $P(\omega \leq b) = b$ for $0 \leq b \leq 1$. Then it is easy to check that $X_n \to_p 0$ (and in fact that $X_n \xrightarrow{\text{a.s.}} 0$). On the other hand, $E[|X_n - 0|^2] = E[X_n^2] = n^2 \cdot (1/n) = n \to \infty$. ⊠

Digression. Why bother with all these different convergence concepts? In practice, we mostly just use convergence in probability and in distribution. Almost sure convergence is useful for some technical arguments when we are proving theorems (notably, it's useful whenever we try to prove convergence of one random variable conditional on the realization of another random variable). Convergence in L_2 is useful for two reasons: as a way of proving convergence in probability (Theorem 8), and to remind ourselves that convergence in probability or in distribution does not imply convergence of moments! Later in the course, you'll see plenty of estimators that have this property: as $n \to \infty$, they converge to the normal distribution (which has all moments), but for any given n, the estimator only has a few (perhaps no) moments: one needs the concept of uniform integrability to ensure convergence of moments, that is $\lim_{M\to\infty} \limsup_{n\to\infty} E[|X_n| \mathbb{1}\{|X_n| \geq M\}] = 0$. If this holds, then $X_n \Rightarrow X$ implies $E[X_n] \to E[X]$. The uniform integrability condition ensures that we don't have the issue of escaping mass. A sufficient condition is that $E[|X_n|^{1+\delta}] \leq C$ for all n. ⊠

3. SLUTSKY'S THEOREM. CONTINUOUS MAPPING THEOREM.

Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables, and let X, Y also be random variables. Let g be some continuous function. Let c be some constant. Then:

1. If $X_n \to_p X$ and $Y_n \to_p Y$, then $X_n + Y_n \to_p X + Y$ and $X_n Y_n \to_p XY$.
2. If $X_n \Rightarrow X$ and $Y \to_p c$, then $X_n + Y_n \Rightarrow X + c$ and $X_n Y_n \Rightarrow cX$.
3. If $X_n \to_p X$, then $g(X_n) \to_p g(X)$.
4. If $X_n \Rightarrow X$, then $g(X_n) \Rightarrow g(X)$

The first and second statements are known as the Eugen Slutsky's theorem. The third and forth statements are known as the Continuous mapping theorem (or sometimes Mann-Wald after Mann and Wald (1943)). These theorems are widely used in statistics.

Remark 12. Note that, in general, $X_n \Rightarrow X$ and $Y_n \Rightarrow Y$ does not imply $X_n + Y_n \Rightarrow X + Y$ or $X_n Y_n \Rightarrow XY$.

Remark 13. We can generalize the concept of convergence in probability and in distribution to random vectors in $\mathbb{R}^k$. Let X_n be a sequence of random vectors with joint CDF F_n, and let X be random vector with CDF F. Then

1. X_n converges to X in probability if for all $\epsilon > 0$, $P(\|X_n - X\| > \epsilon) \to 0$. Here $\|a\| = \sqrt{\sum_{i=1}^k a_i^2}$ denotes the Euclidean norm of a vector. Note that according to this definition, $X_n \xrightarrow{p} X$ as a vector if and only if it's true for all components $j = 1, \ldots, k$ of X_n that $X_{n,j} \xrightarrow{p} X_j$.
2. We say that X_n converges to X in distribution if $F_n(x) \to F(x)$ at all continuity points x of F (this is the same definition as before, except now x is a vector, so $F\colon \mathbb{R}^k \to [0,1]$).

With these definitions, the continuous mapping theorem can be stated as $X_n \Rightarrow X \implies g(X_n) \Rightarrow g(X)$ and $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$ for all continuous functions $g\colon \mathbb{R}^k \to \mathbb{R}^p$ (van der Vaart 1998, Theorem 2.3). Since $g(x,y) = x + y$ and $g(x,y) = xy$ are continuous, this vector version of the continuous mapping theorem directly implies part 1 of Slutsky's theorem.

