A Guided Tour of Sets, Functions, and Random Variables

Hemant D. Tagare

January 23, 2013

> Dongshan asked a monk, "What is the most painful thing in the world?"
> The monk said, "Hell is the most painful."
> Dongshan said, "Not so. If you wear monk's robes, and underneath, you have not clarified the great matter, that's the most painful thing."
>
> - *Essential Zen*, K. Tanahashi and T. D. Schneider, 1994.

> But what were his solutions? Enigmas of a higher order.
>
> - *Ka*, Roberto Calasso, 1998.

1 Introduction

This note discusses elementary set theory, function theory, and random variables. I have defined everything as clearly as I can, but there is one problem that is not easy to resolve: a mathematically consistent definition of a random variable requires very high powered mathematics called *measure theory*. On the other hand, most of the applications we are interested in can be handled without measure theory. So I will introduce random variables informally without measure theory. This requires you to accept some intuitive "definitions", but there aren't too many of them, and you have probably seen most of them before if you have taken an undergraduate course in probability.

In contrast, I will present set and function theory more rigorously. You will not need fancy mathematics to understand this; rather the challenge is to change your point of view regarding some familiar things.

2 Set Theory

Informally, a *set* is a well-defined collection of distinct mathematical objects, which are called the *elements* of the set. This informal definition is sufficient for us.

Some comments:

1. By "well-defined" we mean that given any mathematical object, we can unambiguously decide whether or not the object is in the set.

   If an object a is contained in a set A, we write $a \in A$. If a is not contained in A, we write $a \notin A$.

2. All elements of the set are distinct.

3. There are two common ways of defining sets:

   (a) Exhaustive listing: Define a set by listing all of its elements, e.g. $A = \{a, b, c, d\}$. Such definitions are common when the set contains a finite number of elements.

   (b) Definition by property: Define a set by the properties its elements must have, e.g. $A = \{x \mid 3x^2 + 2x - 1 = 0\}$ is the set of all numbers x that satisfy the property that $3x^2 + 2x - 1 = 0$.

4. The empty set $\emptyset$ has no elements, $\emptyset = \{\}$.

A *finite set* is a set with a finite number of elements. An *infinite set* is a set with an infinite number of elements.

Standard notation for commonly used sets:

$$\begin{aligned}
\mathcal{N} &= \text{the set of natural numbers } = \{1, 2, 3, \ldots\} \\
\mathcal{N}_0 &= \text{the set of non-negative integers } = \{0, 1, 2, \ldots\} \\
\mathcal{Z} &= \text{the set of integers } = \{\ldots, -1, 0, 1, \ldots\} \\
\mathcal{Q} &= \text{the set of rational numbers} \\
\mathcal{R} &= \text{the set of real numbers} \\
\mathcal{R}^n &= \text{the set of ordered n-tuples of real numbers .}
\end{aligned}$$

The set $\mathcal{R}^n$ deserves special attention, since much of what we will be doing will be in $\mathcal{R}^n$.

2.1 $\mathcal{R}^n$

An element x of $\mathcal{R}^n$ is often written as $(x_1, \cdots, x_n)$. We can also think of an element of $\mathcal{R}^n$ as a column vectors with n components or as a row vector with n components, such as

$$\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} \text{ or } [x_1 \; x_2 \; \cdots \; x_n].$$

$\mathcal{R}$ is identical as $\mathcal{R}^1$. We will visualize $\mathcal{R}$ as a line, $\mathcal{R}^2$ as a plane, etc.

The important thing about $\mathcal{R}^n$ is that besides being a set, it is a vector space. That is, we can think of each element x of $\mathcal{R}^n$ as an $n-$dimensional vector. The vector space operations of addition, scalar multiplication, inner product are defined as:

1. **Addition:** If $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$, then $x + y = (x_1 + y_1, \cdots, x_n + y_n)$.

2. **Scalar Multiplication:** If $x = (x_1, \cdots, x_n)$ and α is a real-number (i.e. a scalar), then $\alpha x = (\alpha x_1, \cdots, \alpha x_n)$.

3. **Inner product:** If $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$, then the *inner product* of x and y, denoted $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$. If the elements of $\mathcal{R}^n$ are written as column vectors then $\langle x, y \rangle = x^T y$, and if the elements of $\mathcal{R}^n$ are written as row vectors then $\langle x, y \rangle = x y^T$.

4. **Norm:** The norm of an element x of $\mathcal{R}^n$ is written as $\|x\|$ and is defined as $\|x\| = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}$. By the Pythagorian formula, we may interpret $\|x\|$ as the distance from the orgin of $\mathcal{R}^n$ to the tip of the vector x.

