Math 341: From Generating Functions to the Central Limit Theorem

Steven J. Miller

November 10, 2009

Contents

Chapter 1

From Generating Functions to the Central Limit Theorem

The purpose of this note is to describe the theory and applications of generating functions, in particular, how they can be used to prove the Central Limit Theorem (CLT) in certain special cases. Unfortunately a proof in general requires some results from complex or Fourier analysis; we will state these needed results and discuss how the proof proceeds in general. We give several examples, including how, appropriately scaled, the mean of n independent Poisson variables converges to the standard normal distribution $N(0,1)$.

1.1 Generating Functions

1.1.1 Motivation

Frequently in mathematics we encounter complex data sets, and then do operations on it to make it even more complex! For example, imagine the first data set is the probabilities that the random variable X_1 takes on given values, and the second set is the probabilities of another random variable X_2 taking on given values. From these we can, painfully through brute force, determine the probabilities of $X_1 + X_2$ equaling anything; however, if at all possible we would like to avoid these tedious computations.

Let's consider the case when X_1 has the Poisson distribution with parameter 5 and X_2 is a Poisson with parameter 7. This means

$$\begin{aligned} \text{Prob}(X_1 = m) &= 5^m e^{-5}/m! \\ \text{Prob}(X_2 = n) &= 7^n e^{-7}/n!, \end{aligned} \tag{1.1.1}$$

where m and n range over the non-negative integers. Our answer is thus

$$\text{Prob}(X_1 + X_2 = k) \;=\; \sum_{\ell=0}^{k} \text{Prob}(X_1 = \ell)\text{Prob}(X_2 = k-\ell) \;=\; \sum_{\ell=0}^{k} \frac{5^\ell e^{-5}}{\ell!} \cdot \frac{7^{k-\ell} e^{-7}}{(k-\ell)!}. \tag{1.1.2}$$

For general sums of random variables, it would be hard to write this in a more illuminating manner; however, we're lucky for sums of Poisson random variables *if we happen to think of the following sequence of simplifications!*

1. First, note that we have a factor of $1/\ell!(k-\ell)!$. This is almost $\binom{k}{\ell}$, which is $k!/\ell!(k-\ell)!$. We do one of the most useful tricks in mathematics, we multiply cleverly by 1, where we write 1 as $k!/k!$. Thus this factor becomes $\binom{k}{\ell}/k!$. As our sum is over ℓ, we may pull the $1/k!$ outside the ℓ-sum.

2. The e^{-5} and e^{-7} inside the sum do not depend on ℓ, so we may pull them out, giving us an e^{-12}.

3. We now have $\frac{e^{-12}}{k!}\sum_{\ell=0}^{k}\binom{k}{\ell}5^{\ell}7^{k-\ell}$. Recalling the Binomial Theorem, we see the ℓ-sum is just $(5+7)^k$, or just 12^k.

Putting all the pieces together, we find

$$\text{Prob}(X_1 + X_2 = k) \;=\; \frac{12^k e^{-12}}{k!}; \tag{1.1.3}$$

note this is the probability density for a Poisson random variable with parameter 12 (and $12 = 5+7$). There is nothing special about 5 and 7 in the argument above. Working more generally, we see the sum of two Poisson random variables with parameters λ_1 and λ_2 is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

Exercise 1.1.1. *Using induction, prove the sum of n Poisson random variables with parameters $\lambda_1, \dots, \lambda_n$ is a Poisson random variable with parameter $\lambda_1 + \cdots + \lambda_n$.*

We were fortunate in this case in that we found a 'natural' way to manipulate the algebra so that we could recognize the answer. What would happen if we considered other sums of random variables? We want a procedure that will work in general, which will *not* require us to see these clever algebra tricks.

Fortunately, there is such an approach. It's the theory of generating functions. We'll first describe what generating functions are (there are several variants; depending on what you are studying, some versions are more useful than others), and then show some applications.

1.1.2 Definitions

Definition 1.1.2 (Generating Function)**.** *Given a sequence $\{a_n\}_{n=0}^{\infty}$, we define its generating function by*

$$G_a(s) \;=\; \sum_{n=0}^{\infty} a_n s^n \tag{1.1.4}$$

for all s where the sum converges.

Depending on our data, it's possible for the generating function to exist for all s, for only some s, or sadly only $s = 0$ (as $G_s(0) = a_0$, this isn't really saying much!). For example,

1. If $a_n = 1/n!$, then $G_a(s) = \sum_{n=0}^{\infty} s^n/n!$. This is the definition of e^s, and hence $G_a(s)$ exists for all s.

2. If $a_n = 2^n$, then $G_a(s) = \sum_{n=0}^{\infty}(2s)^n$. This is a geometric series with ratio $2s$; the series converges for $|2s| < 1$ and diverges if $|2s| > 1$. Thus $G_a(s) = (1-2s)^{-1}$ if $|s| < 1/2$.

3. If $a_n = n!$, a little inspection shows $G_a(s)$ diverges for any $|s| > 0$. Probably the easiest way to see that this series diverges is to note that the terms do not tend to zero. Stirling's formula gives $n! \sim (n/e)^n\sqrt{2\pi n}$, so $n!s^n > (ns/e)^n$, which doesn't go to zero as whenever $n > e/|s|$ we have $|n!s^n| > 1$. **ADD REF TO STIRLING**

If we are given a sequence $\{a_m\}_{m=0}^{\infty}$, then clearly we know its generating function (it may not be *easy* to write down a closed form expression for $G_a(s)$, but we do have a formula for it). The converse is also true: if we know a generating function $G_a(s)$ (which converges for $|s| < r$ for some r), then we can recover the original sequence. This is easy if we can differentiate $G_a(s)$ arbitrarily many times, as then $a_m = \frac{1}{m!}\frac{d^m G_a(s)}{ds^m}$. This result is extremely important; as we'll use it frequently later, it's worth isolating as a theorem.

Theorem 1.1.3 (Uniqueness of generating functions of sequences)**.** *Let $\{a_m\}_{m=0}^{\infty}$ and $\{b_m\}_{m=0}^{\infty}$ be two sequences of numbers with generating functions $G_a(s)$ and $G_b(s)$ which converge for $|s| < r$. Then the two sequences are equal (i.e., $a_i = b_i$ for all i) if and only if $G_a(s) = G_b(s)$ for all $|s| < r$. We may recover the sequence from the generating function by differentiating: $a_m = \frac{1}{m!}\frac{d^m G_a(s)}{ds^m}$.*

Proof. Clearly if $a_i = b_i$ then $G_a(s) = G_b(s)$. For the other direction, if we can differentiate arbitrarily many times, we find $a_i = \frac{1}{i!}\frac{d^i G_a(s)}{ds^i}$ and $b_i = \frac{1}{i!}\frac{d^i G_b(s)}{ds^i}$; as $G_a(s) = G_b(s)$, their derivatives are equal and thus $a_i = b_i$. □

Remark 1.1.4. *The division by $n!$ is a little annoying; later we'll see a related generating function that doesn't have this factor. If we don't want to differentiate, then we get a_0 by setting $s = 0$. We can then find a_1 by looking at $(G_a(s) - a_0)/s$ and setting $s = 0$ in this expression; continuing in this manner we can find any a_m. Note how similar this is to differentiating!*

A natural question to ask is why is it worth constructing a generating series. After all, if it is just equivalent to our original sequence of data, what have we gained? There are advantages; the most important is that it helps simplify the algebra we'll encounter in probability. We give two examples to remind the reader how useful it can be to simplify algebra.

The first is from calculus, and involves telescoping series.

Example 1.1.5. *Consider the following addition problem: evaluate*

$$\begin{array}{rrrr} & 12 & - & 7 \\ + & 45 & - & 12 \\ + & 231 & - & 45 \\ + & 7981 & - & 231 \\ + & 9812 & - & 7981. \end{array} \tag{1.1.5}$$

The 'natural' way to do this is to do evaluate each line and then add; if we do this we get

$$5 + 33 + 186 + 7750 + 1831 \;=\; 9805 \tag{1.1.6}$$

(or at least that's what we got when we used Mathematica). A much faster way to do this is to regroup; we have a $+12$ and a -12, and so these terms cancel. Similarly we have a $+45$ and a -45, so these terms cancel. In the end we are left with

$$9812 - 7 \;=\; 9805, \tag{1.1.7}$$

a much simpler problem! (One application of telescoping series is in the proof of the fundamental theorem of calculus, where they are used to show the area under the curve $y = f(x)$ from $x = a$ to b is given by $F(b) - F(a)$, where F is any anti-derivative of f.)

We turn to linear algebra for our second example.

Example 1.1.6. *Consider the matrix*

$$A \;=\; \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}; \tag{1.1.8}$$

what is A^{100}? If your probability (or linear algebra) grade depended on you getting this right, you would be in good shape. So long as you don't make any algebra errors, after a lot of brute force computations (namely 99 matrix multiplications!) you'll find

$$A^{100} \;=\; \begin{pmatrix} 218922995834555169026 & 354224848179261915075 \\ 354224848179261915075 & 573147844013817084101 \end{pmatrix}. \tag{1.1.9}$$

We can find this answer much faster if we diagonalize A. The eigenvalues of A are $\varphi = \frac{1+\sqrt{5}}{2}$ and $-1/\varphi$, with corresponding eigenvectors

$$\vec{v}_1 \;=\; \begin{pmatrix} -1+\varphi \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{v}_2 \;=\; \begin{pmatrix} -1-1/\varphi \\ 1 \end{pmatrix}. \tag{1.1.10}$$

Letting $S = (\vec{v}_1 \; \vec{v}_2)$ and $\Lambda = \begin{pmatrix} \varphi & 0 \\ 0 & -1/\varphi \end{pmatrix}$, we see $A = S\Lambda S^{-1}$. The key observation is that $S^{-1}S = I$, the 2×2 identity matrix. Thus

$$A^2 \;=\; (S\Lambda S^{-1})(S\Lambda S^{-1}) \;=\; S\Lambda(S^{-1}S)\Lambda S^{-1} \;=\; S\Lambda^2 S^{-1}; \tag{1.1.11}$$

more generally,

$$A^n \;=\; S\Lambda^n S^{-1}. \tag{1.1.12}$$

If we only care about finding A^2, this is significantly more work; however, there is a lot of savings if n is large. Note how similar this is to the telescoping example, with all the $S^{-1}S$ terms canceling.

Remark 1.1.7. *As you might have guessed, this is not a randomly chosen matrix! This matrix arises in solving the Fibonacci difference equation, $a_{n+1} = a_n + a_{n-1}$, and φ is the golden mean. If we let*

$$\vec{v}_0 \;=\; \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \vec{v}_n \;=\; \begin{pmatrix} a_n \\ a_{n+1} \end{pmatrix}, \tag{1.1.13}$$

then $\vec{v}_n \;=\; A^n \vec{v}_0$. Thus, if we know A^n, we can quickly compute how many rabbits are alive at time n without *having to compute how many were alive at time 1, time 2, ..., time $n-1$.*

Remark 1.1.8. *There are two reasons to simplify algebra. One is for computational efficiency, the other is to illuminate connections.*

In the next subsection we show how generating functions behave nicely with convolution, and from this we'll finally get our examples of why generating functions are so useful.

1.1.3 Convolutions I: Discrete random variables

If we have two sequences $\{a_m\}_{m=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$, we define their convolution to be the new sequence $\{c_k\}_{k=0}^{\infty}$ given by

$$c_k \;=\; a_0 b_k + a_1 b_{k-1} + \cdots + a_{k-1} b_1 + a_k b_0 \;=\; \sum_{\ell=0}^{k} a_\ell b_{k-\ell}. \tag{1.1.14}$$

We frequently write this as $c = a * b$. This definition arises from multiplying polynomials; if $f(x) = \sum_{m=0}^{\infty} a_m x^m$ and $g(x) = \sum_{n=0}^{\infty} b_n x^n$, then assuming everything converges we have

$$h(x) \;=\; f(x)g(x) \;=\; \sum_{k=0}^{\infty} c_k x^k, \tag{1.1.15}$$

with $c = a * b$. For example, if $f(x) = 2 + 3x - 4x^2$ and $g(x) = 5 - x + x^3$, then $f(x)g(x) = 10 + 13x - 23x^2 + 6x^3 + 3x^4 - 4x^5$. According to our definition, c_2 should equal

$$a_0 b_2 + a_1 b_1 + a_2 b_0 \;=\; 2 \cdot 0 + 3 \cdot (-1) + (-4) \cdot 5 \;=\; -23, \tag{1.1.16}$$

which is exactly what we get from multiplying $f(x)$ and $g(x)$.

Replacing the dummy variable x with s, we find

Lemma 1.1.9. *Let $G_a(s)$ be the generating function for $\{a_m\}_{m=0}^{\infty}$ and $G_b(s)$ the generating function for $\{b_n\}_{n=0}^{\infty}$. Then the generating function of $c = a * b$ is $G_c(s) = G_a(s)G_b(s)$.*

We can now give our first application of how generating functions can simplify algebra.

Example 1.1.10. *What is $\sum_{m=0}^{n} \binom{n}{m}^2$? If we evaluate this sum for small values of n we find that when $n = 1$ the sum is 1, when $n = 2$ it is 6, when $n = 3$ it is 20, then 70 and then 252. We might realize that the answer seems to be $\binom{2n}{n}$, but even if we notice this, how would we prove it? A natural idea is to try induction. We could write $\binom{n}{m}^2$ as $\left(\binom{n-1}{m-1} + \binom{n-1}{m}\right)^2$ (noting that we have to be careful when $m = 0$). If we expand the square we get two sums similar to the initial sum but with an $n-1$ instead of an n, which we would know by induction; the difficulty is that we have the cross term $\binom{n-1}{m-1}\binom{n-1}{m}$ to evaluate, which requires some effort to get this to look like something nice times something like $\binom{n-1}{\ell}^2$.*

Using generating functions, the answer just pops out. Let $a = \{a_m\}_{m=0}^{n}$, where $a_m = \binom{n}{m}$. Thus

$$G_a(s) \;=\; \sum_{m=0}^{n} \binom{n}{m} s^m \;=\; \sum_{m=0}^{n} \binom{n}{m} s^m 1^{n-m} \;=\; (1+s)^n \tag{1.1.17}$$

(when we have binomial sums such as this, it is very *useful to introduce factors such as 1^{n-m}, which facilitates using the Binomial Theorem).*

Let $c = a * a$, so by Lemma 1.1.9 we have $G_c(s) = G_a(s)G_a(s) = G_a(s)^2$. At first this doesn't seem too useful, until we note that

$$c_n \;=\; \sum_{\ell=0}^{n} a_\ell a_{n-\ell} \;=\; \sum_{\ell=0}^{n} \binom{n}{\ell}\binom{n}{n-\ell} \;=\; \sum_{\ell=0}^{n} \binom{n}{\ell}^2 \tag{1.1.18}$$

as $\binom{n}{n-\ell} = \binom{n}{\ell}$. Thus the answer to our problem is c_n. We don't know c_n, but we do *know its generating function, and the entire point of this exercise is to show sometimes it is more useful to know one and deduce the other. We have*

$$\sum_{k=0}^{2n} c_k s^k \;=\; G_c(s) \;=\; G_a(s)^2 \;=\; (1+s)^n \cdot (1+s)^n \;=\; (1+s)^{2n} \;=\; \sum_{k=0}^{2n} \binom{2n}{k} s^k. \tag{1.1.19}$$

Thus $c_n = \binom{2n}{n}$ as claimed.

While we have finally found an example where it is easier to study the problem through generating functions, some things are unsatisfying about this problem. The first is we still needed to have some combinatorial expertise, noting $\binom{n}{\ell} = \binom{n}{n-\ell}$; this is minor for two reasons. First, this is one of the most important properties of binomial coefficients (the number of ways of choosing ℓ people from n people when order doesn't matter is the same as the number of ways of excluding $n - \ell$). The second is more severe: *why would one ever consider convolving our sequence a with itself to solve this problem!*

The answer to the second objection is that convolutions arise all the time in probability, and thus it is natural to study any process which is nice with respect to convolution. To see this, we define

Definition 1.1.11 (Probability generating function)**.** *Let X be a discrete random variable taking on values in the integers. Let $G_X(s)$ be the generating function to $\{a_m\}_{m=-\infty}^{\infty}$ with $a_m = \mathrm{Prob}(X = m)$. Then $G_X(s)$ is called the probability generating function. If X is only non-zero at non-negative integers, a very useful way of computing $G_X(s)$ is to note that*

$$G_X(s) \;=\; \mathbb{E}[s^X] \;=\; \sum_{m=0}^{\infty} s^m \mathrm{Prob}(X = m). \tag{1.1.20}$$

The function $G_X(s)$ can be a bit more complicated than the other generating functions we've seen if X takes on negative values; if this is the case, we are no longer guaranteed that $G_X(0)$ makes sense! One way we can get around this problem is by restricting to s with $0 < \alpha < |s| < \beta$ for some α, β; another is to restrict ourselves to random variables taking on non-negative integer values. We concentrate on the latter. While this does restrict a bit the distributions we may study, so many of the common, important probability distributions take on non-negative integer values that we will still have a wealth of examples and applications.

We can now state one of the most important results for probability generating functions.

Theorem 1.1.12. *Let X_1 and X_2 be* independent *discrete random variables taking on non-negative integer values, with corresponding probability generating functions $G_{X_1}(s)$ and $G_{X_2}(s)$. Then $G_{X_1+X_2}(s) = G_{X_1}(s)G_{X_2}(s)$.*

Proof. The proof proceeds from unwinding the definitions. We have

$$\mathrm{Prob}(X_1 + X_2 = k) \quad = \quad \sum_{\ell=0}^{\infty} \mathrm{Prob}(X_1 = \ell)\mathrm{Prob}(X_2 = k - \ell). \tag{1.1.21}$$

If we let $a_m = \mathrm{Prob}(X_1 = m)$, $b_n = \mathrm{Prob}(X_2 = n)$ and $c_k = \mathrm{Prob}(X_1 + X_2 = k)$, we see that $c = a * b$. Thus $G_c(s) = G_a(s)G_b(s)$, or equivalently, $G_{X_1+X_2}(s) = G_{X_1}(s)G_{X_2}(s)$. □

Remark 1.1.13. *Whenever you see a theorem, you should remove a hypothesis and ask if it is still true. Usually the answer is a resounding* NO! *(or, if true, the proof is usually significantly harder). In the theorem above, how important is it for the random variables to be independent? As an extreme example consider what would happen if $X_2 = -X_1$. Then $X_1 + X_2$ is identically zero, but $G_{X_1+X_2}(s) \neq G_{X_1}(s)G_{-X_1}(s)$.*

The above shows why generating functions play such a central role in probability: *the density of the sum of two independent random variables is the convolution of their probabilities!*

Exercise 1.1.14. *Generalize Theorem 1.1.12 to the sum of a finite number of independent random variables. In particular, if $X, X_1, \ldots, X_n$ are independent, identically distributed discrete random variables taking on values in the non-negative integers, prove $G_{X_1+\cdots+X_n}(s) = G_X(s)^n$.*

1.1.4 Convolutions II: Continuous random variables

The results of the previous subsection readily generalize to continuous random variables. We first generalize the notion of convolution, and then show how this applies to continuous random variables.