The relationship between the convergence of marginal distributions and convergence of the joint distribution is given by a theorem known as the Cramér-Wold device, which says that $X_n \Rightarrow X$ if and only if $a'X_n \Rightarrow a'X$ for all $a \in \mathbb{R}^k$ (Billingsley 1995, Theorem 29.4 or Durrett 2019, Theorem 3.9.5). So in general, it's not true that if $X_{1n} \Rightarrow X_1$ and $X_{2n} \Rightarrow X_2$, then $(X_{1n}, X_{2n}) \Rightarrow (X_1, X_2)$—for example, let $X_{1n} = -X_{2n} \sim \mathcal{N}(0,1)$, and let $X_1 = X_2 \sim \mathcal{N}(0,1)$. However, it is true that if $X_n \Rightarrow X$ and $Y_n \Rightarrow c$, then $(X_n, Y_n) \Rightarrow (X, c)$ (van der Vaart 1998, Theorem 2.7(v)). Combining the last statement with the vector version of the continuous mapping theorem with $g(x,y) = x + y$ and $g(x,y) = xy$ then yields the part 2 of Slutsky's theorem.

4. LAW OF LARGE NUMBERS

Theorem 14 (Durrett 2019, Theorem 2.2.3). If $\{X_i\}_{i=1}^{\infty}$ is a sequence of uncorrelated random variables with common mean $E[X_i] = \mu$ and bounded variances $\text{var}(X_i) \leq C < \infty$, then $\overline{X}_n := \sum_{i=1}^{n} X_i/n \to_{L_2} \mu$ and, thus, in probability.

Proof. By linearity of expectation, $E[\overline{X}_n] = E[\sum_{i=1}^{n} X_i/n] = \sum_{i=1}^{n} E[X_i]/n = \mu$. The result then follows since

$$E(\overline{X}_n - \mu)^2 = \text{var}(\overline{X}_n) = \frac{1}{n^2}\,\text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) \leq C/n \to 0. \qquad \square$$

Another version of the law of large numbers that doesn't require existence of second moments is

Theorem 15 (Strong law of large numbers). If $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables with $E[|X_i|] < \infty$, then $\overline{X}_n - E[X_i] \xrightarrow{a.s.} 0$.

Proof. See Theorem 2.4.1 in Durrett (2019) if you're interested. $\square$

Digression. If $\{X_i\}_{i=1}^{\infty}$ are independent with finite means μ_i, then $\overline{X}_n - \overline{\mu}_n \xrightarrow{a.s.} 0$, provided $E[|X_i|^{1+\delta}] < \infty$ for some $\delta > 0$ (see, e.g., Corollary 3.9 in White 2001). So we need a little more than the first moment if the data are independent, but not identically distributed. $\boxtimes$

Example. If $X_1, X_2, \ldots$ are i.i.d. $\chi^2(1)$, then for $n = 100$, $P(|\overline{X}_n - 1| > 0.1) = 0.479$, and with $n = 1000$, $P(|\overline{X}_n - 1| > 0.1) = 0.02533$. Depending on the shape of the distribution of X_n, it may take a short or a long time for the law to "kick in", but it always does eventually. $\boxtimes$

5. CENTRAL LIMIT THEOREM

REFERENCE CB, Chapter 5.5, HMC, Chapter 5.3.

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then $E[\sum_{i=1}^{n}(X_i - \mu)/\sqrt{n}] = 0$ and $\text{var}(\sum_{i=1}^{n}(X_i - \mu)/\sqrt{n}) = \sigma^2$. A much more remarkable result is the central limit theorem:

Theorem 16 (Central limit theorem). If $X_1, X_2, \ldots$ are i.i.d. with finite variance σ^2, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - E[X_i]) \Rightarrow \mathcal{N}(0, \sigma^2).$$

Digression. The first version of the central limit theorem is due to DeMoivre and Laplace, and goes back to 1738. It states that if X_n is Binomial with parameters n and p, then $(X_n - np)/\sqrt{np(1-p)} \Rightarrow \mathcal{N}(0,1)$ as $n \to \infty$. One can prove this directly using Stirling's formula. $\boxtimes$

The central limit theorem (CLT) goes a step further than a law of large numbers by identifying the limiting distribution. This is a remarkable strengthening and tells you the importance of the variance assumption.