If $a, b \in \mathcal{R}^n$, then $a < b$ means that all n-components of a are less than the corresponding components of b. That is, if $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n)$, then $a < b$ implies $a_i < b_i,\ i = 1, \ldots, n$. Repeating this idea we can define what $a \leq b$, $a \geq b$, and $a > b$ mean (I am not repeating the definitions to save space; they are obvious). The meaning of $a = b$ is standard. Of course, given any two elements of $\mathcal{R}^n$ none of these relations might hold. The only exception is in $\mathcal{R}^1$, when at least one of these relations always holds for any two elements.

If $a \in \mathcal{R}^n$, then we will take $a \to \infty$ to mean that all components of a go to infinity. Similarly, $a \to -\infty$ means that all components of a go to minus infinity.

2.2 Images as elements of a set

A critical step in carrying out applied mathematical research is to select a mathematical object to represent the physical thing we want to analyze. Depending on the kind of analysis we want to conduct (e.g. which property of the physical object we want to emphasize) the same physical object can be represented by very different mathematical objects. Here we set out to represent an image as an element of $\mathcal{R}^k$.

Informally speaking a digital image is a finite grid of pixels and each pixel has value (also called its intensity). Suppose the grid is two dimensional and has $n \times m$ pixels. Then we can scan the grid from top left to bottom right in raster fashion and stack all the intensity values in order as an $k = n \times m$ column vector. Conversely, given a $k = n \times m$ long column vector we can sequentially insert the components of the vector in the grid in raster fashion to get an image.

What we have achieved in doing this is the following: We have identified images as elements of $\mathcal{R}^k$. This identification has some interesting properties: Each image is a unique element of $\mathcal{R}^k$, and every element of $\mathcal{R}^k$ is an image. Further, the set of all possible images is identical to the set of all elements of $\mathcal{R}^k$, and "similar" elements of $\mathcal{R}^k$ are images with a "similar" appearance.

On the other hand, this representation of images does not preserve neighborhoods. That is, pixel intensities from neighboring pixels in the grid are not found in neighboring components of its representation as an element of $\mathcal{R}^k$.

Problem: Can you extend this idea to represent a Diffusion Tensor Image (DTI)? At every pixel of a DTI image, we have a 3×3 matrix of numbers instead of a single intensity.

2.3 Subsets and the Power set

To return to set theory, we now develop the notion of a subset and the power set of a set:

Definition: A set A is a *subset* of B if every element of A is an element of B (i.e. $a \in A \Rightarrow a \in B$). The empty set $\emptyset$ is a subset of every set.

"A is a subset of B" is written as $A \subset B$. Some important subsets of $\mathcal{R}$ are:

The *closed interval* denoted $[a, b]$: $[a, b] = \{x \mid a \leq x \leq b\}$.
The *open interval* denoted (a, b): $(a, b) = \{x \mid a < x < b\}$.

The following natural chain of subsets is important:

$$\mathcal{N} \subset \mathcal{N}_0 \subset \mathcal{Z} \subset \mathcal{Q} \subset \mathcal{R}.$$

Definition: The *power set* of a set A is the set of all subsets of A. The power set of A is denoted 2^A.

2.4 Set theoretic operations

You have probably seen the following set theoretic operations before:

1. Intersection: $A \cap B = \{a \mid a \in A \text{ and } a \in B\}$.
2. Union: $A \cup B = \{a \mid a \in A \text{ or } a \in B\}$.
3. Difference: $A \setminus B = \{a \mid a \in A \text{ and } a \notin B\}$.
4. Symmetric Difference: $A \triangle B = (A \setminus B) \cup (B \setminus A)$.

   The symmetric difference of two sets is a set containing elements that only belong to one of sets.
5. Complement: If $A \subset B$, then $\tilde{A} = B \setminus A$.

Two sets A and B are *disjoint* if $A \cap B = \emptyset$.

The following theorem, called deMorgan's theorem will be useful later:

DeMorgan's Theorem: $\widetilde{\cup_n A_n} = \cap_n \widetilde{A}_n$, $\widetilde{\cap_n A_n} = \cup_n \widetilde{A}_n$.

2.5 Partition

Definition: If $A_i, i = 1, \ldots, n$ are subsets of A, i.e. $A_i \subset A$, with the property that $A_i \cap A_j = \emptyset$ for $i \neq j$, and $\cup_i A_i = A$, then the sets A_i are called a *partition* of A.

