Math Review for Stat 110

Prof. Joe Blitzstein (Harvard Statistics Department)

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* (see the *Translating Between Probability and Sets* handout). If S is a set, then the notation $x \in S$ indicates that x is an element (a.k.a. member) of the set S (think of the set as a club, with very 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);
5. {Stat 110} is the set of prerequisites for Stat 123.

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. For practice, check that $\emptyset$ is a subset of every set! 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

> I won't use Google+ until I can do arbitrary unions, intersections, and complements of circles.
> – @stat110

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 $A_1 \cup A_2 \cdots \cup A_n$ is the set of all objects that are

in *at least one* of the A_j's, while the intersection $A_1 \cap A_2 \cdots \cap A_n$ 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 sliver of points that are in both A and B. The complement of A is all points in the rectangle that are outside of A.

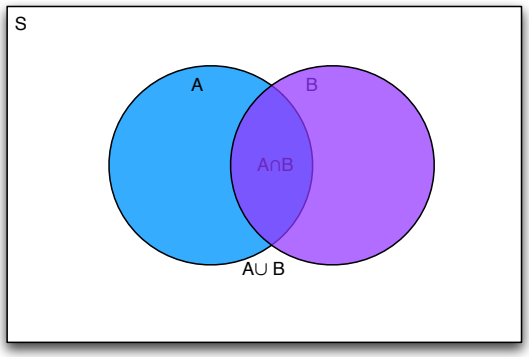

Figure 1: A Venn diagram.

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:

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

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

It is much more important to *understand* De Morgan's laws (why they're true and how to use them) than to *memorize* them! The first 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. For example, let A_j be the set of all 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*.

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

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.

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. A key question to think about then is *where does the randomness come from?* – after all, a function is deterministic!

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 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 $m = n$.

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

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

To *multiply* an m by n matrix A by an n by r matrix B, obtaining an m by r matrix AB (for this to be well-defined, the number of columns of A must equal 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. 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 *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).

4 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

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

and the partial derivative with respect to y is

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

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

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

5 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,

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

5.1 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

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

5.2 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 trig 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

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

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*, a very famous function which we will need later in the course.

6 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 Stat 110: the geometric series, and the Taylor series for e^x.

6.1 Geometric Series

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

6.2 Taylor Series for e^x

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

6.3 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

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

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,

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

6.4 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. We have

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

We'll see in the course that the binomial theorem is closely related to something called the *Binomial distribution*.

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.

7 A Useful Limit

One of the most useful limit results 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 time per year, the growth rate approaches exponential growth. The case $x = 1$ is sometimes taken as the definition of e.

8 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, ... We will see many examples of this kind of thinking in this course. To take a quick math review 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. So we immediately have

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

valid for all real t.

Similarly, suppose we want the Taylor series for $1/(1-x^3)$ (about 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. We then immediately have

$$1/(1-x^3) = \sum_{n=0}^{\infty} x^{3n},$$

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

9 Common Sense and Checking Answers

Whenever possible, check whether your answers make 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 counter-intuitive at first, but which are fun to think about and reward the effort put into trying to understand them. Even if an answer seems plausible, it should be checked whenever possible. Some useful strategies for checking answers, each of which we will see examples of throughout the course, 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. $$\text{“}\int_{-1}^{1} \frac{1}{x^2} dx = (-x^{-1})|_{-1}^{1} = -2.\text{”}$$

   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 also 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 $m \geq n$. Note that in the extreme case $m = n$ there is 1 number in the list. Few people would make the mistake of saying that there are $n - 1$ numbers in $1, 2, \ldots, n$, yet it is very common to make the same mistake when, for example, the sequence starts with 0. This is also one of the most common and insidious programming blunders. But it is easy to avoid by keeping it in mind and always checking simple and extreme cases.