Proof of Theorem 16. The standard proof uses the characteristic function (the characteristic function of a random variable X is $\varphi(t) = E[e^{itX}]$; unlike the moment generating function, it always exists), and is quite short (see, for example, Durrett 2019, Theorem 3.4.1). The textbook proofs (Hogg, McKean, and Craig (2013, Theorem 5.3.1) and Casella and Berger (2002, Theorem 5.5.14)) assume the much stronger condition that the moment generating function (MGF) exists, but use the same logic. However, these proof doesn't give much insight into why the result holds. □

Digression. To gain some insight into why it's the normal distribution that the sample average converges to, I'll give a proof that uses Lindeberg's swapping trick, due to Lindeberg (1922). First observe that the result holds in finite samples if X_i are $\mathcal{N}(\mu, \sigma^2)$. So the theorem certainly holds in that case. Our task is to show that the result is *universal* in the sense that it holds no matter what the distribution of X_i is.

We will use an alternative characterization of convergence in distribution: $Z_n \Rightarrow Z$ iff[1] $E[g(Z_n)] \to E[g(Z)]$ for all g that are bounded and continuous (see, e.g., van der Vaart 1998, Lemma 2.2). The $\implies$ direction follows from the continuous mapping theorem (CMT) and the fact that because g is bounded, we can't have probability masses escaping to infinity. The $\impliedby$ direction follows because $Z_n \Rightarrow Z$ means that for all continuity points x of Z, $E[\mathbb{1}\{Z_n \le x\}] \to E[\mathbb{1}\{Z \le x\}]$, and because the indicator function can be approximated arbitrarily well by a sequence of continuous functions. In fact, because it suffices to look at continuous bounded functions with three bounded derivatives to approximate the indicator function, it suffices to check that $E[g(Z_n)] \to E[g(Z)]$ for g that are bounded with three bounded derivatives. Furthermore, because we can standardize $X_i \mapsto (X_i - E[X_i])/\sigma$, it suffices to check the theorem for X_i with mean zero and variance one. So our task is to show that for X_i with $E[X_i] = 0$, $\text{var}(X_i) = 1$,

$$E[g(n^{-1/2}\textstyle\sum_i X_i)] \to E[g(Z)], \quad Z \sim \mathcal{N}(0,1),$$

for all g that are bounded with three bounded derivatives. We already know that $E[g(Z_n)] = E[g(Z)]$ where $Z_n = n^{-1/2}\sum_i Y_i$, with $Y_i \sim \mathcal{N}(0,1)$, so it suffices to show that $E[g(n^{-1/2}\sum_i X_i)] - E[g(Z_n)] \to 0$. We'll show this under the additional simplifying assumption that $E[|X_i|^3] < \infty$ (without this assumption, the proof gets more complicated). We will do this by replacing each X_i in the sum $g(n^{-1/2}\sum_i X_i)$ with Y_i. If we replace X_n with Y_n, the expectation changes by

$$E[g(S_n + X_n/\sqrt{n})] - E[g(S_n + Y_n/\sqrt{n})], \qquad S_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_{n-1}).$$

Taking a third-order Taylor expansion of both terms around S_n, we get

$$g(S_n + X_n/\sqrt{n}) - g(S_n + Y_n/\sqrt{n}) = \frac{1}{\sqrt{n}}(X_n - Y_n)g'(S_n) + \frac{1}{2n}(X_n^2 - Y_n^2)g''(S_n) + \frac{1}{6n^{3/2}}(X_n^3 g^{(3)}(\tilde{X}_n) - Y_n^3 g^{(3)}(\tilde{Y}_n)).$$

where $\tilde{Y}_n$ is an intermediate point between Y_n and S_n, and $\tilde{X}_n$ is an intermediate point between X_n and S_n. Taking expectations, we get

$$E[g(S_n + X_n/\sqrt{n}) - g(S_n + Y_n/\sqrt{n})] = \frac{1}{6n^{3/2}}E[X_n^3 g^{(3)}(\tilde{X}_n) - Y_n^3 g^{(3)}(\tilde{Y}_n)],$$