Definition 1.1.15 (Convolution)**.** *The convolution of two functions f_1 and f_2, denoted $f_1 * f_2$, is*

$$(f_1 * f_2)(x) \;=\; \int_{-\infty}^{\infty} f_1(t) f_2(x-t) dt. \tag{1.1.22}$$

Let X_1 and X_2 be continuous random variables with densities f_1 and f_2, and set $X = X_1 + X_2$. Consider the convolution of their densities:

$$(f_1 * f_2)(x) \;=\; \int_{-\infty}^{\infty} f_1(t) f_2(x-t) dt. \tag{1.1.23}$$

Note that if we want $X_1 + X_2 = x$, then $X_1 = t$ for some t and X_2 is then forced to be $x - t$. Thus this integral gives the probability density for $X_1 + X_2$, which we denote by f. In other words,

$$f(x) \;=\; (f_1 * f_2)(x) \;=\; \int_{-\infty}^{\infty} f_1(t) f_2(x-t) dt. \tag{1.1.24}$$

We check that f is a density. As f_1 and f_2 are densities, they are non-negative and thus the integral defining $f(x)$ is clearly non-negative. We must show that if we integrate over all x that we get 1. We have

$$\begin{aligned}
\int_{x=-\infty}^{\infty} f(x) dx &= \int_{x=-\infty}^{\infty} \int_{t=-\infty}^{\infty} f_1(t) f_2(x-t) dt dx \\
&= \int_{t=-\infty}^{\infty} f_1(t) \left[\int_{x=-\infty}^{\infty} f_2(x-t) dx \right] dt \\
&= \int_{t=-\infty}^{\infty} f_1(t) \left[\int_{u=-\infty}^{\infty} f_2(u) du \right] dt \\
&= \int_{t=-\infty}^{\infty} f_1(t) \cdot 1 dt \;=\; 1.
\end{aligned} \tag{1.1.25}$$

(In analysis classes we are constantly told to be careful about interchanging orders of integration; this is always permissible in probability theory as our densities take on non-negative values, and thus Fubini's theorem holds.)

This is a natural generalization of the convolution of two sequences, where $c_k = \sum a_\ell b_{k-\ell}$ becomes $(f_1 * f_2)(x) = \int f_1(t) f_2(x-t)$.

It is worth isolating the following result:

Lemma 1.1.16. *The convolution of two sequences or functions is commutative; in other words, $a * b = b * a$ or $f_1 * f_2 = f_2 * f_1$.*

Proof. The proof follows immediately from simple algebra. For example,

$$G_{a*b}(s) \;=\; G_a(s)G_b(s) \;=\; G_b(s)G_a(s) \;=\; G_{b*a}(s); \tag{1.1.26}$$

we could also perform the algebra in the defining sums for c_k, but this is cleaner. We may also see this is true by noting the probabilistic interpretation; if X_1 and X_2 are independent random variables, then $X_1 + X_2 = X_2 + X_1$ (in fact, this is what we're doing when we look at the product of the generating functions). □

1.1.5 Definition and properties of moment generating functions

In Remark 1.1.4 we commented that we can recover our sequence from the generating function through differentiation. In particular, if $a = \{a_m\}_{m=0}^{\infty}$ and $G_a(s) = a_m s^m$, then $a_m = \frac{1}{m!}\frac{d^m G_a(s)}{ds^m}$; however, the factor $1/m!$ is annoying. There is a related generating function that does not have this factor, the moment generating function. Before defining it, we briefly recall the definition of moments.

Definition 1.1.17 (Moments)**.** *Let X be a random variable with density f. Its k^{th} moment, denoted μ_k', is defined by*

$$\mu_k' \;:=\; \sum_{m=0}^{\infty} x_m^k f(x_m) \tag{1.1.27}$$

if X is discrete, taking non-zero values only at the x_m's, and

$$\mu_k' \;:=\; \int_{-\infty}^{\infty} x^k f(x)dx \tag{1.1.28}$$

if X is continuous. In both cases we denote this as $\mu_k' = \mathbb{E}[X^k]$. We define the k^{th} centered moment, μ_k, by $\mu_k := \mathbb{E}[(X - \mu_1')^k]$. We frequently write μ for μ_1' and σ^2 for μ_2.

Whenever we deal with a discrete random variable, we let $\{x_m\}_{m=-\infty}^{\infty}$ or $\{x_m\}_{m=0}^{\infty}$ or $\{x_m\}_{m=1}^{\infty}$ denote the set of points where the probability density is non-zero. In most applications, we have $\{x_m\}_{m=-\infty}^{\infty} = \{0, 1, 2, \dots\}$.

Definition 1.1.18 (Moment generating function)**.** *Let X be a random variable. The moment generating function of X, denoted $M_X(t)$, is given by $M_X(t) = \mathbb{E}[e^{tX}]$. Explicitly, if X is discrete then*

$$M_X(t) \;=\; \sum_{m=-\infty}^{\infty} e^{tx_m} f(x_m), \tag{1.1.29}$$

while if X is continuous then

$$M_X(t) \;=\; \int_{-\infty}^{\infty} e^{tx} f(x)dx. \tag{1.1.30}$$

Of course, it is not clear that $M_X(t)$ exists for any value of t. Frequently what happens is that it exists for some, but not all, t. Usually this is enough to allow us to deduce an amazing number of facts. We now collect many of the nice properties of the moment generating function, which show its usefulness in probability.

Theorem 1.1.19. *Let X be a random variable with moments μ'_k.*

1. *We have*
$$M_X(t) \;=\; 1 + \mu'_1 t + \frac{\mu'_2 t^2}{2!} + \frac{\mu'_3 t^3}{3!} + \cdots; \tag{1.1.31}$$
in particular, $\mu'_k = d^k M_X(t)/dt^k\Big|_{t=0}$.

2. *Let α and β be constants. Then*
$$M_{\alpha X+\beta}(t) \;=\; e^{\beta t} M_X(\alpha t). \tag{1.1.32}$$
Useful special cases are $M_{X+\beta}(t) = e^{\beta t} M_X(t)$ and $M_{\alpha X}(t) = M_X(\alpha t)$; when proving the central limit theorem, it is also useful to have $M_{(X+\beta)/\alpha}(t) = e^{\beta t/\alpha} M_X(t/\alpha)$.

3. *Let X_1 and X_2 be independent random variables with moment generating functions $M_{X_1}(t)$ and $M_{X_2}(t)$ which converge for $|t| < r$. Then*
$$M_{X_1+X_2}(t) \;=\; M_{X_1}(t) M_{X_2}(t). \tag{1.1.33}$$
More generally, if $X_1, \ldots, X_N$ are independent random variables with moment generating functions $M_{X_i}(t)$ which converge for $|t| < \delta$, then
$$M_{X_1+\cdots+X_N}(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_N}(t). \tag{1.1.34}$$
If the random variables all have the same moment generating function $M_X(t)$, then the right hand side becomes $M_X(t)^N$.

Proof. For notational convenience, we only prove the claims when X is a continuous random variable with density f.

1. As the first claim is so important (this is the reason moment generating functions are studied!) we provide two proofs. The two proofs are similar, and both require some results from analysis for general f.

   For our first proof, we use the series expansion for the exponential function: $e^{tx} = \sum_{k=0}^{\infty} (tx)^k/k!$. We have
$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} \sum_{k=0}^{\infty} \frac{x^k t^k}{k!} f(x)\,dx \\ &= \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{\infty} x^k f(x)\,dx; \end{aligned} \tag{1.1.35}$$

the claim follows by noting the integral is just the definition of the k^{th} moment μ'_k. Note this proof requires us to switch the order of an integral and a sum; this can be justified if $M_X(t)$ converges for $|t| < \delta$ for some positive δ.

For our second proof, differentiate $M_X(t)$ a total of k times. Arguing that the derivative of the integral is the integral of the derivative **GIVE REF!**, and noting the only t-dependence in the integrand is the e^{tx} factor, we find

$$\frac{d^k M_X}{dt^k} \;=\; \int_{-\infty}^{\infty} \left[\frac{d^k e^{tx}}{dt^k}\right] f(x)dx \;=\; \int_{-\infty}^{\infty} x^k e^{tx} f(x)dx; \tag{1.1.36}$$

the claim now follows from taking $t = 0$ and recalling the definition of the moments.

2. We now turn to the second claim. We have

$$\begin{aligned} M_{\alpha X+\beta}(t) &= \int_{-\infty}^{\infty} e^{t(\alpha x+\beta)} f(x)dx \\ &= e^{\beta t}\int_{-\infty}^{\infty} e^{t\alpha x} f(x)dx \;=\; e^{\beta t} M_X(\alpha t), \end{aligned} \tag{1.1.37}$$

as the last integral is just the moment generating function evaluated at αt instead of t. The special cases now readily follow.

3. The third property follows from the fact that the expected value of independent random variables is the product of the expected values. If X_1 and X_2 are independent, so too is the pair e^{tX_1} and e^{tX_2} (remember t is fixed). Thus

$$\begin{aligned} M_{X_1+X_2}(t) &= \mathbb{E}[e^{t(X_1+X_2)}] \\ &= \mathbb{E}[e^{tX_1}e^{tX_2}] \\ &= \mathbb{E}[e^{tX_1}]\mathbb{E}[e^{tX_2}] \;=\; M_{X_1}(t)M_{X_2}(t). \end{aligned} \tag{1.1.38}$$

□

1.1.6 Applications of moment generating functions

Let's do some examples where we compute moment generating functions and see how useful they can be.

Example 1.1.20. *Let X be a Poisson random variable with parameter λ and density f; this means that*

$$f(n) \;=\; \text{Prob}(X = n) \;=\; \frac{\lambda^n e^{-\lambda}}{n!} \tag{1.1.39}$$

for $n \geq 0$, and 0 otherwise. The moment generating function is

$$\begin{aligned} M_X(t) &= \sum_{n=0}^{\infty} e^{tn} f(n) \\ &= \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n e^{tn}}{n!} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda e^t} \;=\; e^{\lambda(e^t-1)}. \end{aligned} \tag{1.1.40}$$

From part (3) of Theorem 1.1.19, if X_1 and X_2 are independent Poisson random variables with parameters λ_1 and λ_2, then

$$\begin{aligned} M_{X_1+X_2}(t) &= M_{X_1}(t) M_{X_2}(t) \\ &= e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} \\ &= e^{(\lambda_1+\lambda_2)(e^t-1)}. \end{aligned} \tag{1.1.41}$$

Note this is exactly the moment generating function of a Poisson random variable with parameter $\lambda_1 + \lambda_2$, obtained with significantly less work than the brute force approach! Does this imply that $X_1 + X_2$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$? Yes, because of the following theorem.

Theorem 1.1.21 (Uniqueness of moment generating functions for discrete random variables)**.** *Let X and Y be discrete random variables taking on non-negative integer values (i.e., they are non-zero only in $\{0, 1, 2, \dots\}$) with moment generating functions $M_X(t)$ and $M_Y(t)$, each of which converges for $|t| < \delta$. Then X and Y have the same distribution if and only if $M_X(t) = M_Y(t)$ for $|t| < \delta$.*

In other words, discrete random variables are uniquely determined by their moment generating functions (if they converge).

Proof. One direction is trivial; namely, if X and Y have the same distribution then clearly $M_X(t) = M_Y(t)$. What about the other direction? We'll first prove the claim in a simpler case, and then tackle the general setting.

If X and Y are non-zero only finitely often, the proof is much simpler. Imagine this is the case; let X take on non-zero values $0 \leq x_1 < x_2 < \cdots < x_m$ with positive probabilities $p_1, \dots, p_m$, and let Y take on non-zero values $0 \leq y_1 < y_2 < \cdots < y_n$ with positive probabilities $q_1, \dots, q_n$. As the moment generating functions are equal, by part (3) of Theorem 1.1.19 *all* the moments are equal, as

$$M_X(t) \;=\; 1 + \mu_1' t + \frac{\mu_2' t^2}{2!} + \frac{\mu_3' t^3}{3!} + \cdots \;=\; M_Y(t). \tag{1.1.42}$$

For the k^{th} moment, this means

$$p_1 x_1^k + \cdots + p_m x_m^k \;=\; q_1 y_1^k + \cdots + q_n y_n^k. \tag{1.1.43}$$

As this must hold for all k, it seems absurd that it could be true unless $m = n$, $x_i = y_i$ and $p_i = q_i$. There are many ways to see this. Assume $x_m \neq y_n$; without loss of generality let's assume $y_n > x_m$. For k enormous, the left hand side of (1.1.43) is essentially $p_m x_m^k$, while the right hand side is basically $q_n y_n^k$, and the right hand side will be magnitudes larger. For example, imagine $p_m = .3$, $x_m = 100$, $q_n = .001$ and $y_n = 150$. When $k = 5$, $p_m x_m^k \approx 3 \cdot 10^9$ while $q_n y_n^k \approx 7.5 \cdot 10^7$; when $k = 21$ these numbers become approximately $3 \cdot 10^{41}$ and $5 \cdot 10^{42}$, while if $k = 1001$ it becomes approximately $3 \cdot 10^{2001}$ and $2 \cdot 10^{2175}$. The proof is completed by induction; we leave the details to the reader as we'll give another, more complete proof below. The reason this proof works is that, for a distribution non-zero only finitely often, the high moments are essentially controlled by the largest value.

The following proof is more direct, and works for any discrete distributions. From Theorem 1.1.3, we know that two sequences $\{a_m\}_{m=0}^{\infty}$ and $\{b_n\}_{n=0}^{\infty}$ are equal if and only if their generating functions are equal. Let $a_m = \text{Prob}(X = m)$ and $b_n = \text{Prob}(Y = n)$. The generating functions are (see Definition 1.1.2)

$$\begin{aligned} G_a(s) &= \mathbb{E}[s^X] \;=\; \sum_{m=0}^{\infty} s^m \text{Prob}(X = m) \\ G_b(s) &= \mathbb{E}[s^Y] \;=\; \sum_{n=0}^{\infty} s^n \text{Prob}(Y = n); \end{aligned} \tag{1.1.44}$$

however, the generating functions are trivially related to the moment generating functions through

$$M_X(t) \;=\; \mathbb{E}[e^{tX}], \quad M_Y(t) \;=\; \mathbb{E}[e^{tY}]. \tag{1.1.45}$$

If we let $s = e^t$, we find $G_a(e^t) = M_X(t)$ and $G_b(e^t) = M_Y(t)$; as $M_X(t) = M_Y(t)$, $G_a(e^t) = G_b(e^t)$. We now know that the generating functions are equal, and hence by Theorem 1.1.3 the corresponding sequences are equal. But this means $\text{Prob}(X = i) = \text{Prob}(Y = i)$ for all i, and so the two densities are the same. □

Remark 1.1.22. *There is a lot to remark about in the theorem above. It is* very *useful; it says that the moment generating function of a discrete random variable which is non-zero only at the non-negative integers* uniquely *determines the distribution! While there are* a lot *of hypotheses in this statement, these are fairly mild ones. Most of the discrete distributions we study and use are supported on the non-negative integers, so this is not that restrictive an assumption. Arguing as in Remark 1.1.4, however, we can remove this hypothesis. Imagine first that the random variables only take on non-negative values, so we have*

$$\begin{aligned} G_a(s) &= \mathbb{E}[s^X] \;=\; \sum_{m=0}^{\infty} a_m s^{x_m} \\ G_b(s) &= \mathbb{E}[s^Y] \;=\; \sum_{n=0}^{\infty} b_n s^{y_n}. \end{aligned} \tag{1.1.46}$$

Without loss of generality, assume $x_0 \le y_0$. As $G_a(s)/s^{x_0} = G_b(s)/s^{x_0}$ for all s, sending $s \to 0$ gives $a_0 = b_0 \lim_{s\to 0} s^{y_0-x_0}$; as each $a_m \neq 0$, the only way this can hold is if $y_0 = x_0$ and $a_0 = b_0$. We continue in this manner (specifically, we play this game again, except now our two functions are $G_a(s) - a_0 s^{x_0}$ and $G_b(s) - a_0 s^{x_0}$).

We now return to Example 1.1.20. Using moment generating functions, we saw the sum of two Poisson random variables with parameters λ_1 and λ_2 had its moment generating function equal to $e^{(\lambda_1+\lambda_2)(e^t-1)}$. As the moment generating function of a Poisson random variable with parameter λ is just $e^{\lambda(e^t-1)}$, by Theorem 1.1.21 we can now conclude that the sum of two Poisson random variables with parameters λ_1 and λ_2 is a Poisson random variable, with parameter equal to the $\lambda_1 + \lambda_2$ (see also Exercise 1.1.1).

We now consider a continuous example.

Example 1.1.23. *Let X be an exponentially distributed random variable with parameter λ, so its density function is $f(x) = \lambda^{-1}e^{-x/\lambda}$ for $x \ge 0$ and 0 otherwise. We can calculate its moment generating function:*

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \cdot \frac{e^{-x/\lambda}}{\lambda} dx \\ &= \frac{1}{\lambda}\int_{-\infty}^{\infty} e^{-(\lambda^{-1}-t)x} dx. \end{aligned} \tag{1.1.47}$$

We change variables by setting $u = (\lambda^{-1} - t)x$, so $dx = du/(\lambda^{-1} - t)$. As long as $\lambda^{-1} > t$ (in other words, so long as $t < 1/\lambda$) the exponential has a negative argument, and thus converges. We find

$$M_X(t) = \frac{1}{\lambda}\int_{-\infty}^{\infty} e^{-u}\frac{du}{\lambda^{-1}-t} = \frac{1}{1-\lambda t}\int_{-\infty}^{\infty} e^{-u}du = (1-\lambda t)^{-1}. \tag{1.1.48}$$

In our analysis we needed $t < 1/\lambda$; note that for such t, the resulting expression for $M_X(t)$ makes sense. (While $(1 - \lambda t)^{-1}$ makes sense for all $t \neq 1/\lambda$, clearly something is happening when t goes from below $1/\lambda$ to above.)

If X_i ($i \in \{1, 2\}$) are independent exponentially distributed random variables with parameters λ_i, from the example above and part (3) of Theorem 1.1.19 we find $M_{X_1+X_2}(t) = (1 - \lambda_1 t)^{-1}(1 - \lambda_2 t)^{-1}$. What does this imply about the distribution of $X_1 + X_2$? Is it anything nice? What if we restrict to the special case $\lambda_1 = \lambda_2 = \lambda$? Can we say anything here?