You should visualize a partition as shown in figure 1 (page 5):

2.6 Cartesian product

We now come to an important idea:

Definition: The *Cartesian product* or simply the *product* of two sets A and B, denoted $A \times B$, is the set of all ordered pairs of elements of A and B, i.e.

$$A \times B = \{(a, b) \mid a \in A, b \in B\}.$$

You should visualize Cartesian products as shown in figure 2 (page 5).

Warning: In general, $A \times B$ is not equal to $B \times A$.

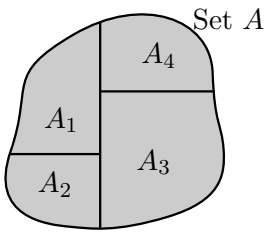


Figure 1: Set A partitioned by A_1, A_2, A_3, A_4

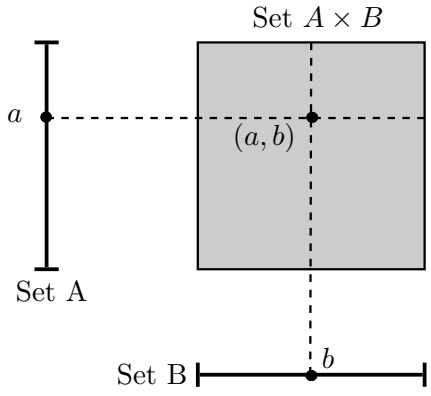


Figure 2: The Cartesian product of two sets

Analogous to the definition of the product of two sets, we can define a product of n sets as

$$A_1 \times A_2 \times \ldots \times A_n = \{(a_1, \ldots, a_n) \mid a_k \in A_k,\ k = 1, \ldots, n\}.$$

Comment: Given a set A we will often take products of the set A with itself, e.g. $A \times A$ or $A \times A \times A \times \ldots \times A$.

If $[a_i, b_i],\ i = 1, \ldots, n$ are intervals in $\mathcal{R}$, then $\prod_{i=1}^{n}[a_i, b_i] = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n]$ is a cube in $\mathcal{R}^n$. If $(-\infty, a_i],\ i = 1, \ldots, n$ are half-infinite intervals, then

$$H_a = \prod_{i=1}^{n}(-\infty, a_i] \tag{1}$$

is an half-infinite cube in $\mathcal{R}^n$.

Finally, $\mathcal{R}^n \times \mathcal{R}^m = \mathcal{R}^{n+m}$.

3 Functions

We want to define functions on arbitrary sets, and not just numbers. If you think of functions as something that is given by a numerical formula, such as $f(x) = x^2$, then I urge you to read this section very carefully and change your point of view.

Given two arbitrary sets A and B, we want to define a function f from set A to set B in the following way: the function takes every element $a \in A$ and associates with it a unique element $b \in B$. By "a unique element $b \in B$" we mean that if the function associates the element b_1 with a, then it cannot associate an element $b_2 \neq b_1$ with the same a. Since associating an element $a \in A$ with an element $b \in B$ is the same as creating the ordered pair (a, b), we have the definition:

Definition: A *function* from set A to set B, denoted $f : A \to B$, is the set of ordered pairs

$$f = \{(a, b) \mid a \in A,\ b \in B,\ (a, b_1) \in f \text{ and } (a, b_2) \in f \text{ only if } b_1 = b_2\}.$$

Some terminology and notation used with functions: We say that the function $f : A \to B$ *maps* the set A to set B. The set A is called the *domain* of the function. The set B is called the *co-domain* or the *range* of the function. We will take the notation $f(a)$ to mean the element b that a is associated with. We call $f(a)$ the "value of the function f at a" or the "evaluation of the function at a". Please do not refer to $f(a)$ as the function. The function is f just by itself, while $f(a)$ is one element in B.

I emphasize that this definition works for any two sets A and B. The elements of these sets **need not be numbers**. The elements can be abstract mathematical objects. Here is one example of this:

Example: Let $A = \{\sin, \cos, \tan\}$ be a set containing ordinary trignometric functions $\sin, \cos, \tan$, and let $B = \{-\sin, \cos, \sec^2\}$ be a set containing ordinary trignometric functions $-\sin, \cos, \sec^2$. Define a function $f : A \to B$ by:

$$f = \{(\sin, \cos), (\cos, -\sin), (\tan, \sec^2)\},$$

so that $f(\sin) = \cos$, $f(\cos) = -\sin$, and $f(\tan) = \sec^2$. You will easily recognize the function f as the familiar "differentiation" of elementary calculus. Differentiation is a function from A to B. **End of example**

It might help your intuition to visualize any function $f : A \to B$ as a table with the first column containing all elements of A, and the second column containing corresponding elements of B:

Element of A	Paired element of B
a_1	b_1
a_2	b_2
$\ldots$	$\ldots$

The requirement that each element of A be paired with a single element of B simply means that each element of A can occur in only one row of the table. An element of B can occur in more than one row.

