Math Review for Stat 110

Prof. Joe Blitzstein

1 Sets

> A set is a Many that allows itself to be thought of as a One.
> – Georg Cantor

> Amazon should put their cloud in a cloud, so the cloud will have the redundancy of the cloud.
> – @dowens

A set is a collection of objects. The objects can be anything: numbers, people, cats, courses, even other sets! The language of sets allows us to talk precisely about *events*. If S is a set, then the notation $x \in S$ indicates that x is an element or member of the set S; we can think of the set as a club, with precisely defined criteria for membership. The set may be finite or infinite. If A is a finite set, we write $|A|$ for the number of elements in A, which is called its *cardinality*.

For example:

1. $\{1, 3, 5, 7, \dots\}$ is the set of all odd numbers;
2. {Worf, Jack, Tobey} is the set of Joe's cats;
3. $[3, 7]$ is the closed interval consisting of all real numbers between 3 and 7;
4. $\{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$ is the set of all possible outcomes if a coin is flipped twice (where, for example, HT means the first flip lands Heads and the second lands Tails).

To describe a set (when it's tedious or impossible to list out its elements), we can give a rule that says whether each possible object is or isn't in the set. For example, $\{(x, y) : x \text{ and } y \text{ are real numbers and } x^2 + y^2 \leq 1\}$ is the disc in the plane of radius 1, centered at the origin.

1.1 The empty set

> Bu Fu to Chi Po: "No, no! You have merely painted what is! Anyone can paint what is; the real secret is to paint what isn't."
> Chi Po: "But what is there that isn't?"
> – Oscar Mandel, *Chi Po and the Sorcerer: A Chinese Tale for Philosophers and Children*

> 'Take some more tea,' the March Hare said to Alice very earnestly. 'I've had nothing yet,' Alice replied in an offended tone, 'so I can't take more.'
> 'You mean you can't take less,' said the Hatter: 'It's very easy to take more than nothing.'
> – Lewis Carroll

The smallest set, which is both subtle and important, is the *empty set*, which is the set that has no elements whatsoever. It is denoted by $\emptyset$ or by $\{\}$. Make sure not to confuse $\emptyset$ with $\{\emptyset\}$! The former has no elements, while the latter has one element. If we visualize the empty set as an empty paper bag, then we can visualize $\{\emptyset\}$ as a paper bag inside of a paper bag.

1.2 Subsets

If A and B are sets, then we say A is a subset of B (and write $A \subseteq B$) if every element of A is also an element of B. For example, the set of all integers is a subset of the set of all real numbers. A general strategy for showing that $A \subseteq B$ is to let x be an arbitrary element of A, and then show that x must also be an element of B. A general strategy for showing that $A = B$ for two sets A and B is to show that each is a subset of the other.

1.3 Unions, intersections, and complements

The *union* of two sets A and B, written as $A \cup B$, is the set of all objects that are in A or B (or both). The *intersection* of A and B, written as $A \cap B$, is the set of all objects that are in both A and B. We say that A and B are *disjoint* if $A \cap B = \emptyset$. For n sets $A_1, \ldots, A_n$, the union $A_1 \cup A_2 \cup \cdots \cup A_n$ is the set of all objects that are in *at least one* of the A_j's, while the intersection $A_1 \cap A_2 \cap \cdots \cap A_n$ is the set of all objects that are in *all* of the A_j's.

In many applications, all the sets we're working with are subsets of some set S (in probability, this may be the set of all possible outcomes of some experiment).

When S is clear from the context, we define the *complement* of a set A to be the set of all objects in S that are *not* in A; this is denoted by A^c.

Unions, intersections, and complements can be visualized easily using Venn diagrams, such as the one below. The union is the entire shared region, while the intersection is the football-shaped region of points that are in both A and B. The complement of A is all points in the rectangle that are outside of A.

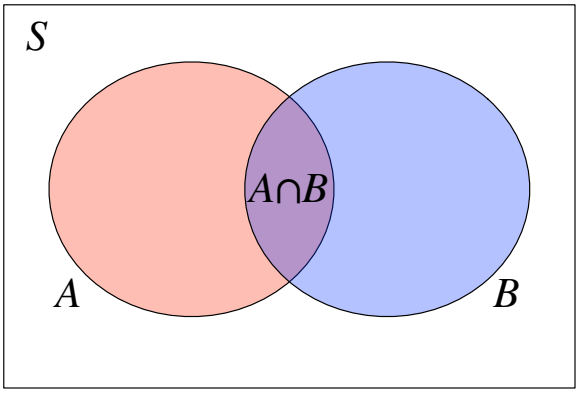

Figure 1: A Venn diagram.

