The Central Limit Theorem 7

7.A Appendix: Proof of the Central Limit Theorem

The central limit theorem.

Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables with mean $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$. Let $S_n = X_1 + \ldots + X_n$ be the sum of these random variables, and let $Z \sim N(0, 1)$ be a standard normal random variable. Then

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z \text{ as } n \to \infty.$$

In words, the central limit theorem says that if the random variables $\{X_i\}_{i=1}^n$ are i.i.d., then when the sample size n is large enough, the distribution of their sum S_n is approximately normal with the appropriate mean and variance. What is remarkable is that this limit result does not depend on the distribution of the underlying trials.[1] Our aim here is to prove this fundamental result.

7.A.1 Moment generating functions

The first step in proving the central limit theorem is to introduce a new and unexpected way of representing the distribution of a random variable. Its key advantage is the ease with which it allows us to work with sums of independent random variables.

Definition.

The **moment generating function** of the random variable X is a function M_X defined on some interval $(-\varepsilon, \varepsilon)$ by

$$M_X(t) = \mathrm{E}(e^{tX}). \tag{7.A.1}$$

For the moment generating function of a random variable to exist, the expected value in (7.A.1) must be finite for all t in an interval around 0. Moment generating functions fail to exist for random variables that put too much probability

[1] This is not to say that the distribution does not matter at all. As we discussed in Chapter 7, if the distribution of the trials is far from normal, then a larger number of trials is needed before the normal approximation becomes accurate.

on values approaching infinity or negative infinity (see property (ii) below). Because of this, our proof of the central limit theorem using moment generating functions will not cover all cases in which the theorem is true; but as we explain at the end of the section, this limitation can be remedied.[2]

Example Suppose that X has a *Bernoulli*(p) distribution: $\mathrm{P}(X=1)=p$ and $\mathrm{P}(X=0)=1-p$. Then the moment generating function of X is defined for all $t\in(-\infty,\infty)$

$$M_X(t)=\mathrm{E}(\mathrm{e}^{tX})=\sum_{x\in\{0,1\}}\mathrm{e}^{tx}\,\mathrm{P}(X=x)=(1-p)+p\mathrm{e}^t.$$

Example Suppose that T has an *exponential*(λ) distribution, so that its density function is $f(t)=\lambda\mathrm{e}^{-\lambda t}$ for $t\geq 0$ and $f(t)=0$ otherwise. Then the moment generating function of T is defined for $t\in(-\infty,\lambda)$ by

$$M_T(t)=\mathrm{E}(\mathrm{e}^{tX})=\int_0^\infty \mathrm{e}^{ts}\lambda\mathrm{e}^{-\lambda s}\,\mathrm{d}s=\int_0^\infty \lambda\mathrm{e}^{(t-\lambda)s}\,\mathrm{d}s=\frac{\lambda}{t-\lambda}\mathrm{e}^{(t-\lambda)s}\Big|_{s=0}^{\infty}=\frac{\lambda}{\lambda-t}.$$

Example Suppose that Z has a standard normal distribution, so that its density function is $f(z)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}$. Then by completing the square, we find that the moment generating function of Z is defined for all $t\in(-\infty,\infty)$ by

$$\begin{aligned}
M_Z(t)&=\mathrm{E}(\mathrm{e}^{tZ})\\
&=\int_{-\infty}^{\infty}\mathrm{e}^{tz}\cdot\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-z^2/2}\,\mathrm{d}z\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(z^2-2tz)/2}\,\mathrm{d}z\\
&=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(z^2-2tz+t^2)/2}\mathrm{e}^{t^2/2}\,\mathrm{d}z\\
&=\mathrm{e}^{t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(z-t)^2/2}\,\mathrm{d}z\\
&=\mathrm{e}^{t^2/2}, \qquad (7.A.2)
\end{aligned}$$

where the final equality is true because $\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(z-t)^2/2}$ is the density of a random variable with a $N(t,1)$ distribution.

[2] Away from probability theory, the moment generating function is known as the (*bilateral*) *Laplace transform*.

We now state five important properties of moment generating functions. Intuitions and proofs are offered afterward.

Properties of moment generating functions.

Let X, Y, and $\{X_i\}_{i=1}^{\infty}$ *be random variables whose moment generating functions exist.*

(i) (*Uniqueness*) *If* $M_X(t) = M_Y(t)$ *for all* $t \in (-\varepsilon, \varepsilon)$*, then X and Y have the same distribution.*

(ii) (*Moments*) $E(X^k) = M_X^{(k)}(0)$*, where* $M_X^{(k)}$ *denotes the kth derivative of* M_X.