In some applications, it helps to visualize the function as a slot machine in which we can put any single element of A and from which comes a single element of B:

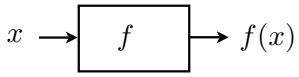


Figure 3: Function

When the domain of a function is a finite set, we often define the function by writing it out as follows:

$$f(a) = \begin{cases} b_1 & \text{if } a = a_1 \\ b_2 & \text{if } a = a_2 \\ \ldots & \ldots \\ b_n & \text{if } a = a_n \end{cases}$$

We will now extend the notation used with functions. Suppose $f : A \to B$ is a function from A to B. Let $A' \subset A$, then we take $f(A')$ to be the subset of B defined by $f(A') = \{b \mid b = f(a), a \in A'\}$.

Similarly, if $B' \subset B$ then we take $f^{-1}(B')$ to be the subset of A defined by $f^{-1}(B') = \{a \mid f(a) \in B'\}$. Note that f^{-1} may or may not be a function (why?).

Finally, note that if the elements of A and B are numbers, then (if we are lucky) we might have a closed form formula for $f(a)$ such as $f(a) = a^2$. This is how we are introduced to functions in high school, but as you can see, it is a very special case.

3.1 Image as a function

Images can also be represented by functions. Suppose that the grid $\mathcal{G}$ has $n \times m$ pixels. Then, we can consider an image as something that associates a number (the pixel intensity) with each pixel. That is, we can model the image as a function $I : \mathcal{G} \to \mathcal{R}$. Thinking of images as functions is a very powerful tool.

3.2 Functions from $\mathcal{R}^n$ to $\mathcal{R}^m$

Suppose $f : \mathcal{R}^n \to \mathcal{R}^m$. Then, f accepts as an input an ordered n-tuple of real numbers (the element of $\mathcal{R}^n$) and produces as an output an ordered m-tuple of real numbers (the element of $\mathcal{R}^m$). Any such function can be regarded as a collection of m functions $f_1, f_2, \cdots, f_m : \mathcal{R}^n \to \mathcal{R}$ as follows: If $y = f(x)$, then

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} f_1(x_1, \cdots, x_n) \\ f_2(x_1, \cdots, x_n) \\ \vdots \\ f_m(x_1, \cdots, x_n) \end{bmatrix} = f(x).$$

The functions $f_1, \cdots, f_m$ are called the *components of the function* f.

If the component functions are differentiable, then the matrix

$$J_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

is the *Jacobian matrix* or simply the *Jacobian* of f. Then $J_f(a)$ denotes the value of J_j at $x = a$.

If f maps $\mathcal{R}^n$ to $\mathcal{R}^n$ (equal dimensions), then J_f is a square matrix. If A is a "small" subset in the first copy of $\mathcal{R}^n$ and $B = f(A)$, and $a \in A$, then

$$\frac{\text{vol } (A)}{\text{vol } (B)} \cong \mid \det J_f(a) \mid, \tag{2}$$

where, vol $(A) = \int_A dx$, vol $(B) = \int_B dy$. The above relation becomes exact in the limit as vol $(A) \to 0$, but is quite useful in the approximate form as well.

3.3 Composition of functions

If $f : A \to B$, and $g : B \to C$, then we can define a function, called the *composition* of f and g, and denoted $g \circ f$, as $g \circ f : A \to C$ given by $(g \circ f)(a) = g(f(a))$ for all $a \in A$. The notation $(g \circ f)(a)$ means the following: the function $g \circ f$ is evaluated at a. The brackets around $g \circ f$ are to suggest that you think of $g \circ f$ as a single function. The composition of functions is illustrated in figure 4.

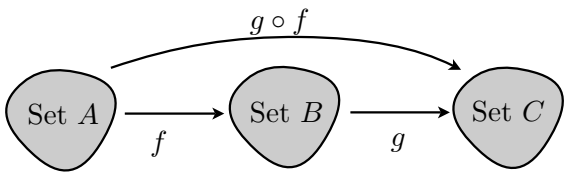


Figure 4: Composition of functions

4 Probability and Random Variables

This section is very intuitive and informal. We begin by adopting the following informal description of randomness:

1. There is an "experiment" that we carry out. Every time we carry out the experiment, we measure a quantity. The measured quantity is an element of $\mathcal{R}^n$.