Note that the area of the region $A \cup B$ is the area of A plus the area of B, minus the area of $A \cap B$ (this is a basic form of what is called the *Inclusion-Exclusion principle*).

De Morgan's laws give an elegant, useful duality between unions and intersections:

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$

$$(A_1 \cap A_2 \cap \cdots \cap A_n)^c = A_1^c \cup A_2^c \cup \cdots \cup A_n^c$$

It is much more important to *understand* De Morgan's laws than to memorize them! The first law says that not being in at least one of the A_j is the same thing as not being in A_1, nor being in A_2, nor being in A_3, etc. For example, let A_j be the set of people who like the jth Star Wars prequel (for $j \in \{1, 2, 3\}$). Then $(A_1 \cup A_2 \cup A_3)^c$ is the set of people for whom it is *not* the case that they like at least one of the prequels, but that's the same as $A_1^c \cap A_2^c \cap A_3^c$, the set of people who don't like *The Phantom Menace*, don't like *Attack of the Clones*, and don't like *Revenge of the Sith*.

The second law says that not being in all of the A_j is the same thing as being outside at least one of the A_j. For example, let the A_j be defined as in the previous paragraph. If it is not the case that you like all of the Star Wars prequels (making you a member of the set $(A_1 \cap A_2 \cap A_3)^c$), then there must be at least one prequel that you don't like (making you a member of the set $A_1^c \cup A_2^c \cup A_3^c$), and vice versa.

For practice prove the following facts (writing out your reasoning, not just drawing Venn diagrams):

1. $A \cap B$ and $A \cap B^c$ are disjoint, with $(A \cap B) \cup (A \cap B^c) = A$.
2. $A \cap B = A$ if and only if $A \subseteq B$.
3. $A \subseteq B$ if and only if $B^c \subseteq A^c$.
4. $|A \cup B| = |A| + |B| - |A \cap B|$ if A and B are finite sets.

1.4 Cardinality of sets

As we mentioned before, the cardinality of a finite set is just its number of elements. We can also define cardinality for infinite sets. It can be shown that not all infinities are the same size: some infinite sets are larger than others. In particular, the cardinality of the real numbers is larger than the cardinality of the integers. An infinite set with the same cardinality as the integers is called *countable*, and an infinite set with the same cardinality as the real numbers is called *uncountable*. Any interval of positive length on the real line is uncountable.

1.5 Partitions

A collection of subsets $A_1, \ldots, A_n$ of a set S is a *partition* of S if $A_1 \cup \cdots \cup A_n = S$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. In words, a partition is a collection of disjoint subsets whose union is the entire set. For example, the set of even numbers $\{0, 2, 4, \ldots\}$ and the set of odd numbers $\{1, 3, 5, \ldots\}$ form a partition of the set of nonnegative integers.

2 Functions

> The concept of function is of the greatest importance, not only in pure mathematics but also in practical applications. Physical laws are nothing but statements concerning the way in which certain quantities depend on others when some of these are permitted to vary.
> – Courant, Robbins, and Stewart, *What is mathematics?*

Let A and B be sets. A *function* from A to B is a deterministic rule that, given an element of A as input, provides an element of B as an output. That is, a function

from A to B is a machine that takes an x in A and "maps" it to some y in B. Different x's can map to the same y, but each x only maps to one y. Here A is called the *domain* and B is called the *target*. The notation $f : A \to B$ says that f is a function mapping A into B.

Of course, we have many familiar examples, such as the function f given by $f(x) = x^2$, for all real x. It is important to distinguish between f (the function) and $f(x)$ (the value of the function when evaluated at x). That is, f is a rule, while $f(x)$ is a number for each number x. The function g given by $g(x) = e^{-x^2/2}$ is exactly the same as the function g given by $g(t) = e^{-t^2/2}$; what matters is the rule, not the name we use for the input.

A function f from the real line to the real line is *continuous* if $f(x) \to f(a)$ as $x \to a$, for any value of a. It is called *right continuous* if this is true when approaching from the right, i.e., $f(x) \to f(a)$ as $x \to a$ while ranging over values with $x > a$.