The following dream theorem would make life easy: *A probability distribution is uniquely determined by its moments.* This would be the natural analogue of Theorem 1.1.21 for continuous random variables. Is it true? **Sadly, this is not always the case; there exist distinct probability distributions which have the same moments.** The standard example given are the following two densities, defined for $x \ge 0$ by

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi x^2}}\, e^{-(\log^2 x)/2} \\ f_2(x) &= f_1(x)\left[1 + \sin(2\pi \log x)\right]. \end{aligned} \tag{1.1.49}$$

In the next three subsections we explore what goes wrong with the functions from (1.1.49). After seeing what the problem is, we discuss what additional properties we need to assume to prevent such an occurance. The solution involves results from complex analysis, which will tell us when a moment generating function (of a continuous random variable) uniquely determines a probability distribution.

1.2 Complex Analysis Results

1.2.1 Warnings from real analysis

The following example is one of our favorites from real analysis. It indicates why real analysis is hard, almost surely much harder than you might expect.

Example 1.2.1. *Consider the function $g : \mathbb{R} \to \mathbb{R}$ given by*

$$g(x) \;=\; \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{1.2.1}$$

Using the definition of the derivative and L'Hopital's rule, we can show that f is infinitely differentiable, and all of its derivatives at the origin vanish. For example,

$$\begin{aligned} g'(0) &= \lim_{h\to 0} \frac{e^{-1/h^2} - 0}{h} \\ &= \lim_{h\to 0} \frac{1/h}{e^{1/h^2}} \\ &= \lim_{k\to\infty} \frac{k}{e^{k^2}} \\ &= \lim_{k\to\infty} \frac{1}{2ke^{k^2}} \;=\; 0, \end{aligned} \tag{1.2.2}$$

where we used L'Hopital's rule in the last step ($\lim_{k\to\infty} A(k)/B(k) = \lim_{k\to\infty} A'(k)/B'(k)$ if $\lim_{k\to\infty} A(k) = \lim_{k\to\infty} B(k) = \infty$). A similar analysis shows $g^{(n)}(0) = 0$ for any n. If we consider the Taylor series for g about 0, we find

$$g(x) \;=\; g(0) + g'(0)x + \frac{g''(0)x^2}{2!} + \cdots \;=\; \sum_{n=0}^{\infty} \frac{g^{(n)}(0)x^n}{n!} \;=\; 0; \tag{1.2.3}$$

however, clearly $g(x) \neq 0$ if $x \neq 0$. We are thus in the ridiculous case where the Taylor series (which converges for all x!) only agrees with the function when $x = 0$. This isn't that impressive, as the Taylor series is forced to agree with the original function at 0, as both are just $g(0)$.

There is a lot we can learn from the above example. The first is that it is possible for a Taylor series to converge for all x, but only agree with the function at one point! The second, which is far

more important, is that *a Taylor series does not uniquely determine a function!* For example, both $\sin x$ and $\sin x + g(x)$ (with $g(x)$ the function from the previous example) have the same Taylor series about $x = 0$.

The reason this is so important for us is that we want to understand when a moment generating function uniquely determines a probability distribution. If our distribution was discrete, there was no problem (Theorem 1.1.21). For continuous distributions, however, it is much harder, as we saw in (1.1.21) where we met two densities that had the same moments.

It is therefore apparant that we must impose some additional conditions for continuous random variables. For discrete random variables, it was enough to know all the moments; this doesn't suffice for continuous random variables. What should those conditions be?

Let's consider again the pair of functions in (1.1.21). A nice calculus exercise shows that $\mu'_k = e^{k^2/2}$. This means that the moment generating function is

$$M_X(t) \;=\; \sum_{k=0}^{\infty} \frac{\mu'_k t^k}{k!} \;=\; \sum_{k=0}^{\infty} \frac{e^{k^2/2} t^k}{k!}; \tag{1.2.4}$$

for what t does this series converge? We claim it converges *only* when $t = 0$! To see this, it suffices to show that the terms do not tend to zero. As $k! \le k^k$, for any fixed t, for k sufficiently large $t^k/k! \ge (t/k)^k$; moreover, $e^{k^2/2} = (e^{k/2})^k$, so the k^{th} term is at least as large as $(e^{k/2}t/k)$. For any $t \neq 0$, this clearly does not tend to zero, and thus the moment generating function has a radius of convergence of zero!

This leads us to the following conjecture: *If the moment generating function converges for $|t| < \delta$ for some r, then it uniquely determines a density.* We'll explore this conjecture in the following subsections.

1.2.2 Complex analysis definitions

Our purpose here is to give a flavor of what kind of inputs are needed to ensure that a moment generating function uniquely determines a probability density. We first collect some definitions, and then state some useful results from complex analysis.

Definition 1.2.2 (Complex variable, complex function)**.** *Any complex number z can be written as $z = x + iy$, with x and y real. A complex function is a map f from $\mathbb{C}$ to $\mathbb{C}$; in other words $f(z) \in \mathbb{C}$. Frequently one writes $x = \mathfrak{Re}(z)$, $y = \mathfrak{Im}(z)$, and $f(z) = u(x, y) + iv(x, y)$ with u and v functions from $\mathbb{R}^2$ to $\mathbb{R}$.*

Definition 1.2.3 (Differentiable)**.** *We say a complex function f is differentiable at z_0 if it is differentiable with respect to the complex variable z, which means*

$$\lim_{h \to 0} \frac{f(z_0 + h) - f(z_0)}{h} \tag{1.2.5}$$

exists, where h tends to zero along any *path in the complex plane. If the limit exists we write $f'(z_0)$ for the limit. If f is differentiable, then f satisfies the Cauchy-Riemann equations:*

$$f'(z) \;=\; \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} \;=\; -i\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} \tag{1.2.6}$$

(one direction is easy, arising from sending $h \to 0$ along the paths $\widetilde{h}$ and $i\widetilde{h}$, with $\widetilde{h} \in \mathbb{R}$).

Many of the theorems below deal with open sets. We briefly review their definition and give some examples.

Definition 1.2.4 (Open set, closed set)**.** *A subset U of $\mathbb{C}$ is an open set if for any $z_0 \in U$ there is a δ such that whenever $|z - z_0| < \delta$ then $z \in U$ (note δ is allowed to depend on z_0). A set C is closed if its complement, $\mathbb{C} \setminus C$, is open.*

Example 1.2.5. *The following are examples of open sets in $\mathbb{C}$:*

1. *$U_1 = \{z : |z| < r\}$ for any $r > 0$. This is usually called the ball of radius r centered at the origin.*
2. *$U_2 = \{z : \mathfrak{Re}(z) > 0\}$. To see this is open, if $z_0 \in U_2$ then we can write $z_0 = x_0 + iy_0$, with $x_0 > 0$. Letting $\delta = x_0/2$, for $z = x + iy$ we see that if $|z - z_0| < \delta$ then $|x - x_0| < x_0/2$, which implies $x > x_0/2 > 0$. U_2 is often called the open right half-plane.*

For examples of closed sets, consider

1. *$C_1 = \{z : |z| \leq r\}$. Note that if we take z_0 to be any point on the boundary, then the ball of radius δ centered at z_0 will contain points more than r units from the origin, and thus C_1 is not open. A little work shows, however, that C_1 is closed (in fact, C_1 is called the closed ball of radius r about the origin).*
2. *$C_2 = \{z : \mathfrak{Re}(z) \geq 0\}$. To see this set is not open, consider any $z_0 = iy$ with $y \in \mathbb{R}$. A similar calculation as the one we did for U_2 shows C_2 is closed.*

For a set that is neither open nor closed, consider $S = U_1 \cup C_2$.

Definition 1.2.6 (Holomorphic, analytic)**.** *Let U be an open subset of $\mathbb{C}$, and let f be a complex function. We say f is holomorphic on U if f is differentiable at every point $z \in U$, and we say f is analytic on U if f has a series expansion that converges and agrees with f on U. This means that for any $z_0 \in U$, for z close to z_0 we can choose a_n's such that*

$$f(z) \;=\; \sum_{n=0}^{\infty} a_n (z - z_0)^n. \tag{1.2.7}$$

Saying a function of a complex variable is differentiable turns out to imply *far* more than saying a function of a real variable is differentiable, as the following theorem shows us.

Theorem 1.2.7. *Let f be a complex function and U an open set. Then f is holomorphic on U if and only if f is analytic on U, and the series expansion for f is its Taylor series.*

The above theorem is amazing; its result seems to good to be true. Namely, as soon as we know f is differentiable once, it is infinitely differentiable and f agrees with its Taylor series expansion! This is very different than what happens in the case of functions of a real variable. For instance, the function

$$h(x) \;=\; x^3 \sin(1/x) \tag{1.2.8}$$

is differentiable once and only once at $x = 0$, and while the function $g(x)$ from (1.2.1) is infinitely differentiable, the Taylor series expansion only agrees with $g(x)$ at $x = 0$.

The next theorem provides a very nice condition for when a function is identically zero. It involves the notion of a limit or accumulation point, which we define first.

Definition 1.2.8 (Limit or accumulation point)**.** *We say z is a limit (or an accumulation) point of a sequence $\{z_n\}_{n=0}^{\infty}$ if there exists a subsequence $\{z_{n_k}\}_{k=0}^{\infty}$ converging to z.*

Example 1.2.9. *We give some examples.*

1. *If $z_n = 1/n$, then 0 is a limit point.*
2. *If $z_n = \cos(\pi n)$ then there are two limit points, namely 1 and -1. (If $z_n = \cos(n)$ then* every *point in $[-1, 1]$ is a limit point of the sequence, though this is harder to show.)*
3. *If $z_n = (1 + (-1)^n)^n + 1/n$, then 0 is a limit point. We can see this by taking the subsequence $\{z_1, z_3, z_5, z_7, \dots\}$; note the subsequence $\{z_0, z_2, z_4, \dots\}$ diverges to infinity.*
4. *Let z_n denote the number of distinct prime factors of n. Then every positive integer is a limit point! For example, let's show 5 is a limit point. The first five primes are 2, 3, 5, 7 and 11; consider $N = 2 \cdot 3 \cdot 5 \cdot 7 \cdot 11 = 2310$. Consider the subsequence $\{z_N, z_{N^2}, z_{N^3}, z_{N^4}, \dots\}$; as N^k has exactly 5 distinct prime factors for each k, 5 is a limit point.*
5. *If $z_n = n^2$ then there are no limit points, as $\lim_{n\to\infty} z_n = \infty$.*
6. *Let z_0 be any odd, positive integer, and set*

$$z_{n+1} \;=\; \begin{cases} 3z_n + 1 & \text{if } z_n \text{ is odd} \\ z_n/2 & \text{if } z_n \text{ is even.} \end{cases} \tag{1.2.9}$$

It is conjectured *that 1 is always a limit point (and if some $z_m = 1$, then the next few terms have to be $4, 2, 1, 4, 2, 1, 4, 2, 1, \dots$, and hence the sequence cycles). This is the famous $3x+1$ problem. Kakutani called it a conspiracy to slow down American mathematics because of the amount of time people spent on this; Erdos said mathematics is not yet ready for such problems.* ***ADD REFS TO LAGARIAS FOR MORE INFO.***

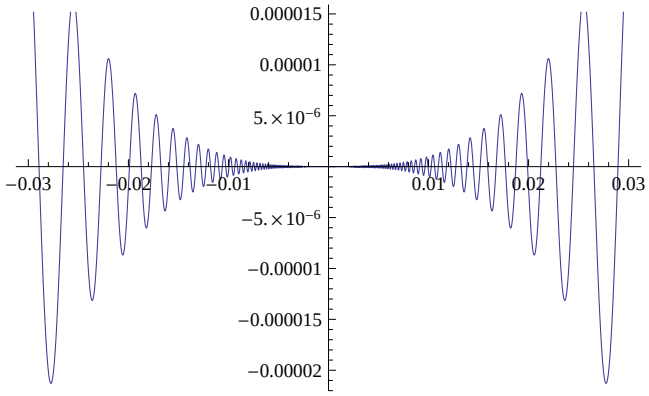

Figure 1.1: Plot of $x^3 \sin(1/x)$.

Theorem 1.2.10. *Let f be an analytic function on an open set U, with infinitely many zeros $z_1, z_2, z_3, \ldots$. If $\lim_{n\to\infty} z_n \in U$, then f is identically zero on U. In other words, if a function is zero along a sequence in U whose accumulation point is also in U, then that function is identically zero in U.*

Note the above is *very* different than what happens in real analysis. Consider again the function from (1.2.8),

$$h(x) \;=\; x^3 \sin(1/x). \tag{1.2.10}$$

This function is continuous and differentiable. It is zero whenever $x = \frac{1}{\pi n}$ with n an integer. If we let $z_n = \frac{1}{\pi n}$, we see this sequence has 0 as a limit point, and our function is alos zero at 0 (see Figure 1.1). It is clear, however, that this function is *not* identically zero. Yet again, we see a stark difference between real and complex valued functions. As a nice exercise, show that $x^3 \sin(1/x)$ is *not* complex differentiable. It will help if you recall $e^{i\theta} = \cos\theta + i \sin\theta$, or $\sin\theta = (e^{i\theta} - e^{-i\theta})/2$.

1.2.3 Integral transforms

Given a function $K(x, y)$ and an interval I (which is frequently $(-\infty, \infty)$ or $[0, \infty)$), we can construct a map from functions to functions as follows: send f to $\int_I f(x)K(x, y)dx$. As the integrand depends on the two variables x and y and we only integrate out x, the result will be a function of y. Obviously it does not matter what letters we use for the dummy variables; other common choices are $K(t, x)$ or $K(t, s)$ or $K(x, \xi)$.

Integral transforms are useful for studying a variety of problems. Their utility stems from the fact that the related function leads to simpler algebra for the problem at hand. We define two of the most important integral transforms, the Laplace and the Fourier transforms.

Definition 1.2.11 (Laplace Transform)**.** *Let $K(t, s) = e^{-ts}$. The Laplace transform of f, denoted $\mathcal{L}f$, is given by*

$$(\mathcal{L}f)(s) \;=\; \int_0^\infty f(t)e^{-st}dt. \tag{1.2.11}$$

Given a function g, its inverse Laplace transform, $\mathcal{L}^{-1}g$, is

$$(\mathcal{L}^{-1}g)(t) \;=\; \lim_{T\to\infty}\frac{1}{2\pi i}\int_{c-iT}^{c+iT} e^{st}g(s)ds \;=\; \lim_{T\to\infty}\frac{1}{2\pi i}\int_{-T}^{T} e^{(c+i\tau)t}g(c+i\tau)id\tau. \tag{1.2.12}$$

Definition 1.2.12 (Fourier Transform)**.** *Let $K(x,y) = e^{-2\pi ixy}$. The Fourier transform of f, denoted $\mathcal{F}f$ or $\widehat{f}$, is given by*

$$\widehat{f}(y) \;=\; \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx, \tag{1.2.13}$$

where

$$e^{i\theta} \;:=\; \sum_{n=0}^{\infty}\frac{\theta^n}{n!} \;=\; \cos\theta + i\sin\theta. \tag{1.2.14}$$

The inverse Fourier transform of g, denoted $\mathcal{F}^{-1}g$, is

$$(\mathcal{F}^{-1}g)(x) \;=\; \int_{-\infty}^{\infty} g(y)e^{2\pi ixy}dy. \tag{1.2.15}$$

Note other books define the Fourier transform differently, sometimes using $K(x,y) = e^{-ixy}$ or $K(x,y) = e^{-ixy}/\sqrt{2\pi}$.

Remark 1.2.13. *The Laplace and Fourier transforms are related. If we let $s = 2\pi iy$ and consider functions $f(x)$ which vanish for $x \le 0$, we see the Laplace and Fourier transforms are equal.*

Given a function f we can compute its transform. What about the other direction? If we are told g is the transform of some function f, can we recover f from knowing g? If yes, is the corresponding f unique? Fortunately, the answer to both questions turns out to be 'yes', provided f and g satisfy certain nice conditions. A particularly nice set of functions to study is the Schwartz space.

Definition 1.2.14 (Schwartz space)**.** *The Schwartz space, $\mathcal{S}(\mathbb{R})$, is the set of all infinitely differentiable functions f such that, for any non-negative integers m and n,*

$$\sup_{x\in\mathbb{R}}\left|(1+x^2)^m\frac{d^nf}{dx^n}\right| \;<\; \infty, \tag{1.2.16}$$

where $\sup_{x\in\mathbb{R}}|g(x)|$ is the smallest number B such that $|g(x)| \le B$ for all x (think 'maximum value' whenever you see supremum).

Theorem 1.2.15 (Inversion Theorems)**.** ***ADD STUFF ON LAPLACE!*** *Let $f \in \mathcal{S}(\mathbb{R})$, the Schwartz space. Then*

$$f(x) \;=\; \int_{-\infty}^{\infty}\widehat{f}(y)e^{2\pi ixy}dy, \tag{1.2.17}$$

where $\widehat{f}$ is the Fourier transform of f. In particular, if f and g are Schwartz functions with the same Fourier transform, then $f(x) = g(x)$.

This interplay between a function and its transform will be very useful for us when we study probability distributions, as the moment generating function is an integral transform of the density! Recall the moment generating function is defined by $M_X(t) = \mathbb{E}[e^{tX}]$, or

$$M_X(t) \;=\; \int_{-\infty}^{\infty} e^{tx} f(t)dt. \tag{1.2.18}$$

If $f(x) = 0$ for $x \leq 0$, this is just the Laplace transform of f. Alternatively, if we take $t = -2\pi iy$ then it is the Fourier transform of f. This is trivially related to (yet another!) generating function, the characteristic function of X, which is defined by $\phi(t) = \mathbb{E}[e^{itX}]$.

We now see why these results from complex analysis will save the day. The inversion formulas above tell us that, if our initial distribution is nice, then knowing its integral transform is the same as knowing it; in other words, knowing the integral transform uniquely determines the distribution.

1.2.4 Complex analysis and moment generating functions

We conclude our technical digression by stating a few more very useful facts. The proof of these requires properties of the Laplace transform, which is defined by $(\mathcal{L}f)(s) = \int_0^\infty e^{-sx} f(x)dx$. The reason the Laplace transform plays such an important role in the theory is apparent when we recall the definition of the moment generating function of a random variable X with density f:

$$M_X(t) \;=\; \mathbb{E}[e^{tX}] \;=\; \int_{-\infty}^{\infty} e^{tx} f(x)dx; \tag{1.2.19}$$

in other words, the moment generating function is the Laplace transform of the density evaluated at $-s = t$.

Before stating our results, we recall some notation.

Definition 1.2.16. *Let F_X and G_Y be the cumulative distribution functions of the random variables X and Y with densities f and g. This means*

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f(t)dt \\ G_Y(y) &= \int_{-\infty}^{y} g(v)dv. \end{aligned} \tag{1.2.20}$$

Our results from complex analysis imply the following two *very* important and useful theorems for determining when we have enough information from the moments to uniquely determine a probability density.