   Each occurrence of the experiment is called a *trial* of the experiment.

2. Repeated trials of the experiment do not necessarily give the same measurement. This is randomness.

4.1 Random Variable

By a *random variable* x we mean a variable which can take all of the values that occur in infinite trials of the experiment. These values are all in $\mathcal{R}^n$. Suppose A is a subset of $\mathcal{R}^n$. Then by "the probability that a random variable x takes values in A" we mean the following: Repeat the experiment N times and in each trial check whether the value of x is in A. Divide the number of times this happens by N, and take the limit of the ratio as $N \to \infty$. This limit is the probability that x takes values in A, denoted $P\{x \in A\}$:

$$P\{x \in A\} = \lim_{N\to\infty} \frac{\text{no. of times } x \in A}{N}.$$

The set A in the above definition is called an *event*. In a trial if $x \in A$ we say that *event* A *happened*. We also call the probability that x takes values in A the *probability of event* A. This terminology is standard.

Some further comments:

1. If x takes values in $\mathcal{R}^n$, then we say that x is an $n-$dimensional ($n-$dim) random variable.

2. We often simplify the notation $P\{x \in A\}$ to $P\{A\}$ when the random variable x is clear from context. When there is the possibility of ambiguity, we revert back to using $P\{x \in A\}$.

3. The measured value of x in a trial is called a *sample of* x or a *realization of* x. Sometimes the notation $x = a$ is used to indicate that a is a sample of x in an experiment.

Warning: The above is a very simplified version of the standard mathematical definition of a random variable. There are many shortcomings of this intuitive approach. The most significant shortcoming you should be especially aware of is this: measure theory shows that it is not possible to define the probability $P\{A\}$ for <u>all</u> events, i.e. for <u>all</u> subsets $A \subset \mathcal{R}^n$ (any attempt to do so leads to logical inconsistencies). It is <u>only</u> possible to define probabilities of certain events which are called *measurable*. Defining a measurable event is a complicated matter, and we will avoid the definition. Instead we will adopt the following idea: The only events $A \in \mathcal{R}^n$ whose probabilities we will calculate will be those that have a well-defined n-dimensional volume

$$\int_A dx_1 \dots dx_n$$

(sets with volume 0 and ∞ are included). Note that this means events in $\mathcal{R}^1$ must have a length, events in $\mathcal{R}^2$ must have an area, etc. Such sets are measurable, and as long as we stick with them, we have no fear of inconsistency.

A second simplification we made above is the idea that the probability is defined as the limit of a quantity as the number of trials goes to infinity. In complicated problem this definition runs into trouble. Modern probability theory gets around this by adopting an axiomatic approach (due to the Russian mathematician Kolmogorov) to define probability. This too is too complicated for us, and we will proceed by adopting the above simplification. The simplification will suffice for us, because we will not be dealing with complex problems that require the axiomatic approach.

4.2 Diagrams

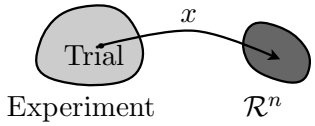


Figure 5: The convention for drawing diagrams

Later in these notes, we will create complex experiments from simple experiments. The definitions of complex experiments will be easier to understand with the help of diagrams. The convention we will adopt for diagrams is the following: We will draw two sets as shown in figure 5. The first set will be the set of all trials (any graphical point in this set will represent a trial), and we will simply call this set the *experiment*. The second set will be the space $\mathcal{R}^n$ and a labeled arrow, say labeled x, from the experiment to $\mathcal{R}^n$ will indicate the value that the random variable x takes for the trial which is the beginning of the arrow.

4.3 A formula for calculating probability

If events $A_1, \ldots, A_n$ partition event A, then

$$P\{A\} = \sum_{i=1}^{n} P\{A_i\}.$$

4.4 Conditional probability

The conditional probability of event A given event B, denoted $P\{A \mid B\}$, is the probability that event A occurs given that event B occurs

$$P\{A \mid B\} = \frac{P(A \cap B)}{P(B)}. \tag{3}$$

This is illustrated in figure 6.