In general though, A needn't consist of numbers, and f needn't be given by an explicit formula. For example, let A be the set of all positive-valued, continuous functions on $[0, 1]$, and f be the rule that takes a function in A as input, and gives the area under its curve (from 0 to 1) as output.

In probability, it is extremely useful to consider functions whose domains are the set of all possible outcomes of some experiment. It may be very difficult to write down a formula for the function, but it's still valid as long as it's defined unambiguously.

2.1 Even and odd functions

Let f be a function whose domain and target are subsets of the real numbers. We say f is an *even function* if $f(x) = f(-x)$ for all x in the domain of f, and we say f is an *odd function* if $-f(x) = f(-x)$ for all x in the domain of f. If neither of these conditions is satisfied, then f is neither even nor odd.

Even and odd functions have nice symmetry properties. The graph of an even function remains the same if you reflect it about the vertical axis, and the graph of an odd function remains the same if you rotate it 180 degrees around the origin. Figure 2 shows the graphs of two even functions and two odd functions.

Even functions have the property that for any a,

$$\int_{-a}^{a} f(x)dx = 2\int_{0}^{a} f(x)dx,$$

assuming the integral exists. This is because the area under the function from $-a$ to 0 is equal to the area under the function from 0 to a. Odd functions have the

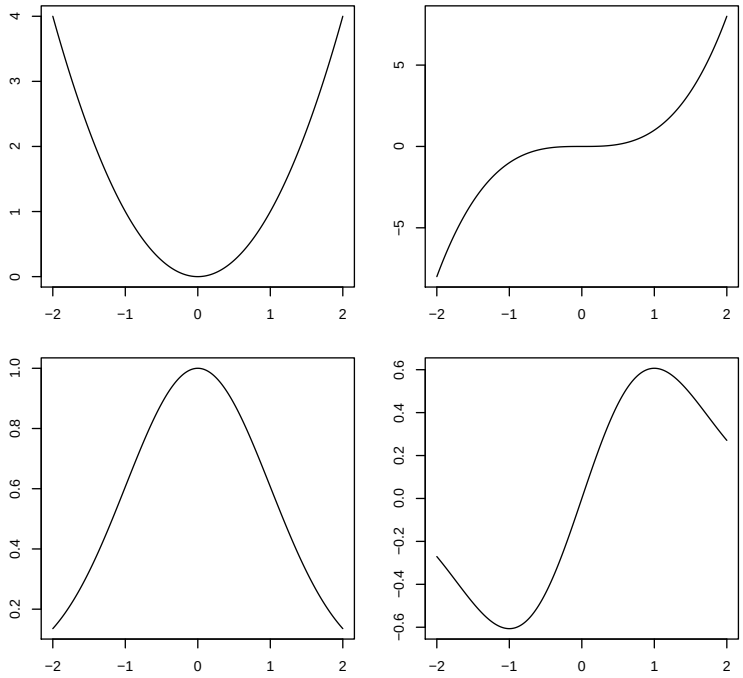

Figure 2: Even and odd functions. The graphs on the left are even functions: $f(x) = x^2$ on the top and $f(x) = e^{-x^2/2}$ on the bottom. The graphs on the right are odd functions: $f(x) = x^3$ on the top and $f(x) = xe^{-x^2/2}$ on the bottom.

property that for any a,

$$\int_{-a}^{a} f(x)dx = 0,$$

again assuming the integral exists. This is because the area under the function from $-a$ to 0 cancels the area under the function from 0 to a.

2.2 Exponential and logarithmic functions

Exponential functions are functions of the form $f(x) = a^x$ for some number $a > 0$. If $a > 1$, the function is increasing, and if $0 < a < 1$, the function is decreasing. The most common exponential function we'll work with is $f(x) = e^x$, and a very useful limit result to know is that

$$\left(1 + \frac{x}{n}\right)^n \to e^x$$

as $n \to \infty$, for any real number x. This has an interpretation in terms of a bank paying compound interest on a deposit: as compounding occurs more and more times per year, the growth rate approaches exponential growth. The case $x = 1$ is sometimes taken as the definition of e.

Here are some properties of exponents:

1. $a^x a^y = a^{x+y}$
2. $a^x b^x = (ab)^x$
3. $(a^x)^y = a^{xy}$

The inverse of an exponential function is a logarithmic function: for positive y, $\log_a y$ is defined to be the number x such that $a^x = y$. Throughout this book, when we write $\log y$ without specifying the base, we are referring to the *natural logarithm* (base e), not the base-10 logarithm.