Theorem 1.2.17. *Assume the moment generating functions $M_X(t)$ and $M_Y(t)$ exist in a neighborhood of zero (i.e., there is some δ such that both functions exist for $|t| < \delta$). If $M_X(t) = M_Y(t)$ in this neighborhood, then $F_X(u) = F_Y(u)$ for all u. As the densities are the derivatives of the cumulative distribution functions, we have $f = g$.*

Theorem 1.2.18. *Let $\{X_i\}_{i\in I}$ be a sequence of random variables with moment generating functions $M_{X_i}(t)$. Assume there is a $\delta > 0$ such that when $|t| < \delta$ we have $\lim_{i\to\infty} M_{X_i}(t) = M_X(t)$ for some moment generating function $M_X(t)$, and all moment generating functions converge for $|t| < \delta$. Then there exists a unique cumulative distribution function F whose moments are determined from $M_X(t)$ and for all x where $F_X(x)$ is continuous, $\lim_{n\to\infty} F_{X_i}(x) = F_X(x)$.*

The proof of these theorems follow from results in complex analysis, specifically the Laplace and Fourier inversion formulas.

To give an example as to how the results from complex analysis allow us to prove results such as these, we give most of the details in the proof of the next theorem. We *deliberately* do not try and prove the following result in as great generality as possible!

Theorem 1.2.19. *Let X and Y be two continuous random variables on $[0,\infty)$ with continuous densities f and g, all of whose moments are finite and agree. Suppose further that:*

1. *There is some $C > 0$ such that for all $c \le C$, $e^{(c+1)t}f(e^t)$ and $e^{(c+1)t}g(e^t)$ are Schwartz functions (see Definition 1.2.14). This is not a terribly restrictive assumption; f and g need to have decay in order for all moments to exist and be finite. As we are evaluating f and g at e^t and not t, there is enormous decay here. The meat of the assumption is that f and g are infinitely differentiable and their derivatives decay.*

2. *The (not necessarily integral) moments*

$$\mu'_{r_n}(f) \;=\; \int_0^\infty x^{r_n} f(x)dx \quad \text{and} \quad \mu'_{r_n}(g) \;=\; \int_0^\infty x^{r_n} g(x)dx \tag{1.2.21}$$

   agree for some sequence of non-negative real numbers $\{r_n\}_{n=0}^\infty$ which has a finite accumulation point (i.e., $\lim_{n\to\infty} r_n = r < \infty$).

Then $f = g$ (in other words, knowing all these moments uniquely determines the probability density).

Proof. We sketch the proof. Let $h(x) = f(x) - g(x)$, and define

$$A(z) \;=\; \int_0^\infty x^z h(x)dx. \tag{1.2.22}$$

Note that $A(z)$ exists for all z with real part non-negative. To see this, let $\mathfrak{Re}(z)$ denote the real part of z, and let k be the unique non-negative integer with $k \le \mathfrak{Re}(z) < k+1$. Then $x^{\mathfrak{Re}z} \le x^k + x^{k+1}$, and

$$\begin{aligned} |A(z)| &\le \int_0^\infty x^{\mathfrak{Re}(z)} \left[|f(x)| + |g(x)|\right] dx \\ &\le \int_0^\infty (x^k + x^{k+1}) f(x)dx + \int_0^\infty (x^k + x^{k+1}) g(x)dx \;=\; 2\mu'_k + 2\mu'_{k+1}. \end{aligned} \tag{1.2.23}$$

Results from analysis now imply that $A(z)$ exists for all z. The key point is that A is also differentiable. Interchanging the derivative and the integration (which can be justified), we find

$$A'(z) \;=\; \int_0^\infty x^z (\log x) h(x) dx. \tag{1.2.24}$$

To show that $A'(z)$ exists, we just need to show this integral is well-defined. There are only two potential problems with the integral, namely when $x \to \infty$ and when $x \to 0$. For x large, $x^z \log x \le x^{\lceil \Re\mathfrak{e}(z) \rceil}$ (where $\rceil w \lceil$ is the smallest integer at least as large as w) and thus $\left|\int_1^\infty x^z (\log x) h(x) dx\right| < \infty$. For x near 0, $h(x)$ looks like $h(0)$ plus a small error (remember we are assuming f and g continuous). There is a constant $\mathcal{C}_{f,g}$ such that

$$\lim_{\epsilon \to 0} \int_\epsilon^1 \left| \int_0^\infty x^z (\log x) h(x) dx \right| \quad \le \quad \lim_{\epsilon \to 0} \int_\epsilon^1 (\log x) \mathcal{C}_{f,g} dx; \tag{1.2.25}$$

the reason it is less than or equal to the right hand side is that $\Re\mathfrak{e}(z) > 0$ so $|x^z| \le 1$, and since f and g are Schwartz functions they are bounded. The anti-derivative of $\log x$ is $x \log x - x$, and $\lim_{\epsilon \to 0} (\epsilon \log \epsilon - \epsilon) s = 0$. This is enough to prove that this integral is bounded, and thus from results in analysis we get $A'(z)$ exists.

We (finally!) use our results from complex analysis. As A is differentiable once, it is infinitely differentiable and it equals its Taylor series for z with $\Re\mathfrak{e}(z) > 0$. Therefore A is an analytic function which is zero for a sequence of z_n's with an accumulation point, and thus it is identically zero. This is amazing – initially we only knew $A(z)$ was zero if z was a positive integer or if z was in the sequence $\{r_n\}$; we now know it is zero for all z with $\Re\mathfrak{e}(z) > 0$.

We change variables, and replace x with e^t and dx with $e^t dt$. The range of integration is now $-\infty$ to ∞, and we set $\mathfrak{h}(t) dt = h(e^t) e^t dt$. We now have

$$A(z) \;=\; \int_{-\infty}^\infty e^{tz} \mathfrak{h}(t) dt \;=\; 0. \tag{1.2.26}$$

Choosing $z = c + 2\pi i y$ with c less than the C from our hypotheses gives

$$A(c + 2\pi i y) \;=\; \int_{-\infty}^\infty e^{2\pi i t y} \left[e^{ct} \mathfrak{h}(t)\right] dt \;=\; 0. \tag{1.2.27}$$

Our assumptions imply that $e^{ct}\mathfrak{h}(t)$ is a Schwartz function, and thus it has a unique inverse Fourier transform. As we know this transform is zero, it implies that $e^{ct}\mathfrak{h}(t) = 0$, or $h(x) = 0$, or $f(x) = g(x)$. □

Remark 1.2.20. *What if we lessen our restrictions on f and g; perhaps one of them is not continuous? Perhaps there is a unique continuous probability distribution attached to a given sequence of moments such as in the above theorem, but if we allow non-continuous distributions there could be additional possibilities. This topic is beyond the scope of this book, requiring more advanced results from analysis; however, we wanted to point out where the dangers lie, where we need to be careful.*

Exercise 1.2.21. *If we are told that all the moments of f are finite and f is infinitely differentiable, must there be some C such that for all $c < C$ we have $e^{(c+1)t} f(e^t)$ is a Schwartz function?*

After proving Theorem 1.2.19, it's natural to go back to the two densities that are causing so much trouble, namely (see (1.1.49))

$$\begin{aligned} f_1(x) &= \frac{1}{\sqrt{2\pi x^2}}\, e^{-(\log^2 x)/2} \\ f_2(x) &= f_1(x)\left[1 + \sin(2\pi \log x)\right]. \end{aligned} \tag{1.2.28}$$

We know these two densities have the same integral moments (their k^{th} moments are $e^{k^2/2}$ for k a non-negative integer). These functions have the correct decay; note

$$e^{(c+1)t} f_1(e^t) \;=\; e^{(c+1)t} \cdot \frac{e^{-t^2/2}}{\sqrt{2\pi} e^t}, \tag{1.2.29}$$

which decays fast enough for any c to satisfy the assumptions of Theorem 1.2.19. As these two densities are not the same, *some* condition must be violated. The only condition left to check is whether or not we have a sequence of numbers $\{r_n\}_{n=0}^{\infty}$ with an accumulation point $r > 0$ such that the ${r_n}^{\text{th}}$ moments agree. Using complex analysis (specifically, contour integration), we can calculate the $(a+ib)^{\text{th}}$moments. We find

$$(a+ib)^{\text{th}} \text{ moment of } f_1 \text{ is } \;\; e^{(a+ib)^2/2} \tag{1.2.30}$$

and

$$(a+ib) \text{ moment of } f_1 \text{ is } \;\; e^{(a+ib)^2/2} + \frac{i}{2}\left(e^{(a+i(b-2\pi))^2/2} - e^{(a+i(b+2\pi))^2/2}\right). \tag{1.2.31}$$

While these moments agree for $b = 0$ and a a positive integer, there is no sequence of real moments having an accumulation point where they agree. To see this, note that when $b = 0$ the a^{th}moment of f_2 is

$$e^{a^2/2} + e^{(a-2i\pi)^2/2}\left(1 - e^{4ia\pi}\right), \tag{1.2.32}$$

and this is never zero unless a is a half-integer (i.e., $a = k/2$ for some integer k). In fact, the reason we wrote (1.2.32) as we did was to highlight the fact that it is only zero when a is a half-integer. Exponentials of real or complex numbers are never zero, and thus the only way this can vanish is if $1 = e^{4ia\pi}$. Recalling that $e^{i\theta} = \cos\theta + i\sin\theta$, we see that the vanishing of the a^{th}moment is equivalent to $1 - \cos(4\pi a) - i\sin(4\pi a) = 0$; the only way this can happen is if $a = k/2$ for some k. If this happens, the cosine term is 1 and the sine term is 0.

1.3 The Central Limit Theorem

1.3.1 Means, Variances and Standard Deviations

The Central Limit Theorem is one of the true gems of probability. The hypotheses are quite weak, and are frequently met in practice. What is so amazing is the universality of the result. Before

stating the Central Limit Theorem, we set some notation and motivate why we study the quantities we do.

Recall that the mean μ and variance σ^2 of a random variable X with density f is given by

$$\begin{aligned} \mu &= \mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \\ \sigma^2 &= \mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 dx \end{aligned} \tag{1.3.1}$$

if X is a continuous random variable, and

$$\begin{aligned} \mu &= \mathbb{E}[X] = \sum_{n=1}^{\infty} x_n f(x_n) \\ \sigma^2 &= \mathbb{E}[(X-\mu)^2] = \sum_{n=1}^{\infty} (x_i-\mu)^2 \end{aligned} \tag{1.3.2}$$

if X is discrete. We often write $\mathrm{Var}(X)$ for the variance of X. The mean measures the average value of X, and the variance how spread out it is (the larger the variance, the more spread out the density).

Example 1.3.1. *Consider the following two data sets:*

$$\begin{aligned} S_1 &= \{0,0,0,0,0,0,0,0,0,0,100,100,100,100,100,100,100,100,100,100\} \\ S_2 &= \{50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50\}. \end{aligned} \tag{1.3.3}$$

Both data sets have a mean of 50, but the first is clearly more spread out than the second. If we try to compute the variances of these two sets, we run into a problem, namely: what are the probabilities $f(x_i)$? *Unless there is information to the contrary, one typically assumes that all data points are equally likely. There are two ways now to determine the probabilities. The first way is to treat each observation as a different measurement. In that case, we have $x_1 = x_2 = \cdots = x_{10} = 0$, all with probabilities $1/20$, and $x_{11} = x_{12} = \cdots x_{20} = 100$, all with probability $1/20$. Alternatively, we could consider $x_1 = 0$ with probability $1/2$ and $x_2 = 100$ with probability $1/2$. Note, however, that while the* number *of data points is different in the two interpretations, all computed quantities will be the same. For example, let's calculate the variance using both methods. Using the first, we find the variance is*

$$\sum_{n=1}^{10} (0-50)^2 \cdot \frac{1}{20} + \sum_{n=11}^{20} (100-50)^2 \cdot \frac{1}{20} = 10 \cdot 50^2 \cdot \frac{1}{20} + 10 \cdot 50^2 \cdot \frac{1}{20} = 50^2, \tag{1.3.4}$$

while the second method gives

$$(0-50)^2 \cdot \frac{1}{2} + (100-50)^2 \cdot \frac{1}{2} = 50^2. \tag{1.3.5}$$

The second set, S_2, is significantly easier to compute. All values are the same, and thus the variance is clearly zero.

Not surprisingly, the second data set has significantly smaller variance than the first; however, there is something a bit unsettling about using the variance to quantify how spread out a data set is. The difficulty comes when our numbers have physical meaning. For example, imagine the two data sets are recording how wait time (in seconds) for a bank teller. Thus we either have a wait of 0 seconds, of 50 seconds, or of 100 seconds. In both banks the average waiting time of customers is the same; however, in the second bank all customers have the same experience, while in the first some are presumably very happy with no wait, while others are almost surely upset at a very long wait. This can be seen by noting the variance in the second set is zero while in the first it is $50^2 = 2500$; however, it is not quite right to say this. In this situation, there are *units* attached to the variance. As time is measured in seconds, the variance is measured in seconds-squared. To be honest, I have no real clue what a second-squared is. I can imagine a meter-squared (area), but a second-squared? Yet this is *precisely* the unit that arises here. To see this, note that the x_i and μ are measured in seconds, the probabilities are unitless numbers, so the variance is a sum of expressions such as $(0\text{sec} - 50\text{sec})^2$, $(50\text{sec} - 50\text{sec})^2$ and $(100\text{sec} - 50\text{sec})^2$. Thus, the variance is measured in second-squared.

If I want to find out how long I need to wait, I'm expecting an answer such as 'say 10 minutes, plus or minus a minute or two'. I'm *not* expecting anyone to respond with 'say 10 minutes, with a variance of 1 or four minutes-squared'. Fortunately, there is a simple solution to this problem; instead of reporting the variance, it is frequently more appropriate to report the standard deviation, which is the square-root of the variance.

Returning to our earlier example, we would say that for the first data set, the mean wait time was 50 seconds, with a standard deviation of 50 seconds, while in the second it was also a mean wait time of 50 seconds, but with a standard deviation of 0 seconds.

The point of the above is that the standard deviation and the mean have the same units, while the variance and the mean do not; we always want to compare apples and apples (i.e., objects with the same dimensions). In fact, this is why the notation for the variance is σ^2, highlighting the fact that the quantity we will frequently care about is σ, its square-root. Similar to writing $\mathrm{Var}(X)$ for the variance of X, we occasionally write $\mathrm{StDev}(X)$ for its standard deviation.

1.3.2 Normalizations

In the previous subsection we saw that the variance of a random variable is not the right scale to look at fluctuations, as the units were wrong. In particular, if X is measured in seconds then the variance is in the physically mysterious unit of seconds-squared; it is the standard deviation that has the same units, and thus it is the standard deviation that we use to discuss how spread out a data set is.

Finding the correct scale or units to discuss a problem is very important. For example, imagine we have two sections of calculus with identical students in each but very different professors (admittedly this is not an entirely realistic situation as no two classes are identical; however, if the classes are large then this is approximately true). Let's assume one professor writes really easy exams, and another writes very challenging ones. If we're told that Hari from the first section has

a 92 average and Daneel from the second section has an 84, which is the better student? Without more information, it is very hard to judge – how does a 92 in the 'easier' section compare to an '84' in the harder?

Let's assume we know more about the two classes. Let's say that in section 1 (the one with the easier exams) the average grade is a 97 and the standard deviation is 1, while in section 2 the average grade is a 64 and the standard deviation is 10. Once we know this, it's clear that Daneel is the superior student (remember in this pretend example we're assuming the two classes are identical in terms of ability; the only difference is that one takes easier tests than the other). Hari is actually below average (by 5 standard deviations, a sizeable number), while Daneel is significantly above average (by 2 standard deviations).

We are warned about comparing apples and oranges, and that's what happened here – we have two different scales, and an 84 on one scale does not mean the same as an 84 on the other. To avoid problems like this (i.e., to compare apples and apples), we frequently normalize our data to have mean zero and variance 1. This puts different data sets on the same scale. This is done as follows:

Definition 1.3.2 (Normalization of a random variable)**.** *Let X be a random variable with mean μ and standard deviation σ, both of which are finite. The normalization, Y, is defined by*

$$Y \ := \ \frac{X - \mathbb{E}[X]}{\mathrm{StDev}(X)} \ = \ \frac{X - \mu}{\sigma}. \tag{1.3.6}$$

Note that

$$\mathbb{E}[Y] \ = \ 0 \quad \text{and} \quad \mathrm{StDev}(Y) \ = \ 1. \tag{1.3.7}$$

Remark 1.3.3. *Instead of calling the above a normalization we could call it a standardization or a renormalization; after some thought we decided to call it a normalization. One reason is that this process is used all the time in problems involving the Central Limit Theorem, and thus this is setting the stage for the result there (which involves the normal distribution). For a typical X the normalization will not be normally distributed.*

The normalization process we've discussed is quite natural; it rescales any 'nice' random variable to a new one having mean 0 and variance 1. The only assumption we need is that it have finite mean and standard deviation. This a mild assumption, but not all distributions satisfy it. For example, consider the Cauchy distribution

$$f(x) \ = \ \frac{1}{\pi}\frac{1}{1+x^2}. \tag{1.3.8}$$

It is debateable whether or not this distribution has a mean; it clearly doesn't have a finite variance. Why is the mean of this distribution problematic? It's because we have an improper integral where the integrand is sometimes positive and sometimes negative. This means *how* we go to infinity matters. For example,

$$\lim_{A\to\infty} \int_{-A}^{A} \frac{x dx}{\pi(1+x^2)} \ = \ \lim_{A\to\infty} 0 \ = \ 0, \tag{1.3.9}$$

while

$$\lim_{A\to\infty}\int_{-A}^{2A}\frac{xdx}{\pi(1+x^2)} \;=\; \lim_{A\to\infty}\int_{A}^{2A}\frac{xdx}{\pi(1+x^2)}, \tag{1.3.10}$$

and the last integral is, for A enormous, essentially $\int_A^{2A} dx/x = \log(2A) - \log(A) = \log(2)$. Thus, *how* we tend to infinity matters!

Remark 1.3.4. ***We cannot stress enough how important and useful it is to normalize a random variable. We will discuss this again below, but given any random variable*** X***, sending*** X ***to*** $(X - \mathbb{E}[X])/\mathrm{StDev}(X)$ ***is an extremely natural and often useful thing to do.***

1.3.3 Statement of the Central Limit Theorem

Of the many distributions encountered in probability, perhaps the most important is the normal distribution. One way to measure how important a distribution is to a subject is to count how many different names are used to refer to it; in this case, names include the normal distribution, the Gaussian distribution and the bell curve.

Definition 1.3.5 (Normal distribution)**.** *A random variable X is normally distributed (or has the normal distribution, or is a Gaussian random variable) with mean μ and variance σ^2 if the density of X is*

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.3.11}$$

We often write $X \sim N(\mu, \sigma^2)$ to denote this. If $\mu = 0$ and $\sigma^2 = 1$, we say X has the standard normal distribution.