Since $P\{A \cap B\} = P\{A \mid B\} \times P(B)$, and $P\{A \cap B\} = P\{B \cap A\} = P\{B \mid A\} \times P\{A\}$,

$$P\{A \mid B\} = P\{B \mid A\} \times \frac{P\{A\}}{P\{B\}}. \tag{4}$$

This is known as Bayes' Theorem.

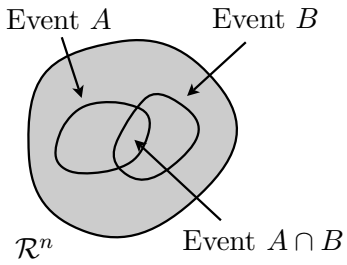


Figure 6: Conditional probability

4.5 Independent Events

Events A and B are *independent* if $P\{A \cap B\} = P\{A\} \times P\{B\}$. For independent events A and B, $P\{A \mid B\} = P\{A\}$ and $P\{B \mid A\} = P\{B\}$. Thus, knowing that one event occurred does not change the probability of the other event occuring. Loosely speaking, we say that independent events do not provide any information about each other.

4.6 Probability density and distribution functions

With every n-dim random variable x we associate a *probability density function* (pdf) $f_x : \mathcal{R}^n \to [0, \infty)$ so that the probability of an event A is obtained as

$$P\{A\} = P\{x \in A\} = \int_A f_x(u)du.$$

The pdf is required to satisfy $\int_{-\infty}^{\infty} f_x(u)du = 1$. We will allow the pdf to contain delta functions.

The strategy to calculate a pdf of an n-dim random variable x from the basic description of the problem is this. Let $a \in \mathcal{R}^n$ be a point and A_a be a small cube centered at a with sides $\Delta x_1, \Delta x_2, \ldots, \Delta x_n$. Let vol $= \Delta x_1 \times \Delta x_2 \times \ldots \times \Delta x_n$. First calculate the probability of the event A_a, i.e. $P\{A_a\}$. Then set

$$f_x(a) = \lim_{\text{vol } (A_a) \to 0} \frac{P\{A_a\}}{\text{vol } (A_a)}.$$

A very useful approximation for calculating $P\{x \in A\}$ is possible when A has a small (infinitesimal) volume, if a is any element of A then,

$$P\{x \in A\} = \int_A f_x(u)du \cong f_x(a) \text{ vol}(A).$$

We will use this approximation frequently below.

We also associate another function with a random variable x. It is called the *cumulative distribution function* (cdf), and is denoted $F_x : \mathcal{R}^n \to [0, 1]$. The value $F_x(a)$ of the cdf at $a \in \mathcal{R}^n$ is the probability that the random variable x falls in the half-infinite cube H_a, i.e.

$$F_x(a) = P\{H_a\} = \int_{H_a} f_x(u)du. \tag{5}$$

That is, $F_x(a)$ is the probability that $x \leq a$.

The cdf has the following properties:

1. The cdf F_x only takes values between 0 and 1,
2. $\lim_{a\to\infty} F_x(a) = 1$,
3. $\lim_{a\to-\infty} F_x(a) = 0$.

Given the pdf, we can calculate the cdf from equation (5). The converse is harder to do, unless the cdf is differentiable. If it is differentiable, then

$$f_x(a) = \frac{\partial^n}{\partial x_1 \dots \partial x_n} F_x \mid_{x=a},$$

where $\frac{\partial^n}{\partial x_1 \dots \partial x_n}$ is the mixed partial derivative of F_x.

From now on, whenever we say "x is a random variable", we will assume that it has a pdf and a cdf.

4.7 Function of a random variable

Suppose that x is an n-dim random variable and $f : \mathcal{R}^n \to \mathcal{R}^m$ is a function. Then $y = f(x)$ is a m-dim random variable – in every trial, y takes the value $f(x)$. The cdf and pdf of y can be calculated from the cdf and pdf of x as follows:

CDF:

The cdf of y is

$$F_y(a) = P\{y \in H_a\} = P\{x \in f^{-1}(H_a)\}.$$

This idea is illustrated in figure 7.