(iii) (*Linear functions of random variables*) $M_{aX+b}(t) = e^{bt} M_X(at)$.

(iv) (*Sums of independent random variables*) *If X and Y are independent, then* $M_{X+Y}(t) = M_X(t)\,M_Y(t)$*. Likewise, if* $\{X_i\}_{i=1}^{n}$ *are independent and* $S_n = \sum_{i=1}^{n} X_i$*, then* $M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$.

(v) (*Continuity*) *If* $\lim_{i\to\infty} M_{X_i}(t) = M_X(t)$ *for all* $t \in (-\varepsilon, \varepsilon)$*, then* $\{X_i\}_{i=1}^{n}$ *converges in distribution to X.*

Property (i) shows that if a random variable admits a moment generating function, then this function determines the random variable's distribution. For instance, if we can show that some random variable Y has moment generating function $M_Y(t) = \mathrm{e}^{t^2/2}$, we can conclude from this and (7.A.2) that Y has a standard normal distribution.

We can provide an idea of why this is true by focusing on random variables with finite numbers of outcomes. Let X and Y be such random variables, and suppose that their moment generating functions, which are defined for all real numbers, are equal:

$$M_X(t) = \sum_x \mathrm{e}^{tx}\,\mathrm{P}(X = x) = \sum_y \mathrm{e}^{ty}\,\mathrm{P}(Y = y) = M_Y(t). \tag{7.A.3}$$

To show that X and Y have the same distribution, notice first that if t is large enough, the summands in (7.A.3) with the largest exponents will be much larger than all of the others. Therefore, if we let $\hat{x}$ and $\hat{y}$ be the largest positive probability outcomes of X and Y, then for large enough t we have

$$\mathrm{e}^{t\hat{x}}\,\mathrm{P}(X = \hat{x}) \approx M_X(t) = M_Y(t) \approx \mathrm{e}^{t\hat{y}}\,\mathrm{P}(Y = \hat{y}).$$

This can only be true if $\hat{x} = \hat{y}$ and $\mathrm{P}(X = \hat{x}) = \mathrm{P}(Y = \hat{x})$: that is, X and Y must have the same largest outcome, and this outcome must occur with the same probability. If we now remove the terms corresponding to this outcome from the sums in (7.A.3), we can use the same argument to deduce that X and Y have the same second-largest outcome, and that these occur with the same probability. And so on.[3]

[3] The general proof of uniqueness uses a quite different argument, not least because that moment generating functions are only required to exist in a neighborhood of 0. See Patrick Billingsley, *Probability and Measure*, 3rd ed., Wiley, 1995, Section 30.

Property (ii) explains the name of the moment generating function: the values of $E(X^k)$ for nonegative integers k, known as the **moments** of X, can be are recovered from M_X by evaluating its derivatives at 0. To see why this is true, we evaluate the first few moments in the discrete case:[4]

$$M_X(0) = \sum_x e^{tx}\,P(X = x)\Big|_{t=0} = 1 = E(X^0);$$

$$\begin{aligned}
M'_X(0) &= \frac{d}{dt}\sum_x e^{tx}\,P(X = x)\Big|_{t=0} \\
&= \sum_x xe^{tx}\,P(X = x)\Big|_{t=0} \\
&= E(X);
\end{aligned}$$

$$\begin{aligned}
M''_X(0) &= \frac{d^2}{(dt)^2}\sum_x e^{tx}\,P(X = x)\Big|_{t=0} \\
&= \frac{d}{dt}\sum_x xe^{tx}\,P(X = x)\Big|_{t=0} \\
&= \sum_x x^2e^{tx}\,P(X = x)\Big|_{t=0} \\
&= E(X^2).
\end{aligned}$$

Proceeding inductively shows that $E(X^k) = \sum_x x^k e^{tx}\,P(X = x)\Big|_{t=0} = M_X^{(k)}(0)$.

Example If $Z \sim N(0,1)$, then $E(Z) = 0$ and $E(Z^2) = \mathrm{Var}(Z) + (E(Z))^2 = 1$. What about higher moments? Since the moment generating function of Z exists, all of these moments exist. The symmetry of the standard normal distribution about 0 implies that the $E(X^k) = 0$ when k is odd. (Why?)

What if k is even? To determine $E(Z^4)$ using property (ii), we compute as follows:

$$M_X^{(4)}(0) = \frac{d^4}{(dt)^4}e^{t^2/2}\Big|_{t=0} = \left(t^4e^{t^2/2} + 6t^2e^{t^2/2} + 3e^{t^2/2}\right)\Big|_{t=0} = 3.$$

Proceeding inductively, one can show that $E(Z^{2n}) = (2n!)/(2^n n!)$.