There are many versions of the Central Limit Theorem. The differences range from the hypotheses assumed and the type of convergence obtained; not surprisingly, the more nice properties one assumes, the stronger the convergence. We state a theorem which, while not the most general one possible, is easy to state and has hypotheses satisfied by most of the common distributions we encounter.

Theorem 1.3.6 (Central Limit Theorem)**.** *Let $X_1, \dots, X_N$ be independent, identically distributed random variables whose moment generating functions converge for $|t| < \delta$ for some $\delta > 0$ (this implies all the moments exist and are finite). Denote the mean by μ and the variance by σ^2, let*

$$\overline{X}_N \;=\; \frac{X_1 + \cdots + X_N}{N} \tag{1.3.12}$$

and set

$$Z_N \;=\; \frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}}. \tag{1.3.13}$$

Then as $N \to \infty$, the distribution of Z_N converges to the standard normal (see Definition 1.3.5 for a statement).

One way to interpret the above is as follows: imagine $X_1, \dots, X_N$ are N independent measurements of some process or phenomenon. Then $\overline{X}_N$ is the average of the observed values. As the X_i's are drawn from a common distribution with mean μ, and as expectation is linear, we have

$$\mathbb{E}[\overline{X}_N] \;=\; \mathbb{E}\left[\frac{X_1 + \cdots + X_N}{N}\right] \;=\; \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_n] \;=\; \frac{1}{N}\cdot N\mu \;=\; \mu. \tag{1.3.14}$$

Since the X_n's are independent, the variance of $\overline{X}_N$ is

$$\mathrm{Var}(\overline{X}_N) \;=\; \mathrm{Var}\left(\frac{X_1 + \cdots + X_N}{N}\right) \;=\; \frac{1}{N^2}\sum_{n=1}^{N}\mathrm{Var}(X_n) \;=\; \frac{1}{N^2}\cdot N\sigma^2 \;=\; \frac{\sigma^2}{N}, \tag{1.3.15}$$

so the standard deviation of $\overline{X}_N$ is just $\sigma/\sqrt{N}$. Note as $N \to \infty$ the standard deviation of $\overline{X}_N$ tends to zero. This leads to the following interpretation: as we take more and more measurements, the distribution of the average value is living in a tighter and tighter band about the true mean. We chose to write $\overline{X}_N$ for the average value of the X_n's to emphasize that we have a sum of N random variables.

Most probability books have tables with the values of the standard normal. For example, imagine we want to compute the probability that a random variable with the standard normal distribution is within one standard deviation of 0. We can just turn to the back of the book and grab this information; however, it is extremely unlikely you'll ever find a book with the tabulation of probabilities for a normally distributed random variable with mean $\sqrt{2}$ and variance π.

Why aren't there such tables? There are two reasons today. The first, of course, is that computers are very powerful and accessible, and thus the need for printed tables like the last one alluded to is greatly lessened, as a few lines of code will give us the answer. While this might be a satisfactory answer today, why were there no such tables before computers? Perhaps our example is a bit absurd, but what about a normally distributed random variable with mean 0 and variance 4; surely that must have occurred in someone's research?

The reason we don't need such tables is that, if we know the probabilities for the standard normal, we can use those to compute the probabilities for *any* normally distributed random variable. For definiteness, imaging $W \sim N(3,4)$, which means W has a mean of 3 and a variance of 4 (or a standard deviation of 2). Imagine we want to know the probability that $W \in [2,10]$. We normalize W (see Definition 1.3.2) by setting

$$Z \;=\; \frac{W - \mathbb{E}[W]}{\mathrm{StDev}(W)} \;=\; \frac{W-3}{2}. \tag{1.3.16}$$

Thus asking that $W \in [2,10]$ is equivalent to asking Z to be in a certain interval. Which interval? Well, $W \in [2,10]$ is the same as $Z \in [-1/2, 7/2]$. If we have a table of probabilities for the standard normal, we can now compute this probability, and hence find the probability that $W \in [2,10]$.

We thus see that we need only have *one* table of probabilities for *one* normally distributed random variables, as we can deduce the probabilities for any other with simple algebra. In the days

before computers, this was a *very* important observation. It meant people needed only calculate *one* table of probabilities in order to study *any* normally distribution.

This is very similar to logarithm tables. Most books only had logarithms base e (sometimes base 10 was given, or perhaps base 2). Through a similar normalization process, if we have a table of logarithms in one base we can compute logarithms in any base. This is because of the following log-law (commonly called the Change of Variable formula): For any $b, c, x > 0$ we have

$$\log_c x \;=\; \frac{\log_b x}{\log_b c}. \tag{1.3.17}$$

Imagine we know logarithms base b. Then using the right hand side of the above formula, we can compute the logarithm of any x base c. Thus it suffices to compile just one table of logarithms (as base e and base 10 are often both used, it might be a kindness to assemble both tables, but just one would suffice).

Before proving the Central Limit Theorem, we'll analyze some special cases where the proof is simpler. As our hypotheses include statements about moment generating functions, it should come as no surprise that we'll need to know the moment generating function of the standard normal.

Theorem 1.3.7 (Moment generating function of normal distributions)**.** *Let X be a normal random variable with mean μ and variance σ^2. Its moment generating function satisfies*

$$M_X(t) \;=\; e^{\mu t + \frac{\sigma^2 t^2}{2}}. \tag{1.3.18}$$

In particular, if Z has the standard normal distribution, its moment generating function is

$$M_Z(t) \;=\; e^{t^2/2}. \tag{1.3.19}$$

Sketch of the proof. While we could try to directly compute $M_X(t)$ through $M_X(t) = \mathbb{E}[e^{tX}]$, clearly we would much rather compute $M_Z(t) = \mathbb{E}[e^{tZ}]$. The reason is clear: Z has zero mean and variance 1, and thus the numbers are a little cleaner. We could set up the equation for $M_X(t)$ and then do some change of variables, or we could note that we can deduce $M_X(t)$ from $M_Z(t)$ through part (2) of Theorem 1.1.19. Specifically, we have

$$Z \;=\; \frac{X-\mu}{\sigma}, \tag{1.3.20}$$

or equivalently

$$X \;=\; \sigma Z + \mu. \tag{1.3.21}$$

We then use $M_{\alpha Z+\beta}(t) \;=\; e^{\beta t} M_Z(\alpha t)$.

Thus we are reduced to computing $M_Z(t)$, or

$$M_Z(t) \;=\; \mathbb{E}[e^{tZ}] \;=\; \int_{-\infty}^{\infty} e^{tz} \cdot \frac{e^{-z^2/2} dx}{\sqrt{2\pi}}. \tag{1.3.22}$$

We solve this by completing the square. The argument of the exponential is

$$tz - \frac{z^2}{2} \;=\; -\frac{z^2-2tz}{2} \;=\; -\frac{z^2-2tz+t^2-t^2}{2} \;=\; -\frac{(z-t)^2}{2} + \frac{t^2}{2}. \tag{1.3.23}$$

Note the second term is independent of z, the variable of integration. We find

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2} + \frac{t^2}{2}\right) dz \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{e^{-(z-t)^2/2} dz}{\sqrt{2\pi}} \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{e^{-u^2/2} du}{\sqrt{2\pi}} \;=\; e^{t^2/2}, \end{aligned} \tag{1.3.24}$$

as the last integral is 1 as it is the integral of the standard normal's density from $-\infty$ to ∞. □

1.3.4 Proof of the CLT for sums of Poisson random variables via MGF

As a warm-up for the general proof, we'll show that the normalized sum of Poisson random variables converges to the standard normal distribution. The proof will include the key ideas of the general case, and will involve moment generating functions. We know from Theorem 1.3.7 that the moment generating function of the standard normal is $e^{t^2/2}$. We computed the moment generating function of a Poisson random variable X with mean λ in Example 1.1.20. We showed that it is

$$M_X(t) \;=\; e^{\lambda(e^t-1)} \;=\; 1 + \mu t + \frac{\mu_2' t^2}{2!} + \cdots. \tag{1.3.25}$$

Note the mean is λ and the variance is λ. To see this, we differentiate the moment generating function and then set $t = 0$:

$$\begin{aligned} \mu &= \frac{dM_X}{dt}\Big|_{t=0} \;=\; \left(\lambda e^t \cdot e^{\lambda(e^t-1)}\right)\Big|_{t=0} \;=\; \lambda \\ \mu_2' &= \frac{d^2M_X}{dt^2}\Big|_{t=0} \\ &= \left(\lambda e^t \cdot e^{\lambda(e^t-1)} + \lambda^2 e^{2t} \cdot e^{\lambda(e^t-1)}\right)\Big|_{t=0} \;=\; \lambda + \lambda^2; \end{aligned} \tag{1.3.26}$$

as

$$\sigma^2 \;=\; \mathbb{E}[(X-\mu)^2] \;=\; \mathbb{E}[X^2] - \mathbb{E}[X]^2, \tag{1.3.27}$$

we see that

$$\sigma^2 \;=\; (\lambda + \lambda^2) - \lambda^2 \;=\; \lambda. \tag{1.3.28}$$

Remark 1.3.8. *Alternatively, we could have found the mean and variance by Taylor expanding the moment generating function. We would have*

$$e^{\lambda(e^t-1)} \;=\; 1 + \lambda(e^t-1) + \frac{(\lambda(e^t-1))^2}{2!} + \frac{(\lambda(e^t-1))^3}{3!} + \cdots. \tag{1.3.29}$$

While at first this looks very complicated, we note that a Taylor expansion of $e^t - 1$ gives $t + t^2/2! + \cdots = t(1 + t/2 + \cdots)$; in other words, $(e^t - 1)^k$ is divisible by t^k. This means

$$\begin{aligned} e^{\lambda(e^t-1)} &= 1 + \lambda t\left(1 + \frac{t}{2} + \cdots\right) + \lambda^2 t^2 \frac{(1 + t/2 + \cdots)^2}{2!} + \lambda^3 t^3 \frac{(1 + t/2 + \cdots)^3}{3!} + \cdots \\ &= 1 + \lambda t + \lambda\frac{t^2}{2} + \lambda\frac{t^2}{2} + \text{terms in } t^3 \text{ or higher} \\ &= 1 + \lambda t + \frac{\lambda^2 t^2}{2} + \cdots . \end{aligned} \tag{1.3.30}$$

Thus, by knowing the Taylor series expansion of e^x, we can find the first two moments through algebra and avoid differentiation; we leave it to the reader to determine which approach they like more (or hate less!).

Theorem 1.3.9. *Let $X, X_1, \ldots, X_N$ be Poisson random variables with parameter λ. Let*

$$\overline{X}_N = \frac{X_1 + \cdots + X_N}{N}, \quad Z_N = \frac{\overline{X}_N - \mathbb{E}[\overline{X}_N]}{\text{StDev}(\overline{X}_N)}. \tag{1.3.31}$$

Then as $N \to \infty$, Z_N converges to having the standard normal distribution.

Proof. We expect $\overline{X}_N$ to be approximately equal to the mean of the Poisson random variable, which in this case is λ. This follows from the linearity of expected value:

$$\mathbb{E}[\overline{X}_N] = \mathbb{E}\left[\frac{X_1 + \cdots + X_N}{N}\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[X_i] = \frac{1}{N}\cdot N\lambda = \lambda. \tag{1.3.32}$$

We will write μ for the mean (and not λ); this keeps the argument a bit more general, and the resulting calculations will look like the general case for a bit longer if we do this.

Let σ^2 denote the variance of the X_n's (Poisson distributions with parameter λ). We know $\sigma = \sqrt{\lambda}$; however, we again choose to write σ below so that these calculations will look a lot like the general case. The variance of $\overline{X}_N$ is computed similarly; since the X_n are independent we have

$$\text{Var}(\overline{X}_N) = \text{Var}\left(\frac{X_1 + \cdots + X_N}{N}\right) = \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}(X_n) = \frac{1}{N^2}\cdot N\sigma^2 = \frac{\sigma}{N}. \tag{1.3.33}$$

As always, the natural quantity to study is

$$Z_N = \frac{\overline{X}_N - \mathbb{E}[\overline{X}_N]}{\text{StDev}(\overline{X}_N)} = \frac{\frac{X_1+\cdots+X_N}{N} - \mu}{\sigma/\sqrt{N}} = \frac{(X_1 + \cdots + X_N) - N\mu}{\sigma\sqrt{N}}. \tag{1.3.34}$$

We now use

$$M_{\frac{X+a}{b}}(t) = e^{at/b}M_X(t/b) \tag{1.3.35}$$

and the moment generating function of a sum of independent variables is the product of the moment generating (Theorem 1.1.19) functions to find the moment generating function of Z_N. We have

$$\begin{aligned} M_{Z_N}(t) &= M_{\frac{(X_1+\cdots+X_N)-N\mu}{\sigma\sqrt{N}}}(t) \\ &= M_{\sum_{n=1}^N \frac{X_n-\mu}{\sigma\sqrt{N}}}(t) \\ &= \prod_{n=1}^N M_{\frac{X_n-\mu}{\sigma\sqrt{N}}}(t) \\ &= \prod_{n=1}^N e^{\frac{-\mu t}{\sigma\sqrt{N}}}\, M_X\left(\frac{t}{\sigma\sqrt{N}}\right) \\ &= \prod_{n=1}^N e^{\frac{-\mu t}{\sigma\sqrt{N}}}\, e^{\mu\left(e^{\frac{t}{\sigma\sqrt{N}}}-1\right)}, \end{aligned} \tag{1.3.36}$$

where in the final step we take advantage of knowing the moment generating function of $M_X(t)$. We now Taylor expand the exponential, using

$$e^u \;=\; \sum_{k=0}^{\infty} \frac{u^k}{k!} \;=\; 1 + u + \frac{u^2}{2!} + \frac{u^3}{3!} + \cdots . \tag{1.3.37}$$

This is one of the most important Taylor expansion we will encounter. Thus the exponential in (1.3.36) is

$$e^{\frac{t}{\sigma\sqrt{N}}} \;=\; 1 + \frac{t}{\sigma\sqrt{N}} + \frac{t^2}{2\sigma^2 N} + \frac{t^3}{6\sigma^3 N\sqrt{N}} + \cdots . \tag{1.3.38}$$

The important thing to note is that after subtracting 1, the first piece is $\frac{t}{\sigma\sqrt{N}}$, the next piece is $\frac{t^2}{2\sigma^2 N}$, and the remaining pieces are dominated by a geometric series (starting with the cubed term) with $r = \frac{t}{\sigma\sqrt{N}}$. Thus, the contribution from all the other terms is of size at most some constant times $\frac{t^3}{N\sqrt{N}}$. For large N this will be negligible, and we write errors like this as $O\left(\frac{t^3}{N\sqrt{N}}\right)$. (This is called big-Oh notation. The technical definition of $f(x) = O(g(x))$ is that there are constants C, x_0 such that whenever $x > x_0$, $|f(x)| \le Cg(x)$. In other words, from some point onward $|f(x)|$ is dominated by a constant times $g(x)$.)

Thus (1.3.36) becomes

$$\begin{aligned} M_{Z_N}(t) &= \prod_{n=1}^N e^{\frac{-\mu t}{\sigma\sqrt{N}}}\, e^{\lambda\cdot\left(\frac{t}{\sigma\sqrt{N}}+\frac{t^2}{2\sigma^2 N}+O\left(\frac{t^3}{N\sqrt{N}}\right)\right)} \\ &= \prod_{n=1}^N e^{\frac{\mu t^2}{\sigma^2 N}+O\left(\frac{t^3}{N\sqrt{N}}\right)} \\ &= e^{\frac{t^2}{2}+O\left(\frac{t^3}{\sqrt{N}}\right)} \end{aligned} \tag{1.3.39}$$

where the last line follows from the fact that we have a product over N identical terms, and as the mean $\mu = \lambda$ and the variance $\sigma^2 = \lambda$ (for X Poisson with parameter λ), we see $\frac{\mu}{\sigma^2} = 1$. Thus, for all t, as $N \to \infty$ the moment generating function of Z_N tends to $e^{\frac{t^2}{2}}$, which is the moment generating function of the standard normal. The proof is completed by invoking Theorem 1.2.18, one of our black-box results from complex analysis, which states that if a sequence of moment generating functions which exist for $|t| < \delta$ converges to a moment generating function of a density, then the corresponding density converges to that density. In our case, this implies convergence to the standard normal. □

Remark 1.3.10. *We only need to Taylor expand far enough to get the main term (which has a finite limit as $N \to \infty$) and then estimate the size of the error term (which tends to zero as $N \to \infty$).*

Even though this was a special case, some features are visible here that will reappear when we consider the general case. Note that the higher moments of the distribution don't seem to matter; all we used was the first and second moments of X. The higher moments *do* matter; their affect is to control the rate of convergence to the standard normal. They are felt in the $e^{O(t^3/\sqrt{N})}$ term.

1.3.5 Proof of the CLT for general sums via MGF

We deliberately kept the proof of Theorem 1.3.9 (normalized sums of independent identically distributed Poisson random variables converges to the standard normal) as general as possible as long as possible for use in proving the full version of the Central Limit Theorem.

Proof of Theorem 1.3.6 (the Central Limit Theorem). Looking at the proof of Theorem 1.3.9, our arguments held for *any* distribution up until the last line of (1.3.36), where we finally used the fact that we had independent Poisson random variables by substituting for $M_X(t/\sigma\sqrt{N})$. This time, we can't substitute a specific expansion for $M_X(t/\sigma\sqrt{N})$ as we don't know M_X. We thus have

$$M_{Z_N}(t) \;=\; \prod_{n=1}^{N} e^{\frac{-\mu t}{\sigma\sqrt{N}}}\, M_X\left(\frac{t}{\sigma\sqrt{N}}\right) \;=\; e^{\frac{-\mu t\sqrt{N}}{\sigma}} M_X\left(\frac{t}{\sigma\sqrt{N}}\right)^N \tag{1.3.40}$$

(as the random variables are identically distributed).

There are several ways to do the algebra to finish the proof; we chose the following approach as it emphasizes one of the most important tricks in mathematics. Namely, whenever you see a product you should *seriously* consider replacing it with a sum. The reason is we have lots of experience evaluating sums. We have formulas for special sums, and using Taylor series we can expand nice functions as sums. We don't really know that many products, or expansions of functions in terms of products.

How do we convert a product to a sum? We know the logarithm of a product is the sum of the logarithms. Thus, let's take logarithms of (1.3.40), and then when we're done analyzing it we just exponentiate. We find

$$\log M_{Z_N}(t) \quad = \quad -\frac{\mu t\sqrt{N}}{\sigma} + N \log M_X\left(\frac{t}{\sigma\sqrt{N}}\right). \tag{1.3.41}$$

Note the first term in the expansion above is of size $\sqrt{N}$ for fixed t; if it isn't cancelled by something from the other term, the limit won't exist. Forunately it is cancelled, and all we will care about is terms up to size $1/N$. We need to be concerned with terms this small because we multiply by N; however, terms of size $1/N^{3/2}$ or smaller won't contribute in the limit as they are only multiplied by N, and thus are still small.