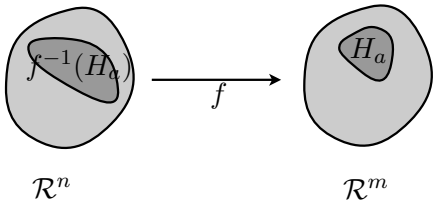


Figure 7: Calculating the cdf of a function of a random variable

Example : Suppose x is a 1-dim random variable, and $f : \mathcal{R}^1 \to \mathcal{R}^1$ is the square function so that $y = f(x) = x^2$. Then, $H_a = (-\infty, a]$ and

$$f^{-1}(H_a) = \begin{cases} \emptyset & \text{if } a < 0 \\ [-\sqrt{a}, \sqrt{a}] & \text{if } a \geq 0. \end{cases}$$

Hence,

$$F_y(a) = \begin{cases} 0 & \text{if } a < 0 \\ F_x(\sqrt{a}) - F_x(-\sqrt{a}) & \text{if } a \geq 0. \end{cases}$$

PDF:

The pdf of y is obtained from the pdf of x as follows. Let $A_a \in R^m$ be a small cube centered at $a \in \mathcal{R}^m$. Assume $f^{-1}(A)$ is an event in $\mathcal{R}^n$. Then

$$f_y(a) = \lim_{\text{vol } (A_a) \to 0} \frac{P_x(f^{-1}(A_a))}{\text{vol } (A_a)}. \tag{6}$$

There is one commonly occurring special case where this formula simplifies. This is the case where $n = m$, i.e. $f : \mathcal{R}^n \to \mathcal{R}^n$, and for every $a \in \mathcal{R}^n$, the set $f^{-1}(a) = \{b_1, \ldots, b_k\}$ is finite (the number of elements of $f^{-1}(a)$ need not be fixed, they can depend on a). Further suppose that J_f, the Jacobian matrix of f, exists everywhere in $\mathcal{R}^n$.

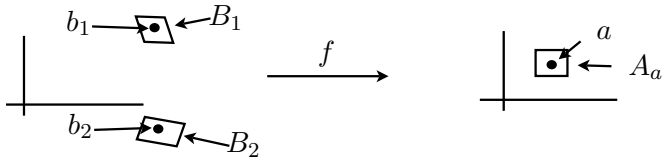


Figure 8: Calculating the pdf of a function of a random variable.

Then we have the situation illustrated in figure 8. Let $J_f(b_i)$ be the $n \times n$ Jacobian matrix at b_i:

$$J_f(b_i) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} (b_i).$$

From elementary multivariate calculus we know the following: For a infinitesimal cube A_a centered at a with sides $\Delta y_1 \Delta y_2 \ldots \Delta y_n$, the set $f^{-1}(A_a)$ consists of k infinitesimal disjoint sets, $f^{-1}(A_a) = B_1 \cup B_2 \cup \ldots \cup B_k$, with $b_i \in B_i$ and $B_i \cap B_j = \emptyset$ for $i \neq j$. Further, vol $(A_a) =$ vol $(B_i) \times \mid \det J_f(b_i) \mid$.

The event $y \in A_a$ is identical to the event that x is in some B_i. Because B_i partition $f^{-1}(A_a)$

$$\begin{aligned} P\{y \in A_a\} &= P\{x \in f^{-1}(A_a)\} \\ &= \sum_{i=1}^{k} P\{B_i\} \\ &\simeq \sum_{i=1}^{k} f_x(b_i) \text{ vol } (B_i) \\ &\simeq \sum_{i=1}^{k} f_x(b_i) \frac{\text{vol } (A_a)}{\mid \det J_f(b_i) \mid}. \end{aligned} \tag{7}$$

Thus,

$$\frac{P\{y \in A_a\}}{\text{vol } (A_a)} \simeq \sum_{i=1}^{k} \frac{f_x(b_i)}{\mid \det J_f(b_i) \mid}.$$

This equation becomes exact in the limit as vol $(A_a) \to 0$. Thus

$$f_y(a) = \lim_{\text{vol }(A_a)\to 0} \frac{P\{y \in A_a\}}{\text{vol }(A_a)} = \sum_{i=1}^{k} \frac{f_x(b_i)}{\mid \det J_f(b_i) \mid}. \tag{8}$$

To use this formula the variables b_i on the right hand side have to be expressed in terms of the variable a. The following example illustrates this:

Example : Let's calculate the pdf for the square function. Since $y = f(x) = x^2$. $J_f(b) = (\frac{df}{dx})(b) = 2b$, and $\mid \det J_f(b) \mid = 2 \mid b \mid$.