[4]The analysis in the continuous case requires results that justify switching the order of differentiation and integration; see George Casella and Roger L. Berger, *Statistical Inference*, 2nd edition, Duxbury/Thomson, 2002, Section 2.4.

Property (iii), the formula for the moment generating function of a linear transformation of a random variable, is both useful and easy to verify:

$$M_{aX+b}(t) = \mathrm{E}(\mathrm{e}^{(aX+b)t}) = \mathrm{e}^{bt}\mathrm{E}(\mathrm{e}^{atX}) = \mathrm{e}^{bt}M_X(at).$$

■ Example *Normal random variables.*

What is the moment generating function of a normal random variable with mean μ and variance σ^2? If $Z \sim N(0, 1)$ is standard normal, then $Y = \mu + \sigma Z$ has a $N(\mu, \sigma^2)$ distribution, so property (iii) yields

$$M_Y(t) = M_{\mu+\sigma Z}(t) = \mathrm{e}^{\mu t}M_Z(\sigma t) = \mathrm{e}^{\mu t}\mathrm{e}^{(\sigma t)^2/2} = \mathrm{e}^{\mu t+\sigma^2 t^2/2}. \tag{7.A.4}$$ ■

Property (iv) is the fundamental fact we noted to start the section. While it is often difficult to compute the distribution or density of a sum of independent random variables directly, computing its moment generating function is merely a matter of multiplication. This property follows easily from two facts: (a) the exponential of a sum equals the product of the exponentials, and (b) the expectation of the product of independent variables equals the product of their expectations (see equation (4.20)). In particular:

$$M_{S_n}(t) = \mathrm{E}\left(\mathrm{e}^{t\sum_{i=1}^n X_i}\right) = \mathrm{E}\left(\prod_{i=1}^n \mathrm{e}^{tX_i}\right) = \prod_{i=1}^n \mathrm{E}(\mathrm{e}^{tX_i}) = \prod_{i=1}^n M_{X_i}(t).$$

■ Example *Binomial random variables.*

Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. *Bernoulli*(p) random variables. As we know from Chapter 5, their sum $S_n = \sum_{i=1}^n X_i$ has a *binomial*(n, p) distribution. Since each summand has moment generating function $M_{X_i}(t) = (1 - p) + p\mathrm{e}^t$, property (iv) implies that the moment generating function of a *binomial*(n, p) random variable is

$$M_{S_n}(t) = \prod_{i=1}^n M_{X_i}(t) = ((1 - p) + p\mathrm{e}^t)^n.$$ ■

■ Example *Sums of independent normal random variables.*

In Chapter 6, we claimed that the sum of independent normal random variables is also normal. We now have the tools to prove this. Suppose that $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent. Then by property (iv) and equation (7.A.4),

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \mathrm{e}^{\mu_X t+\sigma_X^2 t^2/2}\mathrm{e}^{\mu_Y t+\sigma_Y^2 t^2/2} = \mathrm{e}^{(\mu_X+\mu_Y)t+(\sigma_X^2+\sigma_Y^2)t^2/2}.$$

Thus equation (7.A.4) implies that $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. ■

Continuity property (v) says that if the moment generating functions M_{X_i} of a sequence of random variables $\{X_i\}_{i=1}^n$ converges pointwise to the moment generating function M_X of some random variable X, then the former random variables converge in distribution to the latter one. This property, which like property (i) is not easy to prove,[5] justifies the final step in the proof of the central limit theorem.

7.A.2 Proof of the central limit theorem

We now present a proof of the central limit theorem for the case in which the moment generating function of the trials X_i exists on some interval $(-\varepsilon, \varepsilon)$. The proof is based on the properties of moment generating functions discussed above, as well as two facts from calculus. The first is the second-order Taylor approximation for a smooth function g about the point 0:

$$g(x) = g(0) + g'(0)x + \tfrac{1}{2}g''(0)x^2 + O(x^3). \tag{7.A.5}$$

Here $O(x^3)$ is a term that approaches 0 at least as fast as x^3 as x approaches 0. The second is the following fact from calculus, an extension of the continuous-compounding characterization (8.2) of e^x:

$$\text{If } \lim_{n\to\infty} a_n = a, \text{ then } \lim_{n\to\infty} \left(1 + \tfrac{a_n}{n}\right)^n = \mathrm{e}^a. \tag{7.A.6}$$

We prove this fact at the end of the section.