We know

$$M_X(t) \;=\; 1 + \mu t + \frac{\mu_2' t^2}{2!} + \cdots \;=\; 1 + t\left(\mu + \frac{\mu_2' t}{2} + \cdots\right). \tag{1.3.42}$$

We now use the Taylor series expansion for $\log(1+u)$, which is

$$\log(1+u) \;=\; u - \frac{u^2}{2} + \frac{u^3}{3!} - \cdots. \tag{1.3.43}$$

Combining the two gives

$$\begin{aligned} \log M_X(t) &= t\left(\mu + \frac{\mu_2' t}{2} + \cdots\right) - \frac{t^2\left(\mu + \frac{\mu_2' t}{2} + \cdots\right)^2}{2} + \cdots \\ &= \mu t + \frac{\mu_2' - \mu^2}{2} t^2 + \text{terms in } t^3 \text{ or higher.} \end{aligned} \tag{1.3.44}$$

Thus

$$\log M_X\left(\frac{t}{\sigma\sqrt{N}}\right) \;=\; \mu t + \frac{\mu_2' - \mu^2}{2} t^2 + \text{terms in } t^3 \text{ or higher.} \tag{1.3.45}$$

But we do not want to evaluate M_X at t, but rather at $t/\sigma\sqrt{N}$. We find

$$\log M_X\left(\frac{t}{\sigma\sqrt{N}}\right) \;=\; \frac{\mu t}{\sigma\sqrt{N}} + \frac{\mu_2' - \mu^2}{2}\frac{t^2}{\sigma^2 N} + \text{terms in } t^3/N^{3/2} \text{ or lower in } N. \tag{1.3.46}$$

Henceforth we'll denote these lower order terms by $O(N^{-3/2})$, and when we multiply these by N we'll denote the new error by $O(N^{-1/2})$. The entire point of all of this is to simplify (1.3.41), the expansion for $\log M_{Z_N}(t)$. Collecting our pieces, we find

$$\begin{aligned} \log M_{Z_N}(t) &= -\frac{\mu t\sqrt{N}}{\sigma} + N\left(\frac{\mu t}{\sigma\sqrt{N}} + \frac{\mu_2' - \mu^2}{2}\frac{t^2}{\sigma^2 N} + O(N^{-3/2})\right) \\ &= -\frac{\mu t\sqrt{N}}{\sigma} + \frac{\mu t\sqrt{N}}{\sigma} + \frac{\mu_2' - \mu^2}{2}\frac{t^2}{\sigma^2} + O(N^{-1/2}) \\ &= \frac{t^2}{2} + O(N^{-1/2}). \end{aligned} \tag{1.3.47}$$

Why is the last step true? We have $\mu_2' - \mu^2$; this equals $\mathbb{E}[X^2] - \mathbb{E}[X]^2$, which is an alternate way of defining the variance. Thus $\mu_2' - \mu^2 = \sigma^2$, and the claim follows.

So, if $\log M_{Z_N}(t)$ is like $t/2 + O(N^{-1/2})$, then

$$M_{Z_N}(t) \;=\; e^{\frac{t^2}{2}+O(N^{-1/2})}. \tag{1.3.48}$$

Though we took a different route, we end in the same place as in the proof of Theorem 1.3.9. We again appeal to Theorem 1.2.18, one of our black-box results from complex analysis, which states that if a sequence of moment generating functions which exist for $|t| < \delta$ converges to a moment generating function of a density, then the corresponding density converges to that density. In our case, this implies convergence to the standard normal. □

1.4 Fourier Analysis and the Central Limit Theorem

Any theorem as important as the Central Limit Theorem deserves more than one proof. Different proofs emphasize different aspects of the problem. Our first proof was based on properties of the inverse Laplace transform, specifically an appeal to Theorem 1.2.18. The proof below uses the Fourier transform. It is thus very similar to the previous proof, as the Fourier transform and the Laplace transform are equal for certain functions after a change of variable. We choose to present this proof as well for several reasons. **ADD REASONS!**

1.4.1 Needed results from Fourier analysis

Recall the Fourier transform of f is (see Definition 1.2.12)

$$\widehat{f}(y) \;=\; \int_{-\infty}^{\infty} f(x)e^{-2\pi ixy}dx; \tag{1.4.1}$$

(sometimes the Fourier transform is defined with e^{-ixy} or $e^{ixy}/\sqrt{2\pi}$ instead of $e^{-2\pi ixy}$, so always check the convention when you reference a book or use a program such as Mathematica). While $\widehat{f}(y)$ is well defined whenever $\int_{-\infty}^{\infty} |f(x)|dx < \infty$, much more is true for functions with $\int_{-\infty}^{\infty} |f(x)|^2dx < \infty$. Unfortunately, for many applications even this assumption isn't enough, and we need to take f in the Schwartz Space $\mathcal{S}(\mathbb{R})$ (see Definition 1.2.14), the space of all infinitely differentiable functions whose derivatives are rapidly decreasing. One can show the Fourier transform of a Schwartz function is a Schwartz function, and we have the following wonderful theorem on inverting the Fourier transform (Theorem 1.2.15), which for convenience we restate below.

Theorem 1.2.15 (Fourier Inversion Formula). *For $f \in \mathcal{S}(\mathbb{R})$,*

$$f(x) \;=\; \int_{-\infty}^{\infty} \widehat{f}(y)e^{2\pi ixy}dy. \tag{1.4.2}$$

In fact, for any $f \in \mathcal{S}(\mathbb{R})$,

$$\int_{-\infty}^{\infty} |f(x)|^2dx \;=\; \int_{-\infty}^{\infty} |\widehat{f}(y)|^2dy. \tag{1.4.3}$$

Definition 1.4.1 (Compact Support). *A function $f : \mathbb{R} \to \mathbb{C}$ has compact support if there is a finite closed interval $[a, b]$ such that for all $x \notin [a, b]$, $f(x) = 0$.*

Remark 1.4.2 (Advanced). *Schwartz functions with compact support are extremely useful in many arguments. It can be shown that given any continuous function g on a finite closed interval $[a, b]$, there is a Schwartz function f with compact support arbitrarily close to g; i.e., for all $x \in [a, b]$, $|f(x) - g(x)| < \epsilon$. Similarly, given any such continuous function g, one can find a sum of step functions of intervals arbitrarily close to g (in the same sense as above). Often, to prove a result for step functions it suffices to prove the result for continuous functions, which is the same as proving the result for Schwartz functions. Schwartz functions are infinitely differentiable and as the Fourier Inversion formula holds, we can pass to the Fourier transform space, which is sometimes easier to study.*

Example 1.4.3. *Whenever we define a space or a set, it's worthwhile to show that it isn't empty! Let's show there are infinitely many Schwartz functions. We claim the Gaussians $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ are in $\mathcal{S}(\mathbb{R})$ for any $\mu, \sigma \in \mathbb{R}$. By a change of variables, it suffices to study the special case of $\mu = 0$ and $\sigma = 1$. Clearly the standard normal, $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is infinitely differentiable. Its first few derivatives are*

$$\begin{aligned} f'(x) &= -x \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ f''(x) &= (x^2 - 1) \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \\ f'''(x) &= -(x^3 - 3x) \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned} \tag{1.4.4}$$

By induction, we can show that the n^{th}derivative is a polynomial $p_n(x)$ of degree n times $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. To show f is Schwartz, by Definition 1.2.14 we must show

$$\left| (1 + x^2)^m \cdot p_n(x) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right| \tag{1.4.5}$$

is bounded. This follows from the fact that the standard normal decays faster than any polynomial. Say we want to show $|x^m e^{-x^2/2}|$ is bounded. The claim is clear for $|x| \leq 1$. What about larger $|x|$? We know $e^{x^2/2} \geq (x^2/2)^k/k!$ for any k, so $e^{-x^2/2} \leq k! 2^k / x^{2k}$. Thus $|x^m e^{-x^2/2}| \leq k! 2^k / x^{2k-m}$, and if we choose $2k > m$ then this is bounded by $k! 2^k$.

Exercise 1.4.4. *Let $f(x)$ be a Schwartz function with compact support contained in $[-\sigma, \sigma]$ and denote its Fourier transform by $\widehat{f}(y)$. Prove for any integer $A > 0$ that $|\widehat{f}(y)| \leq c_f y^{-A}$, where the constant c_f depends only on f, its derivatives and σ. As such a bound is useless at $y = 0$, one often derives bounds of the form $|\widehat{f}(y)| \leq \frac{\widetilde{c_f}}{(1+|y|)^A}$.*

1.4.2 Convolutions and Probability Theory

An important property of the Fourier transform is that it behave nicely under convolution. Remember we denote the convolution of two functions f and g by $h = f * g$, where

$$h(y) \;=\; \int_{-\infty}^{\infty} f(x)g(y-x)dx \;=\; \int_{I} f(x-y)g(x)dx. \tag{1.4.6}$$

A natural question to ask is: what must we assume about f and g to ensure that the convolution exists? For our purposes, f and g will be probability densities. Thus they are non-negative and integrate to 1. While this is enough to ensure that $h = f * g$ integrates to 1, it is not quite enough to guarantee that $f * g$ is finite. If we assume f and g are square-integrable, namely $\int_{-\infty}^{\infty} f(x)^2 dx$ and $\int_{-\infty}^{\infty} g(x)^2 dx$ are finite, then $f * g$ is well-behaved everywhere. This follows from the Cauchy-Schwarz inequality.

Lemma 1.4.5 (Cauchy-Schwarz Inequality)**.** *For complex-valued functions f and g,*

$$\int_{-\infty}^{\infty} |f(x)g(x)|dx \;\leq\; \left(\int_{-\infty}^{\infty} |f(x)|^2 dx\right)^{1/2} \cdot \left(\int_{-\infty}^{\infty} |g(x)|^2 dx\right)^{1/2}. \tag{1.4.7}$$

Lemma 1.4.6 (Convolutions and the Fourier Transform)**.** *Let f, g be continuous functions on $\mathbb{R}$. If $\int_{-\infty}^{\infty} |f(x)|^2 dx$ and $\int_{-\infty}^{\infty} |g(x)|^2 dx$ are finite then $h = f * g$ exists, and $\widehat{h}(y) = \widehat{f}(y)\widehat{g}(y)$. Thus the Fourier transform converts convolution to multiplication.*

Proof. We first show $h = f * g$ exists. We have

$$\begin{aligned} h(x) &= (f*g)(x) \\ &= \int_{-\infty}^{\infty} f(t)g(x-t)dt \\ |h(x)| &\ll \int_{-\infty}^{\infty} |f(t)| \cdot |g(x-t)|dt \\ &\leq \left(\int_{-\infty}^{\infty} |f(t)|^2 dt\right)^{1/2} \left(\int_{-\infty}^{\infty} |g(x-t)|^2 dt\right)^{1/2} \end{aligned} \tag{1.4.8}$$

by the Cauchy-Schwarz inequality. As we are assuming f and g are square-integrable, both integrals are finite (for x fixed, as t runs from $-\infty$ to ∞ so too does $x - t$).

Now that we know h exists, we can explore its properties. We have

$$\begin{aligned}
\widehat{h}(y) &= \int_{-\infty}^{\infty} h(x)e^{-2\pi ixy}dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)g(x-t)e^{-2\pi ixy}dtdx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)g(x-t)e^{-2\pi i(x-t+t)y}dtdx \\
&= \int_{t=-\infty}^{\infty} f(t)e^{-2\pi ity}\left[\int_{x=-\infty}^{\infty} g(x-t)e^{-2\pi i(x-t)y}dx\right]dt \\
&= \int_{t=-\infty}^{\infty} f(t)e^{-2\pi ity}\left[\int_{u=-\infty}^{\infty} g(u)e^{-2\pi iuy}dx\right]dt \\
&= \int_{t=-\infty}^{\infty} f(t)e^{-2\pi ity}\widehat{g}(y)dt \;=\; \widehat{f}(y)\widehat{g}(y), \qquad (1.4.9)
\end{aligned}$$

where the last line is from the definition of the Fourier transform.

Note that in the argument above we interchanged the order of integration. This is an incredibly common technique, but we must justify it as it is not always possible to switch orders. **ADD STUFF ON JUSTIFYING, GIVE REF, GIVE EXAMPLE WHERE CANNOT.** □

Exercise 1.4.7 (Important)**.** *If for all $i = 1, 2, \dots$ we have f_i is square-integrable, prove for all i and j that $\int_{-\infty}^{\infty} |f_i(x)f_j(x)| < \infty$. What about $f_1*(f_2*f_3)$ (and so on)? Prove $f_1*(f_2*f_3) = (f_1*f_2)*f_3$. Therefore convolution is associative, and we may write $f_1 * \cdots * f_N$ for the convolution of N functions.*

The following lemma is the starting point to the proof of the Central Limit Theorem.

Lemma 1.4.8. *Let X_1 and X_2 be two independent random variables with densities f and g. Assume f and g are square-integrable probability densities, so $\int_{-\infty}^{\infty} f(x)^2dx$ and $\int_{-\infty}^{\infty} g(x)^2dx$ are finite. Then $f * g$ is the probability density for $X_1 + X_2$. More generally, if $X_1, \dots, X_N$ are independent random variables with square-integrable densities p_n, then $p_1 * p_2 * \cdots * p_N$ is the density for $X_1 + \cdots + X_N$. (As convolution is commutative and associative, we don't have to be careful when writing $p_1 * p_2 * \cdots * p_N$.)*

Proof. The probability of $X_i \in [x, x + \Delta x]$ is $\int_x^{x+\Delta x} f(t)dt$, which is approximately $f(x)\Delta x$. The probability that $X_1 + X_2 \in [x, x + \Delta x]$ is just

$$\int_{x_1=-\infty}^{\infty}\int_{x_2=x-x_1}^{x+\Delta x-x_1} f(x_1)g(x_2)dx_2dx_1. \qquad (1.4.10)$$

As $\Delta x \to 0$ we obtain the convolution $f * g$, and find

$$\mathrm{Prob}(X_1 + X_2 \in [a, b]) \;=\; \int_a^b (f * g)(z)dz. \qquad (1.4.11)$$

We must justify our use of the word "probability" in (1.4.11); namely, we must show $f * g$ is a probability density. Clearly $(f * g)(z) \geq 0$ as $f(x), g(x) \geq 0$. As we are assuming f and g are square-integrable,

$$\begin{aligned}\int_{-\infty}^{\infty}(f*g)(x)dx &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x-y)g(y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x-y)g(y)dxdy \\ &= \int_{-\infty}^{\infty}g(y)\left(\int_{-\infty}^{\infty}f(x-y)dx\right)dy \\ &= \int_{-\infty}^{\infty}g(y)\left(\int_{-\infty}^{\infty}f(t)dt\right)dy. \qquad (1.4.12)\end{aligned}$$

As f and g are probability densities, these integrals are 1, which completes the proof. □

Remark 1.4.9. *This section introduced a lot of material and results, but we can now begin to see the big picture. If we take N independent random variables with densities $p_1, \dots, p_N$, then the sum has density $p = p_1 * \cdots * p_N$. While at first this equation looks frightening (what is the convolution of N exponential densities?), there is a remarkable simplification that happens. Using the Fourier transform of a convolution is the product of the Fourier transforms, we find $\widehat{p}(y) = \widehat{p_1}(y) \cdots \widehat{p_N}(y)$; in the special case when the random variables are identically distributed, this simplifies further to just $\widehat{p_1}(y)^N$. Now 'all' (and, sadly, it is a big 'all') we need to do to prove the Central Limit Theorem in the case when all the densities are equal is show that, as $N \to \infty$, $\widehat{p_1}(y)^N$ converges to the Fourier transform of something normally distributed (remember we haven't normalized our sum), and the inverse Fourier transform is uniquely determined and is normally distributed.*

Remark 1.4.10. *It is unusual to have two operations that essentially commute. We have the Fourier transform of a convolution is the product of the Fourier transforms; as convolution is like multiplication, this is saying that using this special type of multiplication, we can switch the orders of the operations. It is rare to have two operations satisfying such a rule. For example, $\sqrt{a+b}$ typically is* not *$\sqrt{a} + \sqrt{b}$.*

We end with the promised proof of the Cauchy-Schwarz inequality.

Proof of the Cauchy-Schwarz inequality. For notational simplicity, assume f and g are non-negative functions. Working with $|f|$ and $|g|$ we see there is no harm in the above assumption. As the proof is immediate if either of the integrals on the right hand side of (1.4.7) is zero or infinity, we assume both integrals are non-zero and finite. Let

$$h(x) = f(x) - \lambda g(x), \quad \lambda = \frac{\int_{-\infty}^{\infty}f(x)g(x)dx}{\int_{-\infty}^{\infty}g(x)^2dx}. \qquad (1.4.13)$$

As $\int_{-\infty}^{\infty} h(x)^2 dx \geq 0$ we have

$$\begin{aligned} 0 &\leq \int_{-\infty}^{\infty} \left(f(x) - \lambda g(x)\right)^2 dx \\ &= \int_{-\infty}^{\infty} f(x)^2 dx \; - \; 2\lambda \int_{-\infty}^{\infty} f(x)g(x)dx \; + \; \lambda^2 \int_{-\infty}^{\infty} g(x)^2 dx \\ &= \int_{-\infty}^{\infty} f(x)^2 dx \; - \; 2\frac{\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2}{\int_{-\infty}^{\infty} g(x)^2 dx} \; + \; \frac{\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2}{\int_{-\infty}^{\infty} g(x)^2 dx} \\ &= \int_{-\infty}^{\infty} f(x)^2 dx \; - \; \frac{\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2}{\int_{-\infty}^{\infty} g(x)^2 dx}. \end{aligned} \tag{1.4.14}$$

This implies

$$\frac{\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2}{\int_{-\infty}^{\infty} g(x)^2 dx} \leq \int_{-\infty}^{\infty} f(x)^2 dx, \tag{1.4.15}$$

or equivalently

$$\left(\int_{-\infty}^{\infty} f(x)g(x)dx\right)^2 \leq \int_{-\infty}^{\infty} f(x)^2 dx \cdot \int_{-\infty}^{\infty} g(x)^2 dx. \tag{1.4.16}$$

Taking square roots completes the proof. □

1.4.3 Proof of the Central Limit Theorem

We can now sketch the proof of the Central Limit Theorem, which for convenience we restate.

Theorem 1.3.6 (Central Limit Theorem). *Let $X_1, \ldots, X_N$ be independent, identically distributed random variables whose moment generating functions converge for $|t| < \delta$ for some $\delta > 0$ (this implies all the moments exist and are finite). Denote the mean by μ and the variance by σ^2, let*

$$\overline{X}_N = \frac{X_1 + \cdots + X_N}{N} \tag{1.4.17}$$

and set

$$Z_N = \frac{\overline{X}_N - \mu}{\sigma/\sqrt{N}}. \tag{1.4.18}$$

Then as $N \to \infty$, the distribution of Z_N converges to the standard normal (see Definition 1.3.5 for a statement).