This is the key point in the derivation: since S_n is independent of X_n and of Y_n, $E[(X_n - Y_n)g'(S_n)] = E[X_n - Y_n]E[g'(S_n)] = 0 \cdot E[g'(S_n)] = 0$, and similarly $E[(X_n^2 - Y_n^2)g''(S_n)] = E[(X_n^2 - Y_n^2)]E[g''(S_n)] = 0 \cdot E[g''(S_n)] = 0$.

Continuing to swap X_{n-1} with Y_{n-1}, X_{n-2} with Y_{n-2}, etc., and summing the telescoping series $E[g(X_1 + \cdots + X_i + Y_{i+1} + \cdots Y_n)] - E[g(X_1 + \cdots + X_{i-1} + Y_i + Y_{i+1} + \cdots Y_n)]$, we get

$$|E[g(S_n)] - E[g(Z_n)]| = \left|\frac{1}{6n^{3/2}}\sum_{i=1}^{n} E[X_i^3 g^{(3)}(\tilde{X}_i) - Y_i^3 g^{(3)}(\tilde{Y}_i)]\right| \le \frac{\sup_x |g^{(3)}(x)|}{6n^{1/2}} E[|X_i|^3 + 2^{3/2}\pi^{-1/2}] \to 0,$$

where the inequality follows by applying the inequality $E|X_i^3 g^{(3)}(\tilde{X}_i) - Y_i^3 g^{(3)}(\tilde{Y}_i)| \le \sup_x |g^{(3)}(x)| \cdot E[|X_n|^3 + 2^{3/2}\pi^{-1/2}]$. Here we use the fact that the third central moment of a standard normal is given by $E[|Y_i|^3] =$

1. iff with the extra "f" is an abbreviation of "if and only if"

$2^{3/2}/\sqrt{\pi}$. This completes the proof.

You can see from the logic of this proof that the reason that sample averages converge to a normal distribution is that the normal distribution is stable (average of two normal random variables is normal). In fact, and this should now not be surprising, it is the only distribution with finite variance that's stable (for instance the Cauchy distribution is also stable, but it doesn't have any moments). ⊠

Multivariate case:

Theorem 17. If $X_1, X_2, \ldots$ are i.i.d. random vectors with finite variance $\text{var}(X_i) = E[(X_i - E[X_i])(X_i - E[X_i])']$, *then* $\sum_{i=1}^n (X_i - E[X_i])/\sqrt{n} \Rightarrow \mathcal{N}(0, \text{var}(X_i))$.

Example. Let the probability of a newborn being a boy be, say, 0.51. What is the probability that at least a half of n newborns will be boys? To answer this question, let $X_i = 1$ if ith newborn is a boy and $X_i = 0$ otherwise. Then $X_i = 1$ with probability $p = 0.51$ and $X_i = 0$ with probability $1 - p = 0.49$. Therefore, $\mu = E[X_i] = 0.51$ and $\sigma^2 = p(1-p) = 0.51 \cdot 0.49$. Moreover, $X_1, \ldots, X_n$ are independent random variables. The total number of boys equals $\sum_{i=1}^n X_i$. Thus,

$$\begin{aligned} P\left(\sum_{i=1}^n X_i \geq n/2\right) &= P\left(\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{p(1-p)}} \geq \frac{\sqrt{n}(1/2-p)}{\sqrt{p(1-p)}}\right) \\ &\approx 1 - \Phi\left(\frac{\sqrt{n}(1/2-p)}{\sqrt{p(1-p)}}\right) \\ &= 1 - \Phi\left(-\frac{0.01\sqrt{n}}{\sqrt{0.51 \cdot 0.49}}\right), \end{aligned}$$

since $\sqrt{n}(\overline{X}_n - p) \Rightarrow \mathcal{N}(0, p(1-p))$ as $n \to \infty$. Note that here we used the fact that $\Phi(x)$ is continuous at all $x \in \mathbb{R}$.