Here are some properties of logarithms:

1. $\log_a x + \log_a y = \log_a xy$
2. $\log_a x^n = n \log_a x$
3. $\log_a x = \frac{\log x}{\log a}$

2.3 Factorial function

The factorial function takes a natural number n and returns the product of all natural numbers from 1 to n, denoted $n!$ and read "n factorial":

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Furthermore, we define $0! = 1$. A famous approximation for factorials is *Stirling's formula*,

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

The *gamma function* Γ generalizes the factorial function to positive real numbers; it is defined as

$$\Gamma(a) = \int_0^\infty x^a e^{-x} \frac{dx}{x}, \quad a > 0,$$

and satisfies $\Gamma(n) = (n-1)!$ for n a positive integer, and $\Gamma(a+1) = a\Gamma(a)$ for all real $a > 0$. Also, it turns out that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

3 Matrices

> Neo: What is the Matrix?
> Trinity: The answer is out there, Neo, and it's looking for you, and it will find you if you want it to.
> – *The Matrix*

A matrix is a rectangular array of numbers, such as $\begin{pmatrix} 3 & 1/e \\ 2\pi & 1 \end{pmatrix}$ or $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix}$. We say that the dimensions of a matrix are m by n (also written $m \times n$) if it has m rows and n columns (so the former example is 2 by 2, while the latter is 2 by 3). The matrix is called *square* if $m = n$. If $m = 1$, we have a *row vector*; if $n = 1$, we have a *column vector*.

3.1 Matrix addition and multiplication

To *add* two matrices A and B with the same dimensions, just add the corresponding entries, e.g.,

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}.$$

When we *multiply* an m by n matrix A by an n by r matrix B, we obtain an m by r matrix AB; note that matrix multiplication is only well-defined when the number of columns of A equals the number of rows of B. The row i, column j entry of AB is $\sum_{k=1}^{n} a_{ik}b_{kj}$, where a_{ij} and b_{ij} are the row i, column j entries of A and B, respectively. For example, here is how to multiply a 2×3 matrix by a 3×1 vector.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 \cdot 7 + 2 \cdot 8 + 3 \cdot 9 \\ 4 \cdot 7 + 5 \cdot 8 + 6 \cdot 9 \end{pmatrix} = \begin{pmatrix} 50 \\ 122 \end{pmatrix}.$$

Note that AB may not equal BA, even if both are defined. To multiply a matrix A by a scalar, just multiply each entry by that scalar.

The *transpose* of a matrix A is the matrix whose row i, column j entry is the row j, column i entry of A. It is denoted by A' and read as "A transpose" or "A prime". The rows of A are the columns of A', and the columns of A are the rows of A'. If A and B are matrices such that the product AB is defined, then $(AB)' = B'A'$.

The *determinant* of a 2 by 2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is defined to be $ad - bc$. Determinants can also be defined for n by n matrices in a recursive manner not reviewed here.

3.2 Eigenvalues and eigenvectors

An *eigenvalue* of an $n \times n$ matrix A is a number λ such that

$$A\mathbf{v} = \lambda\mathbf{v}$$

for some $n \times 1$ column vector $\mathbf{v}$, where the elements of $\mathbf{v}$ are not all zero. The vector $\mathbf{v}$ is called an *eigenvector* of A, or sometimes a *right eigenvector*. (A *left eigenvector* of A would be a row vector $\mathbf{w}$ satisfying $\mathbf{w}A = \lambda\mathbf{w}$ for some λ.) This definition says that when A and $\mathbf{v}$ are multiplied, $\mathbf{v}$ just gets stretched by the constant λ.

Some matrices have no eigenvalues, but the *Perron-Frobenius theorem* tells us that in a special case that is of particular interest in probability, eigenvalues exist and have nice properties. Let A be a square matrix whose entries are nonnegative and whose rows sum to 1. Further assume that for all i and j, there exists $k \geq 1$ such that the row i, column j entry of A^k is positive. Then the Perron-Frobenius theorem says that 1 is an eigenvalue of A and there are is a corresponding left eigenvector whose entries are all positive.