Now consider the i.i.d. sequence $\{X_i\}_{i=1}^\infty$ with $\mathrm{E}(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. To simplify the calculations, we introduce the random variables $Y_i = (X_i - \mu)/\sigma$, so that $\mathrm{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = 1$. Letting Z have a standard normal distribution, we would like to show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \xrightarrow{d} Z \text{ as } n \to \infty. \tag{7.A.7}$$

Equation (7.A.7) is the conclusion of the central limit theorem for the i.i.d. sequence $\{Y_i\}_{i=1}^\infty$. Combining (7.A.7) with the shifting and scaling properties of the normal distribution yields the central limit theorem for the original sequence $\{X_i\}_{i=1}^\infty$.

We analyze the moment generating function of the left-hand side of (7.A.7). Properties (iii) and (iv) imply that

$$M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}(t) = M_{\sum_{i=1}^n Y_i}(\tfrac{t}{\sqrt{n}}) = \left(M_{Y_1}(\tfrac{t}{\sqrt{n}})\right)^n.$$

[5] Again, see Section 30 of Billingsley (footnote 3).

By fact (ii) and the fact that $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$ (see (4.15)),

$$M'_{Y_1}(0) = \mathrm{E}(Y_1) = 0 \text{ and}$$
$$M''_{Y_1}(0) = \mathrm{E}(Y_1^2) = \mathrm{Var}(Y_1) + (\mathrm{E}(Y_1))^2 = 1 + 0 = 1.$$

Applying the Taylor approximation (7.A.5) shows that for each fixed t in some small enough interval $(-\varepsilon, \varepsilon)$,

$$\begin{aligned} M_{Y_1}(\tfrac{t}{\sqrt{n}}) &= M_{Y_1}(0) + M'_{Y_1}(0)(\tfrac{t}{\sqrt{n}} - 0) + \tfrac{1}{2}M''_{Y_1}(0)(\tfrac{t}{\sqrt{n}} - 0)^2 + O(\tfrac{1}{n^{3/2}}) \\ &= 1 + 0 + \tfrac{1}{2}\tfrac{t^2}{n} + O(\tfrac{1}{n^{3/2}}), \end{aligned}$$

where $O(n^{3/2})$ is an expression that approaches 0 at least as fast as $n^{3/2}$ as n grows large. Using (7.A.6), we conclude that for all $t \in (-\varepsilon, \varepsilon)$,

$$\begin{aligned} \lim_{n\to\infty} M_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}(t) &= \lim_{n\to\infty} \left(M_{Y_1}(\tfrac{t}{\sqrt{n}})\right)^n \\ &= \lim_{n\to\infty} \left(1 + \tfrac{1}{n}\left(\tfrac{t^2}{2} + O(\tfrac{1}{\sqrt{n}})\right)\right)^n \\ &= \mathrm{e}^{t^2/2}. \end{aligned}$$

Since this is the moment generating function for the standard normal distribution, this calculation and facts (v) and (i) prove the theorem.

CHARACTERISTIC FUNCTIONS.

The proof of the central limit theorem presented above only applies to sequences of i.i.d. random variables whose moment generating function exists. This rules out random variables for which the probabilities of outcomes quite far from zero do not vanish quickly enough.[6]

To prove the central limit theorem using only the assumption that the trials have finite variance, one replaces arguments based on moment generating functions with ones based on characteristic functions. The **characteristic function** of random variable X is defined for all $t \in (-\infty, \infty)$ by

$$\phi(t) = \mathrm{E}(\mathrm{e}^{itX}), \tag{7.A.8}$$

(*continued*)

[6]One distribution with a finite variance whose moment generating function does not exist is the lognormal distribution, introduced in Exercise 6.M.6.

(*continued*)

where i is the imaginary unit.[7] The characteristic function can separated into its real and imaginary parts by applying Euler's formula, $\mathrm{e}^{\mathrm{i}x} = \cos x + \mathrm{i}\sin x$, to (7.A.8):

(7.A.9) $$\phi(t) = \mathrm{E}(\cos(tX)) + \mathrm{i}\,\mathrm{E}(\sin(tX)).$$

Since characteristic functions are defined using complex numbers, working with them requires some use of complex variable theory, although only a few basic facts are actually needed.

Apart from this complication, working with characteristic functions has many advantages. Characteristic functions satisfy analogues of the five properties of moment generating functions listed above. But unlike moment generating functions, characteristic functions are guaranteed to exist: expression (7.A.9) implies that for all $t \in (-\infty, \infty)$, $\phi(t)$ is a point in the unit disk in the complex plane.