Evaluating the right-hand side gives $1 - 0.456$, $1 - 0.421$, and $1 - 0.263$ at $n = 30, 100$, and 1000. By way of comparison, a direct calculation gives $1 - 0.385, 1 - 0.382$, and $1 - 0.253$. ⊠

6. DELTA METHOD

REFERENCE CB, Chapter 5.5.4, HMC, Chapter 5.2.2.

Suppose that $\sqrt{n}(X_n - a) \Rightarrow \mathcal{N}(0, \sigma^2)$. For example X_n could be the sample mean. The next result is incredibly useful if we're interested in a function of X_n.

Theorem 18 (Delta Method). Let $X_1, X_2, \ldots$ be a sequence of random variables that satisfies $\sqrt{n}(X_n - a) \Rightarrow \mathcal{N}(0, \sigma^2)$, and let $g\colon \mathbb{R} \to \mathbb{R}$ be continuously differentiable at a. Then

$$\sqrt{n}(g(X_n) - g(a)) \Rightarrow \mathcal{N}(0, g'(a)^2\sigma^2).$$

Proof. The mean value theorem implies that $g(X_n) - g(a) = (X_n - a)g'(\tilde{X}_n)$, with $\tilde{X}_n(\omega) \in [a, X_n(\omega)]$. Thus, $X_n \xrightarrow{p} a$ implies $\tilde{X}_n \xrightarrow{p} a$, and by the continuous mapping theorem, $g'(\tilde{X}_n) \xrightarrow{p} g'(a)$. The result follows by Slutsky theorem. □

Digression. In fact, it suffices that g is differentiable, and works even if the limiting distribution is not normal; see van der Vaart (1998, Theorem 3.1). ⊠

Note that this theorem also holds when $g'(a) = 0$ but in this case the asymptotic distribution will be 0 (constant), i.e. degenerate, which is not a particularly useful approximation.

The Delta method has a multidimensional extension that can be proved using similar logic:

Theorem 19. Let $X_1, X_2, \ldots$ be a sequence of random vectors, and let $g\colon \mathbb{R}^k \to \mathbb{R}^p$ be continuously differentiable at a, and let G denote the matrix with elements $G_{ij} = \partial g_i(a)/\partial x_j$. Suppose $\sqrt{n}(X_n - a) \Rightarrow \mathcal{N}(0, \Sigma)$. Then $\sqrt{n}(g(X_n) - g(a)) \Rightarrow \mathcal{N}(G\Sigma G')$.

Example 1. Let $X_1, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. What is the limiting distribution of $(\overline{X}_n)^2$? Let $g(x) = x^2$. Then $g'(\mu) = 2\mu$. Thus, by the Delta method, $\sqrt{n}(\overline{X}_n^2 - \mu^2) \Rightarrow \mathcal{N}(0, 4\mu^2\sigma^2)$. Note that if $\mu = 0$, then the limit distribution is degenerate. A more useful approximation is via the continuous mapping theorem: $n\overline{X}_n^2 \Rightarrow \sigma^2 \cdot \chi^2(1)$.

Let's put some numbers on this. Suppose $X_i \sim \mathcal{N}(\mu, 1)$ i.i.d. Then $\overline{X}_n \sim \mathcal{N}(\mu, 1/n)$, or $\sqrt{n}(\overline{X}_n - \mu) \sim \mathcal{N}(0, 1)$. The delta method says that $\sqrt{n}(\overline{X}_n^2 - \mu^2) \Rightarrow \mathcal{N}(0, 4\mu^2)$. If we're interested in $P(\overline{X}_n^2 \leq z)$, then an exact calculation yields

$$\begin{aligned} P(\overline{X}_n^2 \leq z) = P(-\sqrt{z} \leq \mu + \mathcal{N}(0,1)/\sqrt{n} \leq \sqrt{z}) \\ = \Phi(\sqrt{n}(\sqrt{z} - \mu)) - \Phi(-\sqrt{n}(\sqrt{z} + \mu)), \end{aligned}$$

while the delta method gives

$$P(\overline{X}_n^2 \leq z) = P(\sqrt{n}(\overline{X}_n^2 - \mu^2) \leq \sqrt{n}(z - \mu^2)) \approx \Phi(\sqrt{n}(z - \mu^2)/2\mu).$$