4 Difference equations

A *difference equation* describes a sequence of numbers in a way that involves the differences between successive numbers in the sequence. For example, the difference equation

$$p_{i+1} - p_i = r(p_i - p_{i-1})$$

describes a sequence of numbers p_i whose differences $a_i \equiv p_i - p_{i-1}$ form a geometric sequence with common ratio r. In fact, there are many such sequences, so this difference equation actually describes an entire collection of sequences.

In this section we will describe how to solve equations of the form

$$p_i = p \cdot p_{i+1} + q \cdot p_{i-1}$$

for some constants p and q; the previous equation was a special case of this form. The first step is to *guess* a solution of the form $p_i = x^i$. Plugging this into the above, we have

$$x^i = p \cdot x^{i+1} + q \cdot x^{i-1},$$

which reduces to $x = px^2 + q$ or $px^2 - x + q = 0$. This is called the *characteristic equation*, and the solution to the difference equation depends on whether the characteristic equation has one or two distinct roots. If there are two distinct roots r_1

and r_2, then the solution is of the form

$$p_i = ar_1^i + br_2^i$$

for some constants a and b. If there is only one distinct root r, then the solution is of the form

$$p_i = ar^i + bir^i.$$

In our case, the characteristic equation has roots 1 and q/p, which are distinct if $p \neq q$ and are both equal to 1 if $p = q$. So we have

$$p_i = \begin{cases} a + b\left(\dfrac{q}{p}\right)^i, & p \neq q, \\ a + bi, & p = q. \end{cases}$$

This is called the *general solution* of the difference equation, since we have not yet specified the constants a and b. To get a *specific solution*, we need to know two points in the sequence in order to solve for a and b.

5 Separable differential equations

Differential equations are the continuous version of difference equations. A differential equation uses derivatives to describe a function or collection of functions. For example, the differential equation

$$\frac{dy}{dx} = 3y$$

describes a collection of functions that have the following property: the instantaneous rate of change of the function at any point (x, y) is equal to cy. This is an example of a *separable* differential equation because we can separate the x's and y's, putting them on opposite sides of the equation:

$$\frac{dy}{y} = 3dx.$$

Now we can integrate both sides, giving $\log y = 3x + c$, or equivalently,

$$y = Ce^{3x},$$

where C is any constant. This is called the *general solution* of the differential equation, and it tells us that all functions satisfying the differential equation are of the form $y = Ce^{3x}$ for some C. To get a *specific solution*, we need to specify one point on the graph, which allows us to solve for C. In general, it may not be possible to rearrange the x's and y's to be on opposite sides, in which case the differential equation is *not* separable, and other techniques are needed to solve it.

6 Partial derivatives

If you can do ordinary derivatives, you can do partial derivatives: just hold all the other input variables constant except for the one you're differentiating with respect to. For example, let $f(x, y) = y \sin(x^2 + y^3)$. Then the partial derivative with respect to x is

$$\frac{\partial f(x,y)}{\partial x} = 2xy \cos(x^2 + y^3),$$

and the partial derivative with respect to y is

$$\frac{\partial f(x,y)}{\partial y} = \sin(x^2 + y^3) + 3y^3 \cos(x^2 + y^3).$$

The *Jacobian* of a function which maps $(x_1, \ldots, x_n)$ to $(y_1, \ldots, y_n)$ is the n by n matrix of all possible partial derivatives, given by

$$\frac{d\vec{y}}{d\vec{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}.$$

7 Multiple integrals

If you can do single integrals, you can do multiple integrals: just do more than one integral, holding variables other than the current variable of integration constant. For example,

$$\begin{aligned} \int_0^1 \int_0^y (x-y)^2 dx dy &= \int_0^1 \int_0^y (x^2 - 2xy + y^2) dx dy \\ &= \int_0^1 (x^3/3 - x^2 y + xy^2)\Big|_0^y dy \\ &= \int_0^1 (y^3/3 - y^3 + y^3) dy \\ &= \frac{1}{12}. \end{aligned}$$

7.1 Change of order of integration

We can also integrate in the other order, $dydx$ rather than $dxdy$, as long as we are careful about the limits of integration. Since we're integrating over all (x, y) with x

and y between 0 and 1 such that $x \leq y$, to integrate the other way we write

$$\begin{aligned}\int_0^1 \int_x^1 (x-y)^2 dydx &= \int_0^1 \int_x^1 (x^2 - 2xy + y^2) dydx \\ &= \int_0^1 (x^2 y - xy^2 + y^3/3)\Big|_x^1 dx \\ &= \int_0^1 (x^2 - x + 1/3 - x^3 + x^3 - x^3/3) dx \\ &= \left(x^3/3 - x^2/2 + x/3 - \frac{x^4}{12}\right)\Big|_0^1 \\ &= \frac{1}{12}.\end{aligned}$$