In addition, there are general formulas that allow one to recover a random variable's distribution directly from its characteristic function. As an illustration, suppose that the random variable X has density function f. Given f, we can compute the characteristic function ϕ using the definition

$$\phi(t) = \mathrm{E}(\mathrm{e}^{\mathrm{i}tX}) = \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}tx} f(x)\,\mathrm{d}x.$$

Conversely, if we know the characteristic function ϕ, we can compute the density function f using the **inversion formula**[8]

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}tx} \phi(t)\,\mathrm{d}t.$$

Because of these useful properties, characteristic functions are the standard tool for proving the central limit theorem and related results.

Proof that $\lim_{n\to\infty} a_n = a$ *implies that* $\lim_{n\to\infty} (1 + \frac{a_n}{n})^n = e^a$.

We show first that for $x > 1 - \sqrt{2}/2$,

$$x - x^2 \leq \ln(1 + x) \leq x.$$

To prove these inequalities, observe that $f(x) = x - x^2$, $g(x) = \ln(1 + x)$, and $h(x) = x$ all equal 0 and have slope 1 at $x = 0$. The second inequality then follows from the concavity of the logarithm function. For the first inequality, note that $f''(x) = -2$ is

[7] Away from probability theory, the characteristic function is known as the *Fourier transform*.

[8] There is a general inversion formula that allows one to recover the distribution of any random variable from its characteristic function—see Billingsley (footnote 3), Section 26.

Trim Size: 7.5in x 9.25in Sandholm bapp01.tex V1 - 10/05/2018 3:41pm Page 9

less than $g''(x) = -1/(1+x)^2$ when $x > 1 - \sqrt{2}/2$. Since $f'(0) = g'(0)$, the fundamental theorem of calculus implies that $f'(x) > g'(x)$ when $x \in (1 - \sqrt{2}/2, 0)$ and that $f'(x) < g'(x)$ when $x > 0$. Then since $f(0) = g(0)$, the first inequality follows.

Now taking the logarithm of $(1 + \frac{a_n}{n})^n$ and applying the inequalities shows that for n large enough,

$$n\left(\frac{a_n}{n} - (\frac{a_n}{n})^2\right) \leq n\ln(1 + \frac{a_n}{n}) \leq n(\frac{a_n}{n}).$$

Since the leftmost expression is $a_n - a_n(a_n/n)$ and the rightmost is a_n, the fact that a_n converges to a implies that both of these expressions converge to a. This in turn implies that $n\ln(1 + \frac{a_n}{n})$ converges to a, which proves the claim.

KEY TERMS AND CONCEPTS

moment generating function (p. 1)
moments (p. 4)
characteristic function (p. 7)
inversion formula (p. 8)

Exercises

Exercise 7.A.1. Random variable N has a *Poisson*(μ) distribution if

$$\text{(8.3)} \qquad P(N = k) = \frac{e^{-\mu}\mu^k}{k!} \quad \text{for } k \in \{0, 1, 2, 3, \dots\}$$

Many properties of the Poisson distribution are derived using the series formula for e^x:

$$\text{(8.1)} \qquad \sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x.$$

In Chapter 8, we use this formula to show that (8.3) defines a legitimate probability distribution, and to show that $E(N) = \text{Var}(N) = \mu$.

a. Use formula (8.1) to express the moment generating function for N in a simple form.
b. Using your answer to part (a), show that if S_n is the sum of n independent *Poisson*(μ) random variables it has a *Poisson*$(n\mu)$ distribution.
c. Use your answer to part (a) and formula (8.1) to show that if S_n has a *Poisson*$(n\mu)$ distribution, then $(S_n - n\mu)/\sqrt{n\mu}$ converges in distribution to $Z \sim N(0, 1)$.

d. Explain the connection between your answer to part (c) and the central limit theorem.

Exercise 7.A.2. Suppose that $X \sim \textit{uniform}(0, 1)$, so that $\mathrm{E}(X) = \frac{1}{2}$ and $\mathrm{Var}(X) = \frac{1}{12}$.

a. What is the moment generating function of X? (Be careful not to write down any undefined expressions. Once you see the problem here, you need not be as careful in the remaining parts of the question.)
b. Let S_n be the sum of n i.i.d. *uniform*(0, 1) random variables. What is the moment generating function of S_n?
c. Using your answers to the previous parts, determine the moment generating function of $(S_n - \frac{n}{2})/\sqrt{\frac{n}{12}}$.
d. Using your answer to part (c), the continuity property of moment generating functions, and the Taylor expansion

$$\frac{\mathrm{e}^x - \mathrm{e}^{-x}}{2} = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots,$$

show that $(S_n - \frac{n}{2})/\sqrt{\frac{n}{12}}$ converges in distribution to a standard normal random variable.