Figure 2 plots the approximation error as well as the exact probabilities for different values of μ and z and $n = 100$ and $n = 400$. What do you notice? ⊠

7. HOW TO THINK ABOUT CONVERGENCE RESULTS

The aim of considering the limit of a sequence of random variables is to approximate properties of an element in the sequence with n large, but finite. Large sample results (or "asymptotic theory") is a framework to systematically generate such approximations. The $n \to \infty$ is a thought experiment that tells us what would happen if we had an infinitely large sample—but we never do, so the only interesting thing about this thought experiment is that it often provides accurate approximations for small samples that we actually observe.

Sometimes asymptotic theory is criticized on the grounds that the thought experiment is inconceivable: one cannot literally take the number of US states, the number of

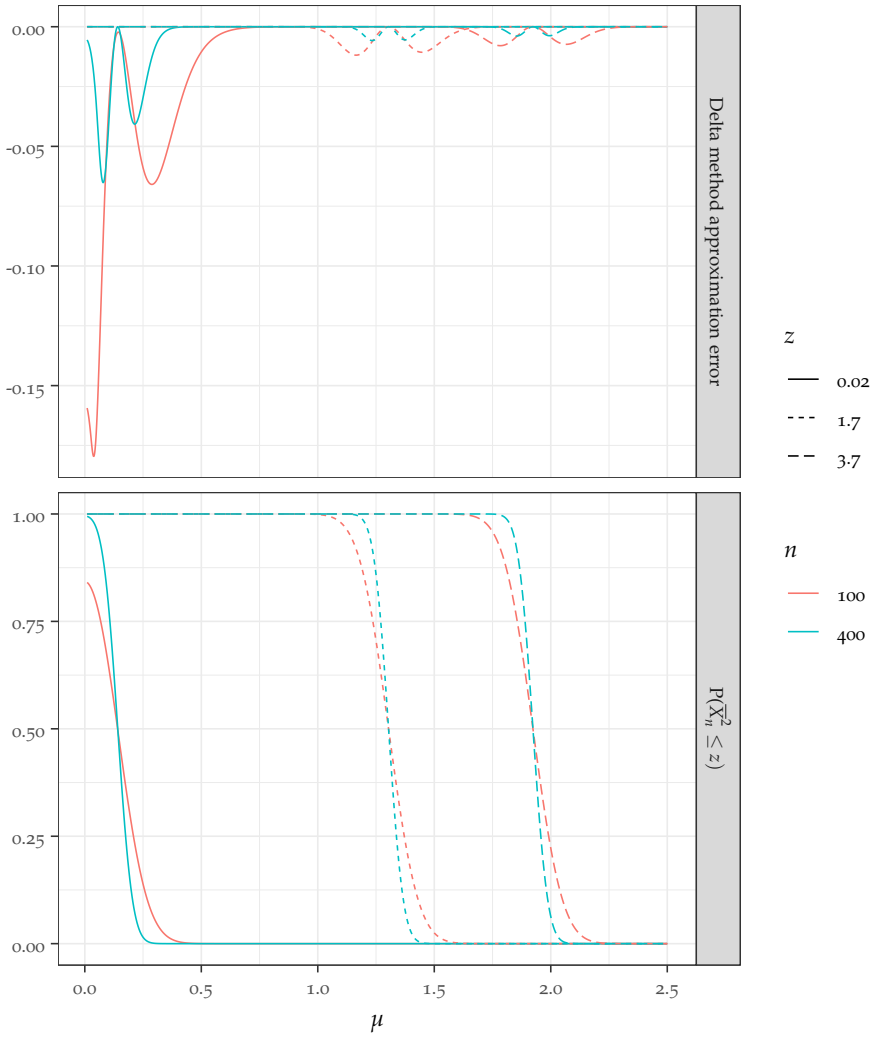

Figure 2: Exact probabilities $P(\overline{X}_n \leq z)$ from Example 1 as well as approximation error of delta method for different values of z, μ, and $n = 100$ and $n = 400$.

countries in the world, or many other samples to infinity, so someone may argue that it makes no sense to let n be larger. But this argument misses the point. We are not interested in making statements about a counterfactual world with many more countries. Rather, we're interested in a (hopefully accurate) approximation to what's going on in the sample at hand, with a particular, finite n. We do this by embedding the statistical analysis in a sequence of ever-increasing n's because it's an accurate approximation in most cases.