7.2 Change of variables

In making a change of variables with multiple integrals, a Jacobian is needed. Let's state the two-dimensional version, for concreteness. Suppose we make a change of variables (transformation) from (x, y) to (u, v), say with $x = g(u, v), y = h(u, v)$. Then

$$\int\int f(x,y)dxdy = \int\int f(g(u,v), h(u,v)) \left|\frac{d(x,y)}{d(u,v)}\right| dudv,$$

over the appropriate limits of integration, where $\left|\frac{d(x,y)}{d(u,v)}\right|$ is the absolute value of the determinant of the Jacobian. We assume that the partial derivatives exist and are continuous, and that the determinant is nonzero.

For example, let's find the area of a circle of radius 1. To find the area of a region, we just need to integrate 1 over that region (so any difficulty comes from the limits of integration; the function we're integrating is just the constant 1). So the area is

$$\iint\limits_{x^2+y^2\leq 1} 1 dxdy = \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} 1 dxdy = 2\int_{-1}^1 \sqrt{1-y^2} dy.$$

Note that the limits for the inner variable (x) of the double integral can depend on the outer variable (y), while the outer limits are constants. The last integral can be done with a trigonometric substitution, but instead let's simplify the problem by transforming to polar coordinates: let

$$x = r\cos\theta, y = r\sin\theta,$$

where r is the distance from (x, y) to the origin and $\theta \in [0, 2\pi)$ is the angle. The Jacobian of this transformation is

$$\frac{d(x,y)}{d(r,\theta)} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta, \end{pmatrix}$$

so the absolute value of the determinant is $r(\cos^2\theta + \sin^2\theta) = r$. That is, $dxdy$ becomes $rdrd\theta$. So the area of the circle is

$$\int_0^{2\pi}\int_0^1 rdrd\theta = \int_0^{2\pi}\frac{1}{2}d\theta = \pi.$$

For a circle of radius r, it follows immediately that the area is πr^2 since we can imagine converting our units of measurement to the unit for which the radius is 1.

This may seem like a lot of work just to get such a familiar result, but it served as illustration and with similar methods, we can get the volume of a ball in any number of dimensions! It turns out that the volume of a ball of radius 1 in n dimensions is $\frac{\pi^{n/2}}{\Gamma(n/2+1)}$, where $\Gamma(a) = \int_0^\infty x^a e^{-x}\frac{dx}{x}$ is the *gamma function*.

8 Sums

> 'So you've got to the end of our race-course?' said the Tortoise. 'Even though it does consist of an infinite series of distances? I thought some wiseacre or another had proved that the thing couldn't be done?'
> 'It can be done,' said Achilles; 'It has been done! Solvitur ambulando. You see, the distances were constantly diminishing.'
> – Lewis Carroll

There are two infinite series results that we use over and over again in this book: the geometric series and the Taylor series for e^x.

8.1 Geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \text{ for } |x| < 1 \text{ (this is called a } \textit{geometric series}\text{)}.$$

The series diverges if $|x| \geq 1$. The analogue when the number of terms is finite is

$$\sum_{k=0}^{n} x^k = \frac{1-x^{n+1}}{1-x} \text{ (this is called a } \textit{finite geometric series}\text{)}.$$

8.2 Taylor series for e^x

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x, \text{ for all } x \text{ (this is the } \textit{Taylor series for } e^x\text{)}.$$

8.3 Harmonic series and other sums with a fixed exponent

It is also useful to know that $\sum_{n=1}^{\infty} 1/n^c$ converges for $c > 1$ and diverges for $c \leq 1$. For $c = 1$, this is called the *harmonic series*. The sum of the first n terms of the harmonic series can be approximated using

$$\sum_{k=1}^{n} \frac{1}{k} \approx \log(n) + \gamma$$

for n large, where $\gamma \approx 0.577$.