We highlight the key steps, but we do not provide detailed justifications (which would require several standard lemmas about the Fourier transform; see for example [**?**]). For simplicity, we consider the case where we have a probability density p on $\mathbb{R}$ that has mean zero and variance one, and is of sufficiently rapid decay so that all convolution integrals that arise converge; see Exercise **??**. As we assume the moment generating function converges for $|t| < \delta$, the third moment is finite (we'll use this later in the error analysis). Specifically, let p be an infinitely differentiable function satisfying

$$\int_{-\infty}^{\infty} xp(x)dx \;=\; 0, \;\; \int_{-\infty}^{\infty} x^2p(x)dx \;=\; 1, \;\; \int_{-\infty}^{\infty} |x|^3p(x)dx \;<\; \infty. \tag{1.4.19}$$

Assume $X_1, X_2, \ldots$ are independent identically distributed random variables (i.i.d.r.v.) drawn from p; thus, $\text{Prob}(X_i \in [a,b]) = \int_a^b p(x)dx$. Define $S_N = \sum_{i=1}^N X_i$. Recall the standard Gaussian (mean zero, variance one) is $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

As we are assuming $\mu = 0$ and $\sigma = 1$, we have $Z_N = \frac{(X_1+\cdots+X_N)/N}{1/\sqrt{N}} = \frac{X_1+\cdots+X_N}{\sqrt{N}}$. Let's define $S_N = \sum_{n=1}^N X_n$, so $Z_N = S_N/\sqrt{N}$. We must show $\frac{S_N}{\sqrt{N}}$ converges in probability to the standard Gaussian:

$$\lim_{N\to\infty} \text{Prob}\left(\frac{S_N}{\sqrt{N}} \;\in\; [a,b]\right) \;=\; \frac{1}{\sqrt{2\pi}}\int_a^b e^{-\frac{x^2}{2}}dx. \tag{1.4.20}$$

We sketch the proof. The Fourier transform of p is

$$\widehat{p}(y) \;=\; \int_{-\infty}^{\infty} p(x)e^{-2\pi ixy}dx. \tag{1.4.21}$$

Clearly, $|\widehat{p}(y)| \le \int_{-\infty}^{\infty} p(x)dx = 1$, and $\widehat{p}(0) = \int_{-\infty}^{\infty} p(x)dx = 1$.

Exercise 1.4.11. *One useful property of the Fourier transform is that the derivative of $\widehat{g}$ is the Fourier transform of $2\pi ixg(x)$; thus, differentiation (hard) is converted to multiplication (easy). Explicitly, show*

$$\widehat{g}'(y) \;=\; \int_{-\infty}^{\infty} 2\pi ix \cdot g(x)e^{-2\pi ixy}dx. \tag{1.4.22}$$

If g is a probability density, note $\widehat{g}'(0) = 2\pi i\mathbb{E}[x]$ and $\widehat{g}''(0) = -4\pi^2\mathbb{E}[x^2]$.

The above exercise shows why it is, at least potentially, natural to use the Fourier transform to analyze probability distributions. The mean and variance (and the higher moments) are simple multiples of the derivatives of $\widehat{p}$ at zero. By Exercise 1.4.11, as p has mean zero and variance one, $\widehat{p}'(0) = 0$, $\widehat{p}''(0) = -4\pi^2$. We Taylor expand $\widehat{p}$ (we do not justify that such an expansion exists and converges; however, in most problems of interest this can be checked directly, and this is the reason we need technical conditions about the higher moments of p), and find near the origin that

$$\widehat{p}(y) \quad = \quad 1 + \frac{p''(0)}{2}y^2 + \cdots \;=\; 1 - 2\pi^2y^2 + O(y^3). \tag{1.4.23}$$

Near the origin, the above shows $\widehat{p}$ looks like a concave down parabola.

From §1.4.2, we know

- The probability that $X_1 + \cdots + X_N \in [a, b]$ is $\int_a^b (p * \cdots * p)(z)dz$.
- The Fourier transform converts convolution to multiplication. If $\text{FT}[f](y)$ denotes the Fourier transform of f evaluated at y, then we have

$$\text{FT}[p * \cdots * p](y) \;=\; \widehat{p}(y) \cdots \widehat{p}(y). \tag{1.4.24}$$

However, we do not want to study the distribution of $X_1 + \cdots + X_N = x$, but rather the distribution of $S_N = \frac{X_1+\cdots+X_N}{\sqrt{N}} = x$.

Exercise 1.4.12. *If $B(x) = A(cx)$ for some fixed $c \neq 0$, show $\widehat{B}(y) = \frac{1}{c}\widehat{A}\left(\frac{y}{c}\right)$.*

Exercise 1.4.13. *Show that if the probability density of $X_1 + \cdots + X_N = x$ is $(p * \cdots * p)(x)$ (i.e., the distribution of the sum is given by $p * \cdots * p$), then the probability density of $\frac{X_1+\cdots+X_N}{\sqrt{N}} = x$ is $(\sqrt{N}p * \cdots * \sqrt{N}p)(x\sqrt{N})$. By Exercise 1.4.12, show*

$$\textit{FT}\left[(\sqrt{N}p * \cdots * \sqrt{N}p)(x\sqrt{N})\right](y) \;=\; \left[\widehat{p}\left(\frac{y}{\sqrt{N}}\right)\right]^N. \tag{1.4.25}$$

The previous exercises allow us to determine the Fourier transform of the distribution of S_N. It is just $\left[\widehat{p}\left(\frac{y}{\sqrt{N}}\right)\right]^N$. We take the limit as $N \to \infty$ for **fixed** y. From (1.4.23), $\widehat{p}(y) = 1 - 2\pi^2 y^2 + O(y^3)$. Thus we have to study

$$\left[1 - \frac{2\pi^2 y^2}{N} + O\left(\frac{y^3}{N^{3/2}}\right)\right]^N. \tag{1.4.26}$$

For any fixed y, we have

$$\lim_{N\to\infty}\left[1 - \frac{2\pi^2 y^2}{N} + O\left(\frac{y^3}{N^{3/2}}\right)\right]^N \;=\; e^{-2\pi y^2}. \tag{1.4.27}$$

There are two definitions of e^x; while we normally work with the infinite sum expansion, in this case the product formulation is far more useful:

$$e^x \;=\; \lim_{N\to\infty}\left(1 + \frac{x}{N}\right)^N \tag{1.4.28}$$

(you might recall this formula from compound interest).

Exercise 1.4.14. *Show that the Fourier transform of $e^{-2\pi y^2}$ at x is $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$.* Hint: *This problem requires contour integration from complex analysis.*

We would like to conclude that as the Fourier transform of the distribution of S_N converges to $e^{-2\pi y^2}$ and the Fourier transform of $e^{-2\pi y^2}$ is $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$, then the distribution of S_N equalling x converges to $\frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$. Justifying these statements requires some results from complex analysis.

We refer the reader to [**?**] for the details, which completes the proof. □

The key point in the proof is that we used Fourier Analysis to study the sum of independent identically distributed random variables, as Fourier transforms convert convolution to multiplication. The universality is due to the fact that *only* terms up to the second order contribute in the Taylor expansions. Explicitly, for "nice" p the distribution of S_N converges to the standard Gaussian, independent of the fine structure of p. The fact that p has mean zero and variance one is really just a normalization to study all probability distributions on a similar scale; see Exercise **??**.

The higher order terms are important in determining the *rate* of convergence in the Central Limit Theorem (see [**?**] for details and [**?**] for an application to Benford's Law).

Exercise 1.4.15. *Modify the proof to deal with the case of p having mean μ and variance σ^2.*

Exercise 1.4.16. *For reasonable assumptions on p, estimate the rate of convergence to the Gaussian.*

Exercise 1.4.17. *Let p_1, p_2 be two probability densities satisfying* (1.4.19)*. Consider $S_N = X_1 + \cdots + X_N$, where for each i, X_1 is equally likely to be drawn randomly from p_1* or *p_2. Show the Central Limit Theorem is still true in this case. What if we instead had a fixed, finite number of such distributions $p_1, \dots, p_k$, and for each i we draw X_i from p_j with probability q_j (of course, $q_1 + \cdots + q_k = 1$)?*

1.5 Generating functions, combinatorics and number theory

The purpose of this section is to give a flavor as to the power of generating functions, as well as their drawbacks. Some of these problems, such as the cookie problem, can be solved byu a clever use of combinatorics; generating functions provide another approach. One of the reasons this approach is worth mastering is that it can be readily generalized to situations where we don't know how to do the combinatorics.

1.5.1 The cookie problem through combinatorics

Earlier we considered the following problem: *How many ways are there to divide N identical cookies among P different people?* In other words, we don't care which cookies a given person gets, only how many. The reader may object to seeing this problem in a probability book, arguing that this problem belongs to combinatorics. The reason we include it, however, is that probability is frequently just one step away from combinatorics. To illustrate this point, here are some problems we can ask in probability, once we know the combinatorics of cookie division.

Exercise 1.5.1. *Consider all the ways to divide N cookies among P people. Let there be $W(N, P)$ ways (not surprisingly, the number of ways is a function of N and P). Choose one of these ways uniformly at random; we may represent the choice as a vector of P non-negative integers. For example, if $N = 10$ and $P = 5$ then $W(10, 5) = 1001$, and thus we choose the division $(2, 0, 2, 5, 1)$ with probability $1/1001$.*

Assume $N = 10$ and $P = 5$ below.

1. *What is the probability that someone receives at least 8 cookies?*
2. *In a fair world, everyone would receive two cookies. What is the probability everyone receives at least one cookie?*
3. *Instead of making sure everyone gets at least one cookie, what about making sure that no one gets more than 3 cookies? Or everyone gets at least one but no more than 3 cookies?*

More generally, imagine we now have N cookies and P people. What are the answers? Remember to generalize the other numbers accordingly. Thus instead of making sure everyone gets at least 1 cookie, to be fair we should have each person receive at least $N/P - 1$ cookies. If everyone gets at least N/P cookies and the average number of cookies each person gets is N, we see that this really *confines the number of assignments, and thus this is not the most natural generalization. Let $\sigma(N, P)$ be the standard deviation in the number of cookies each person can expect to receive. The natural generalization is what is the probability everyone gets at least$N/p - k\sigma(N, P)$ for some k.*

Remark 1.5.2. *There are many reasons why we care about problems like the above. For one example, imagine we have a network and we are randomly assigning incoming signals to different routers or processors. It can be quite important to make sure the work is distributed fairly, as we don't want bottlenecks arising from one part being overworked. Do we have to be careful about how we assign? Do we need to keep track of how busy each machine is or is it extremely likely that no part will be overburdened?*

We calculate the probability of an event by looking at how many ways it can happen, and divide by the total number of ways. We see combinatorics is often an indispensable tool for probability, allowing us to determine these numbers. After reviewing the solution to the cookie problem, we'll return to these questions and see which are hard, which are easy.

For the original cookie problem, if N and P are small then we can solve the problem (painfully) through brute force. For example, imagine $N = 10$ and $P = 5$. Let's try to break the analysis down in terms of the maximum number of cookies someone receives. There are $5 = \binom{5}{1}$ ways one person can get all 10 cookies, and then there are $10 = \binom{5}{2}$ ways for one person to get 9 and one person to get one. While we could continue arguing along these lines, it gets complicated very quickly. The next case is when one person gets 8 (there are $\binom{5}{1}$ ways to choose the person who gets 8), and then we must distribute the remaining 2 cookies among the other four people. We can either give one person 2 cookies (there are $4 = \binom{4}{1}$ ways to do this), or give a cookie each to two different people (and there are $6 = \binom{4}{2}$ ways this can be done). Thus the number of ways when the maximum number of cookies someone receives is 8 is $\binom{5}{1} \cdot (\binom{4}{1} + \binom{4}{2}) = 5 \cdot (4 + 6) = 50$. To truly appreciate how painful and unwieldly this method becomes, consider the case where the maximum number of cookies anyone receives is 4. We then have to distribute 6 more cookies, with no one getting more than 4. But it's possible someone else gets 4, or maybe two people each get 3, or three people each get 2, et cetera.

We saw there was a really elegant way of solving this problem (**ADD REF**), and the answer turns out to be $\binom{N+P-1}{P-1}$. We quickly recap the solution. We'll do the case of $N = 10$ and $P = 5$ again to limit the notation, though the argument readily generalizes. The way to count how many ways there are of dividing 10 cookies among 5 people is to imagine that we have 14 cookies in a line and some kind person, say Cookie Monster, who will help us out by eating four of them. (At least the old, non-politically correct Cookie Monster would do this; the newer, politically correct version may claim that cookies are a sometimes food and pass on eating any.) How does Cookie Monster eating four of them help us? We are now left with 10 uneaten cookies and 4 devoured cookies (okay, if they're eaten we're not left with them – we either have the space where they were, or some crumbs). **CAN WE ADD A PICTURE OF COOKIE MONSTER?** The four spaces divide the 10 remaining cookies into five groups; we give all the cookies (if any) up to the first devoured cookie to the first person, then all cookies (if any) between the first and second devoured cookie to the second person, and so on. For example, if Cookie Monster gobbles up cookies 3, 4, 7 and 13 of the 14 cookies,

$$\odot\ \odot\ \otimes\ \otimes\ \odot\ \odot\ \otimes\ \odot\ \odot\ \odot\ \odot\ \odot\ \otimes\ \odot$$

then person 1 receives two cookies, person 2 receives zero, person 3 receives two, person 4 receives five and person 5 receives one cookie. The number of ways of dividing the 10 cookies among 5 people is equivalent to the number of ways of choosing 4 cookies from 14, as each such choice corresponds to a partitioning of the cookies among the people, and of course every choice corresponds to a partitioning. In general, we add $P - 1$ cookies and have to choose $P - 1$ of them to eat, so the answer is $\binom{N+P-1}{P-1}$.

We isolate the solution above, as we'll use it frequently in studying the other problems.

Lemma 1.5.3. *Imagine we have N identical cookies and P people. The number of ways of distributing the N cookies among the P people is $\binom{N+P-1}{P-1}$; we may also interpret this as the number of solutions to $x_1 + \cdots + x_P = N$ with each $x_i \in \{0, 1, 2, \dots\}$.*

The following interpretation of how we solved the cookie problem will be of great use in studying the other questions in Exercise 1.5.1, as well as in generalizing the cookie problem to picky eaters. We are really counting solutions to the equation

$$x_1 + \cdots + x_P \ = \ N, \quad x_i \in \{0, 1, 2, 3, \dots\}. \tag{1.5.1}$$

We have shown the number of solutions is $\binom{N+P-1}{P-1}$.

Let's discuss Exercise 1.5.1. The first part asks what is the probability someone receives at least 8 of the 10 cookies. We know there are $W(10, 5) = \binom{10+5-1}{5-1} = 1001$ ways of distributing the cookies. We could try to solve this by brute force, as there aren't too many possibilities; however, there is a far more concise way that allows us to avoid these tedious computations. What's nice about these numbers is that we don't have to worry about two people receiving at least 8 cookies; the problem would be harder if we asked what is the probability someone receives at least 3 cookies, as there is a real danger of double (or even triple!) counting. There are $\binom{5}{1} = 5$ ways to choose one person to receive 8, 9 or 10 cookies. We then have either 2, 1 or 0 remaining cookies to distribute

among the other four people. By our solution to the cookie problem (Lemma 1.5.3), the number of ways to do this is $\binom{2+4-1}{4-1} = 10$ if there are 2 cookies left, $\binom{1+4-1}{4-1} = 4$ if there is one cookie left, and $\binom{0+4-1}{4-1} = 1$ if there are no cookies left. Thus the number of ways of distributing the 10 cookies so that someone has at least 8 is $5 \cdot (10 + 4 + 1) = 75$. As there are 1001 ways of distributing the cookies, we see the probability that there is a lucky person getting at least 8 cookies is about 7.49%. It's up to you as to whether or not you view this as a likely event; this means that roughly one out of every 13 times we do this, someone gets *a lot* of cookies (and at least three people get none!).

For the second part of Exercise 1.5.1, we want to make sure each of the 5 people gets at least one cookie. This is equivalent to counting how many solutions there are to

$$x_1 + \cdots + x_5 \;=\; 10, \quad x_i \in \{1, 2, 3, \dots\}; \tag{1.5.2}$$

unfortunately, we only know how to solve the above equation when each $x_i \in \{\mathbf{0}, 1, 2, \dots\}$. Amazingly, we get the solution to this new problem from our old for free. We introduce new variables y_i with $y_i = x_i - 1$. Note that as each $x_i \in \{1, 2, 3, \dots\}$, each $y_i \in \{0, 1, 2, \dots\}$. Our equation (1.5.2) becomes

$$(y_1 - 1) + \cdots + (y_5 - 1) \;=\; 10, \quad y_i \in \{0, 1, 2, \dots\}. \tag{1.5.3}$$

Rearranging gives

$$y_1 + \cdots + y_5 \;=\; 5, \quad y_i \in \{0, 1, 2, \dots\}. \tag{1.5.4}$$

This is just another cookie problem, but now we have 5 cookies and 5 people instead of 10 cookies and 5 people. The solution is just $\binom{5+5-1}{5-1} = 126$, which means the probability that everyone gets at least one cookie is $126/1001$, or about 12.6%. Thus there is almost a 90% chance that at least one person will be deprived of cookies. What's particularly nice about this solution is how easily it generalizes to other problems. If we had a pecking order, and wanted to know how many ways there are such that the first person gets at least 3 cookies, the second, fourth and fifth at least 1 and the third at least 2, then we would have $y_1 = x_1 - 3$, $y_2 = x_1 - 1$, $y_3 = x_3 - 2$, $y_4 = x_4 - 1$ and $y_5 = x_5 - 1$. This leads to

$$y_1 + \cdots + y_5 \;=\; 2, \quad y_i \in \{0, 1, 2, \dots\}, \tag{1.5.5}$$

and the answer is just $\binom{2+5-1}{5-1} = 15$.

After our success with the first two parts, it might come as a surprise that the last part is an extremely difficult challenge. We unfortunately don't have a good way of imposing *upper* bounds on the number of cookies someone receives, only *lower* bounds. This is most unfortunate, as this important problem arises in statistical mechanics. **ADD MORE ON THIS** Using the Central Limit Theorem, however, we will be able to approximate the solution to this problem when N and P are large.