The assumptions necessary to generate law of large numberss (LLNs) and CLTs are remarkably weak. In a set-up with i.i.d. sampling, we have made essentially no assumptions on the distribution of X_i besides assuming its mean and variance exist. This is exciting, since it suggests that one can do asymptotically justified inference about μ under rather weak assumptions. Contrast this to an often close to impossible approach of spelling out exactly what kind of distributions X_i is allowed to obey, and given those distributions, work out the small sample properties of $\overline{X}_n$.

The somewhat hidden cost of doing this is that LLN and CLT don't give us error bounds on the accuracy of the approximation. In fact, there exist impossibility results which show that, for a given n, one can always find a distribution for X_i with a finite variance (and which therefore satisfies the CLT assumptions) that makes the normal approximation arbitrarily bad. Thus, implicitly, when we're doing asymptotics, we're ruling those distributions out. In other words, "asymptotically justified inference" replaces concrete assumptions on X_i with the much vaguer but crucial assumption that for the given n that we actually face, the asymptotics provide a reasonable approximation (sometimes called "the asymptotics kick in").

As an extreme example for how asymptotics can be misleading, consider an example from Davidson (1994, p. 181). A monkey is hitting a keyboard at random. Then the probability that the monkey eventually types out the text of the US Constitution is 1; yet for any reasonable fixed sample size, the probability is near zero—these asymptotics don't kick in for reasonable sample sizes.

To get a rough sense of when the asymptotics kick in, many papers run small sample simulations for some "nice" distributions and compare the resulting small sample properties of the suggested procedure with the asymptotic approximation. If they are close, then we know that at least for some "nice" data generating process, the asymptotics do kick in.

The other possibility is to do some extra work to get the error bounds: you can look up Berry-Eseen, Edgeworth expansions, or tail bounds for sub-Gaussian random variables if you are interested in learning more about these types of results.

REFERENCES

Billingsley, Patrick. 1995. *Probability and Measure.* 3rd ed. New York, NY: John Wiley & Sons.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury/Thomson Learning.

Davidson, James. 1994. *Stochastic Limit Theory.* Oxford, UK: Oxford University Press. https://doi.org/10.1093/0198774036.001.0001.

Durrett, Rick. 2019. *Probability: Theory and Examples.* 5th ed. New York, NY: Cambridge University Press. https://doi.org/doi.org/10.1017/9781108591034.

Hogg, Robert V., Joseph W. McKean, and Allen T. Craig. 2013. *Introduction to Mathematical Statistics.* 7th ed. New York, NY: Pearson.

Lehmann, Erich L., and George Casella. 1998. *Theory of Point Estimation.* 2nd ed. New York, NY: Springer. https://doi.org/doi.org/10.1007/b98854.

Lindeberg, Jarl W. 1922. "Eine Neue Herleitung Des Exponentialgesetzes in Der Wahrscheinlichkeitsrechnung." *Mathematische Zeitschrift* 15, no. 1 (December): 211–225. https://doi.org/10.1007/BF01494395.

Mann, Henry B., and Abraham Wald. 1943. "On Stochastic Limit and Order Relationships." *Annals of Mathematical Statistics* 14, no. 3 (September): 217–226. https://doi.org/10.1214/aoms/1177731415.

Van der Vaart, Aad. 1998. *Asymptotic Statistics.* New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511802256.

White, Halbert. 2001. *Asymptotic Theory for Econometricians.* Revised edition. San Diego, CA: Academic Press.