The sum of the first n positive integers is

$$\sum_{k=1}^{n} k = n(n+1)/2.$$

For squares of integers, we have

$$\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6.$$

For cubes of integers, amazingly, the sum is the square of the sum of the first n positive integers! That is,

$$\sum_{k=1}^{n} k^3 = (n(n+1)/2)^2.$$

8.4 Binomial theorem

The *binomial theorem* states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

where $\binom{n}{k}$ is a *binomial coefficient*, defined as the number of ways to choose k objects out of n, with order not mattering. An explicit formula for $\binom{n}{k}$ in terms of factorials is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

To prove the binomial theorem, expand out the product

$$\underbrace{(x+y)(x+y)\dots(x+y)}_{n \text{ factors}}.$$

Just as $(a+b)(c+d) = ac+ad+bc+bd$ is the sum of terms where we pick the a or the b from the first factor (but not both) and the c or the d from the second factor (but not both), the terms of $(x+y)^n$ are obtained by picking either the x or the y (but not both) from each factor. There are $\binom{n}{k}$ ways to choose exactly k of the x's, and for each such choice we obtain the term $x^k y^{n-k}$. The binomial theorem follows.

9 Pattern recognition

Much of math and statistics is really about *pattern recognition*: seeing the essential structure of a problem, recognizing when one problem is essentially the same as another problem (just in a different guise), noticing symmetry, and so on. We will see many examples of this kind of thinking in this book. For example, suppose we have the series $\sum_{k=0}^{\infty} e^{tk} e^{-\lambda} \lambda^k / k!$, with λ a positive constant. The $e^{-\lambda}$ can be taken out from the sum, and then the structure of the series exactly matches up with the structure of the Taylor series for e^x. Therefore

$$\sum_{k=0}^{\infty} \frac{e^{tk} e^{-\lambda} \lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)},$$

valid for all real t.

Similarly, suppose we want the Taylor series for $1/(1-x^3)$ about $x = 0$. It would be tedious to start taking derivatives of this function. Instead, note that this function is reminiscent of the result of summing a geometric series. Therefore

$$\frac{1}{1-x^3} = \sum_{n=0}^{\infty} x^{3n},$$

valid for $|x^3| < 1$ (which is equivalent to $|x| < 1$). What matters is the structure, not what names we use for variables!

10 Common sense and checking answers

Whenever possible, check whether your answer makes sense intuitively. If it doesn't make sense intuitively, either it is wrong or it is a good opportunity to think harder

and try to explain what's going on. Probability is full of results that seem counterintuitive at first, but which are fun to think about and reward the effort put into trying to understand them. Some useful strategies for checking answers are (a) trying out simple cases, (b) trying out extreme cases, and (c) looking for an alternative method.

For practice, explain what is wrong with each of the following arguments:

1. $$\text{“}\int_{-1}^{1} \frac{1}{x^2} dx = (-x^{-1})\Big|_{-1}^{1} = -2.\text{”}$$

   This makes no sense intuitively, since $1/x^2$ is a *positive quantity*; it would be a miracle if its integral were negative! But where is the mistake?

2. "Let us find $\int \frac{1}{x} dx$ using integration by parts. Let $u = 1/x, dv = dx$. Then

   $$\int \frac{1}{x} dx = uv - \int v du = 1 + \int \frac{x}{x^2} dx = 1 + \int \frac{1}{x} dx,$$

   which implies $0 = 1$."

3. What is wrong with the following "proof" that all horses are the same color? (This example is due to George Pólya, a famous mathematician who wrote the classic problem-solving book *How to Solve It.*) "Let n be the number of horses, and use induction on n to "prove" that in every group of n horses, all the horses have the same color. For the base case $n = 1$, there is only one horse, which clearly must be its own color. Now assume the claim is true for $n = k$, and show that it is true for $n = k + 1$. Consider a group of $k + 1$ horses. Excluding the oldest horse, we have k horses, which by the inductive hypothesis must all be the same color. But excluding the youngest horse, we also have k horses, which again by the inductive hypothesis must have the same color. Thus, all the horses have the same color."

Also, be careful to avoid off-by-one errors, such as thinking that there are $m - n$ numbers in $n, n + 1, \ldots, m$ if n and m are integers with $m \geq n$. This is one of the most common errors in programming, but it is easy to avoid by always checking simple and extreme cases: in the extreme case $m = n$ there is 1 number in the list, not 0, so the length of $n, n + 1, \ldots, m$ is $m - n + 1$.