The last part asks us to generalize to arbitrary N and P. What is the probability someone receives at least $N - 2$ cookies? Let's assume $N \geq 5$ so we don't have to worry about two or more people getting at least $N-2$ cookies (i.e., there is no danger of double counting). We'll also assume that $P \geq 3$ (as otherwise the problem is straightforward). There are $\binom{P}{1}$ ways to choose the lucky

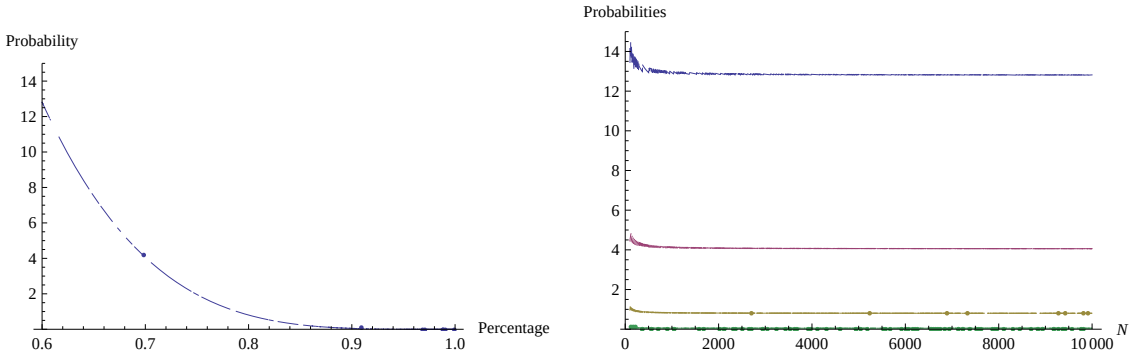

Figure 1.2: Generalizing the cookie problem to calculating the probability one person gets at least $c\%$ of the cookies. In both plots we fix P at 5. The first plot graphs the probabilities as a function of c when $N = 10000$. The second plot graphs the probabilities as N varies, with the upper curve corresponding to $c = 60\%$, the next to $c = 70\%$, then $c = 80\%$ and finally $c = 90\%$ (which is almost indistinguishable from 0).

person. This leaves us with 2 cookies to distribute; there are $\binom{P-1}{1}$ ways to give these two cookies to one person, and $\binom{P-1}{2}$ ways to give them to two people. Thus the probability is

$$\frac{\binom{P}{1}\cdot\left(\binom{P-1}{1}+\binom{P-1}{2}\right)}{\binom{N+P-1}{P-1}} = \frac{P\left(P-1+\frac{(P-1)(P-2)}{2}\right)}{\binom{N+P-1}{P-1}} = \frac{P^2(P-1)}{2\binom{N+P-1}{P-1}}. \tag{1.5.6}$$

Instead of enumerating the two cases, giving us $\binom{P-1}{+}\binom{P-1}{2} = P(P-1)/2$ (after some algebra), we could of course say the number of ways is just the solution to the cookie problem with 2 cookies and $P-1$ people, which is $\binom{2+(P-1)-1}{(P-1)-1} = P(P-1)/2$, which is the same. If P is fixed and $N \to \infty$, we can get a simple upper bound by noting $\binom{N+P-1}{P-1} = \frac{(N+P-1)!}{(P-1)!N!} > N^{P-1}/(P-1)!$. Thus the probability is at most $P(P-1)P!/2N^{P-1}$. For P fixed and N large, this tends to zero rapidly. This approximation isn't too bad, and is easy to use. If we take $N = 100$ and $P = 5$, our approximation says the probability is about .0012%, while the actual answer is approximately .0011%.

An interesting question is to ask about the probability one luck person gets at least $c\%$ of the cookies, with say $c > 1/2$ to simplify life, as $N \to \infty$. Let $\lfloor x \rfloor$ denote the smallest integer greater than or equal to x. The answer would be

$$\frac{\sum_{k=\lfloor cN \rfloor}^{N}\binom{P}{1}\cdot\binom{N-k+(P-1)-1}{(P-1)-1}}{\binom{N+P-1}{P-1}}. \tag{1.5.7}$$

We plot some of these probabilities in Figure 1.2. Looking at the plots in Figure 1.2, it seems that for fixed P and c, as $N \to \infty$ the probability of someone receiving at least $c\%$ converges to a non-zero constant. Can you prove this?

How should we generalize the notion of a fair world? Let's say there are N cookies and P people, and for convenience let's assume P divides N. Thus everyone should get exactly N/P cookies. We would be surprised if everyone got exactly the fair amount; what isn't immediately clear is how far below N/P someone must be before they cry foul. One natural way to solve this is to compute not just the expected number of cookies someone receives, but also the standard deviation. We can do this with binary indicator variables. Let X_n be the random variable which is 1 if the first person receives cookie i and 0 otherwise. As there are P people, we see $X_n = 1$ with probability $1/P$ and 0 with probability $1 - 1/P$. Thus each X_i is a binomial random variable with parameter $p = 1/P$, and thus their expect values are $p = 1/P$ and their variances are $p(1-p) = (P-1)/P^2$. Letting $X = X_1 + \cdots + X_N$, we have

$$\mathbb{E}[X] \;=\; \sum_{n=1}^{N} \mathbb{E}[X_n] \;=\; N \cdot \frac{1}{P} \;=\; \frac{N}{P} \tag{1.5.8}$$

and

$$\mathrm{Var}(X) \;=\; \sum_{n=1}^{N} \mathrm{Var}(X_n) \;=\; N \cdot \frac{P-1}{P^2} \;=\; \frac{(P-1)}{P^2} N. \tag{1.5.9}$$

Thus the standard deviation is $\sqrt{\mathrm{Var}(X)} = \sqrt{(P-1)/P^2}\sqrt{N}$.

Perhaps a reasonable interpretation of making the world fair is that each person should be no worse than one standard deviation below their expected number. What is the probability of this happening if N is large and P is fixed? We now want each person to receive at least

$$\mathfrak{m}(N,P) \;=\; \lfloor N/P - \sqrt{(P-1)/P^2}\sqrt{N} \rfloor \tag{1.5.10}$$

cookies. Thus instead of solving

$$x_1 + \cdots + x_P \;=\; N, \quad x_i \in \left\{ \frac{N}{P} + \mathfrak{m}(N,P), \frac{N}{P} + \mathfrak{m}(N,P) + 1, \frac{N}{P} + \mathfrak{m}(N,P) + 2, \dots \right\}, \tag{1.5.11}$$

as we did previously we make a change of variables by setting $y_i = x_i - \left(\frac{N}{P} - \mathfrak{m}(N,P)\right)$. This leads us to solving

$$y_1 + \cdots + y_P \;=\; P\mathfrak{m}(N,P), \quad y_i \in \{0, 1, 2, \dots\}. \tag{1.5.12}$$

This is just the cookie problem with $N - P\mathfrak{m}(N,P)$ cookies and P people. Thus the probability everyone gets no worse than one standard deviation below their expected number of cookies is

$$\frac{\binom{P\mathfrak{m}(N,P)+P-1}{P-1}}{\binom{N+P-1}{P-1}} \;=\; \frac{\binom{P\lfloor\sqrt{(P-1)/P^2}\sqrt{N}\rfloor+P-1}{P-1}}{\binom{N+P-1}{P-1}}. \tag{1.5.13}$$

If we're willing to accept a little more inequity, we can ask for the probability that each person gets no worse than $\frac{N}{P} - k\mathfrak{m}(N,P)$ cookies (in other words, no one is more than k standard deviations below their expected number). In Figure 1.3 we plot these probabilities for N varying from 1 to

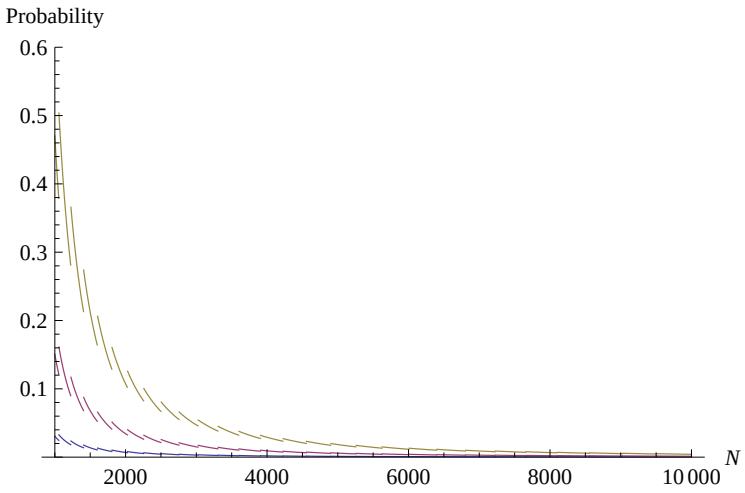

Figure 1.3: Generalizing the cookie problem to calculating the probability everyone receives no worse than k standard deviations below their expected number. The lowest plot is when $k = 2$, the middle is $k = 3$ and the upper is $k = 4$. The zig-zagging nature of our plot is do to the appearance of the floor function, and the fact that we are using $N - P\lfloor N/P \rfloor = 0$.

10000 for 2, 3 and 4 standard deviations. The data suggests that for any fixed P and k, as $N \to \infty$ the probability of everyone being no worse than k standard deviations away tends to zero – in other words, someone is going to be pretty unlucky. Can you prove this?

In conclusion, we see the cookie problem does belong in a probability course. Once we can compute the number of ways of dividing N cookies among P people, we can ask and answer lots of related questions related to the probability of certain assignments. If we distributed the cookies randomly, we can figure out how probable it is that someone will be significantly slighted in getting cookies. As mentioned, there are many examples where we care about such information, such as distributed jobs among computers or routing transmissions through a network.

1.5.2 The cookie problem through generating functions

Let's revisit the cookie problem, this time using the theory of generating functions to find the answer. One of the most difficult steps in using generating functions is figuring out what sequence to use. For many problems in probability, the choice is obvious: let a_m equal the probability our random variable takes on the value m.

Sadly, in the cookie problem it isn't clear what would be a good generating function. Let's assume inspiration strikes us and we decide to consider

$$G(x) \;=\; \sum_{m=0}^{\infty} x^m \;=\; \frac{1}{1-x} \text{ if } |x| < 1. \tag{1.5.14}$$

This is just the generating function of the sequence $\{a_m\}_{m=0}^{\infty}$ with each $a_m = 1$. The generating function can be simplified to $G(x) = (1-x)^{-1}$ for $|x| < 1$ due to the geometric series formula.

We now show how this generating function solves the cookie problem, and then discuss how we can generalize this to solve a variety of more difficult problems. Let $r_{1,s}(N)$ denote the number of solutions to $m_1 + \cdots + m_s = N$ where each m_i is a non-negative integer. Note that $r_{1,s}(N)$ is just the number of ways of dividing N cookies among s people (we switch to denoting the number of people by s and not P as we'll be differentiating with respect to this parameter in a moment); we'll discuss later why we choose to denote this as $r_{1,s}(N)$ and note $r_s(N)$. We claim

$$G(x)^s \;=\; \left(\sum_{m_1=0}^{\infty} x^{m_1}\right)\cdots\left(\sum_{m_s=0}^{\infty} x^{m_s}\right) \;=\; \sum_{N=0}^{\infty} r_{1,s}(N)x^N. \tag{1.5.15}$$

For example, by direct multiplication we see that the first four terms of $G(x)^5$ are $1 + 5x + 15x^2 + 35x^3$, and for $N \in \{0,1,2,3\}$ that the coefficient of x^N in $G(x)^5$ is $\binom{N+5-1}{5-1}$. To prove (1.5.15) for general s and N, we expand the product. We have terms such as $x^{m_1}\cdots x^{m_s}$, which is $x^{m_1+\cdots+m_s} = x^N$ for some N. Assuming everything converges, when we expand the product we obtain x^N many times, once for each choice of $m_1,\ldots,m_s$ that adds to N. Thus the coefficient of x^N in the expansion is $r_{1,s}(N)$. On the other hand, straightforward differentiation shows that

$$G(x)^s \;=\; \left(\frac{1}{1-x}\right)^s \;=\; \frac{1}{(s-1)!}\frac{d^{s-1}}{dx^{s-1}}\frac{1}{1-x}. \tag{1.5.16}$$

Why do we want to write $G(x)^s$ like this? The reason is that we have a nice formula for $G(x)$ as an infinite sum, and differentiating that $s-1$ times is no problem. Note that we are using *both* expressions for $G(x)$ (i.e., we're using the infinite sum formulation as well as the geometric series' answer for what that sum is).

Substituting the geometric series expansion for $\frac{1}{1-x}$ gives

$$G(x)^s \;=\; \frac{1}{(s-1)!}\frac{d^{s-1}}{dx^{s-1}}\sum_{N=0}^{\infty} x^N \;=\; \sum_{N=0}^{\infty}\binom{N+s-1}{s-1}x^N, \tag{1.5.17}$$

which yields $r_{1,s}(N) = \binom{N+s-1}{s-1}$. It is this second method of proof that we generalize. Below we describe a variety of problems and show how to find their generating functions. In most cases, exact formulas such as (1.5.16) are unavailable; we develop sufficient machinery to analyze the generating functions in a more general setting.

Exercise 1.5.4. *With $G(x)$ as in* (1.5.14)*, show by direct multiplication that the first four terms of $G(x)^5$ are $1 + 5x + 15x^2 + 35x^3$, and for $N \le 3$ that the coefficient of x^N in $G(x)^5$ is $\binom{N+5-1}{5-1}$.*

Exercise 1.5.5. *Justify the arguments above. Show all series converge, and prove* (1.5.16) *and* (1.5.17).

1.5.3 The generalized cookie problem

For the original cookie problem, we have a very nice combinatorial perspective that allows us to quickly and cleanly solve it. Let's consider the following twist. We still have N cookies and P people, but now each person wants to have a square number of cookies (the squares are $\{0, 1, 4, 9, 16, 25, \dots\}$). This is significantly harder, as we no longer have a combinatorial interpretation. How do we divide a number N into P squares? Let us know if you find a nice geometric way! Fortunately, there is a solution by using generating functions.

Let $r_{k,s}(N)$ denote the number of ways of writing N as a sum of exactly s integers, where each integer is a k^{th}power. In other words, it is the number of solutions to

$$x_1 + \cdots + x_s \;=\; N, \quad x_i\{0, 1, 2^k, 3^k, 4^k, \dots\} \tag{1.5.18}$$

or equivalently it is the number of solutions to

$$z_1^k + \cdots + z_s^k \;=\; N, \quad z_i \in \{0, 1, 2, 3, 4, \dots\}. \tag{1.5.19}$$

We now see why we denoted the solution to the cookie problem $r_{1,s}(P)$. The solution to the cookie problem in §1.5.2, where we used generating function, immediately generalizes. We have our new generating function is

$$G_k(x) \;=\; \sum_{m=0}^{\infty} x^{m^k}, \tag{1.5.20}$$

and then

$$G_k(x)^s \;=\; \sum_{N=0}^{\infty} r_{k,s}(N)x^n; \tag{1.5.21}$$

unfortunately, if $k \neq 1$ we don't have a simple formula for $G_k(x)$. We don't have an analogue of the geometric series formula to simplify this.

Sadly, this is a very common feature in mathematics. We can reduce the solution to a difficult problem to a difficult sum or integral, which in general we cannot evaluate! All hope is not lost, however, as there are ways to approximate these sums and integrals. To describe these methods in detail is beyond the scope of this book, so we will content ourselves with a brief explanation and some references to the literature. **ADD REFS**.

For technical reasons, it is often convenient to replace x with $e^{2\pi it}$. This converts the problem to one of Fourier analysis. A major pain is the fact that we have infinitely many terms to sum, and this means there are convergence issues. We can fortunately bypass these difficulties with a very simple but quite powerful observation. Let's say we care about how many ways there are to write N as a sum of exactly s numbers that are k^{th}powers; this is known as Waring's problem. Clearly none of the numbers can exceed $N^{1/k}$, so it suffices to look at the truncated generating series

$$G_{k,N}(x) \;=\; \sum_{m=0}^{N^{1/k}} x^{m^k} \quad \text{or} \quad \mathcal{G}_{k,N}(t) \;=\; \sum_{m=0}^{N^{1/k}} e^{2\pi i m^k t}. \tag{1.5.22}$$

We can pull off the solutions by integration. Using

$$\int_0^1 e^{2\pi ikt}dt \;=\; \begin{cases} 1 & \text{if } k=0 \\ 0 & \text{for } k \text{ a non-zero integer} \end{cases} \tag{1.5.23}$$

(one can prove this using either complex analysis, or by writing $e^{i\theta}$ as $\cos\theta + i\sin\theta$), we find

$$r_{k,s}(N) \;=\; \int_0^1 \mathcal{G}_{k,N}(t)e^{-2\pi iNt}dt. \tag{1.5.24}$$

While this *is* the solution to the problem, this is not an easy integral to evaluate! Fortunately there are methods to approximate the function $\mathcal{G}_{k,N}(t)$, which allows us to approximate the integral.

This is just one of many problems that can be solved using this method. One of the most famous problems in number theory is whether or not every 'sufficiently large' even number may be written as the sum of two primes (it is believed that 'sufficiently large' means 'at least 4'). We show that the solution to this problem can be obtained by studying the generating function

$$G_N(x) \;=\; \sum_{p\le N} x^p \tag{1.5.25}$$

where p ranges over primes at most p, or equivalently using the generating function

$$\mathcal{G}_N(t) \;=\; \sum_{p\le N} e^{2\pi ipt}. \tag{1.5.26}$$

To see this, we compute $G_N(t)^2$, which is

$$G_N(t)^2 \;=\; \sum_{p_1\le N} e^{2\pi ip_1t}\sum_{p_2\le N} e^{2\pi ip_2t} \;=\; \sum_{p_1,p_2\le N} e^{2\pi i(p_1+p_2)t} \;=\; \sum_{n=4}^{2N} a_{2,N}(n)e^{2\pi int}, \tag{1.5.27}$$

where $a_{2,N}(n)$ is the number of ways of writing n as the sum of two primes each of which is at most N. Thus we just need to pull off $a_{2,N}(N)$ to solve the problem. We can do this by integrating:

$$\begin{aligned} \int_0^1 G_N(t)^2e^{-2\pi iNt}dt &= \int_0^1 \sum_{n=4}^{N} a_{2,N}(n)e^{2\pi int}e^{-2\pi iNt} \\ &= \sum_{n=4}^{N} a_{2,N}(n)\int_0^1 e^{2\pi i(n-N)t}dt. \end{aligned} \tag{1.5.28}$$

The last integral is 1 if $n = N$ and 0 otherwise, and thus the right hand side becomes $a_{2,N}(N)$.

We have our answer:

$$a_{2,N}(N) \;=\; \int_0^1 G_N(t)^2e^{-2\pi iNt}dt; \tag{1.5.29}$$

unfortunately we can't analyze this integral well enough to solve the problem. We can compute what we believe is the main term, but we cannot show that what is expected to be the error term is smaller. The situation is very different if we try to write N as the sum of three primes; there we *can* show the error term is smaller, and prove that there are *many* ways of writing a large odd number as the sum of three primes.