If $a < 0$, then $f^{-1}(a)$ is empty. If $a = 0$, then $f^{-1}(a) = \{0\}$, and if $a > 0$, then $f^{-1}(a) = \{-\sqrt{a}, \sqrt{a}\}$. Thus, using equation (8):

$$f_y(a) = \begin{cases} 0 & \text{if } a < 0 \\ \infty & \text{if } a = 0 \\ \frac{f_x(-\sqrt{a}) + f_x(\sqrt{a})}{2\sqrt{a}} & \text{otherwise.} \end{cases}$$

Example : Let $f : \mathcal{R}^n \to \mathcal{R}^n$ be defined by $f(x) = Ax$, where A is an $n \times n$ invertible matrix. That is,

$$\begin{aligned} f(x) &= (f_1(x_1, \ldots, x_n) \ , \ \ldots \ , \ f_n(x_1, \ldots, x_n)) \\ &= (a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \ , \ \ldots \ , \ a_{n1}x_1 + a_{n2}x_2 + \ldots a_{nn}x_n). \end{aligned}$$

Thus,

$$\begin{aligned} J_f &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \\ &= A. \end{aligned}$$

Further, because A is an invertible matrix, $f^{-1}(v) = A^{-1}v$, so that

$$f_y(v) = \frac{f_x(A^{-1}v)}{\mid \det A \mid}.$$

4.8 Joint random variables and marginalization

We now arrive at an important concept. Suppose that at each trial of the experiment we measure two quantities instead of one. Now, we have two random variables, say x and y, and they take values together in every trial.

There is a very convenient way of regarding such random variables as a single random variable. Suppose the x is n-dim and y is m-dim. Then, the ordered pair x, y is n+m-dim, We regard the ordered pair x, y as a single random variable called the *joint random variable* x, y. The joint random variable has cdf and pdf $F_{x,y}$ and $f_{x,y}$. These are the called the *joint cdf* and *joint pdf* respectively.

To understand the relation between the joint and individual cdfs, it is convenient to relate the events in $\mathcal{R}^n$ and $\mathcal{R}^m$ with events in $\mathcal{R}^{n+m}$:

1. An event $A \subset \mathcal{R}^n$ and $B \subset \mathcal{R}^m$ defines the event $A \times B \subset \mathcal{R}^{n+m}$. This is illustrated in figure 9. Of course, not every event in $\mathcal{R}^{n+m}$ is defined this way.

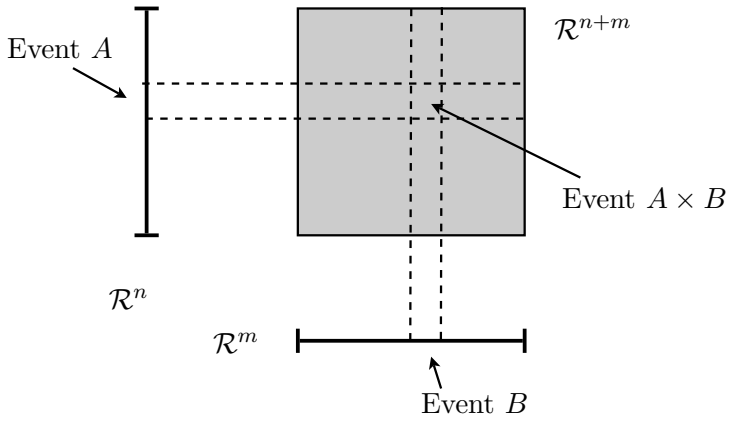


Figure 9: The event $A \times B$ in $\mathcal{R}^{n+m}$.

2. If A is an event in $\mathcal{R}^n$, then $x \in A$ if and only if $(x, y) \in A \times \mathcal{R}^m$. One consequence of this is that in repeated trials, the number of times $x \in A$ is identical to the number of times $(x, y) \in A \times \mathcal{R}^m$, so that $P\{x \in A\} = P\{(x, y) \in A \times \mathcal{R}^m\}$.

3. Similarly, if B is an event in $\mathcal{R}^m$ then, $y \in B$ if and only if $(x, y) \in \mathcal{R}^n \times B$, and similar to the above, $P\{y \in B\} = P\{(x, y) \in \mathcal{R}^n \times B\}$.

We can now relate the cdfs of the joint random variable to the cdfs of the individual random variables

$$F_x(a) = P\{x \in H_a\} = P\{(x, y) \in H_a \times \mathcal{R}^m\} = F_{x,y}(a, \infty), \tag{9}$$
$$F_y(a) = P\{y \in H_a\} = P\{(x, y) \in \mathcal{R}^n \times H_a\} = F_{x,y}(\infty, a). \tag{10}$$

A similar argument (which we skip for now) also relates the pdf of the joint random variable, to the pdfs of the individual random variables:

$$f_x(a) = \int_{-\infty}^{\infty} f_{x,y}(a, v)dv, \tag{11}$$
$$f_y(a) = \int_{-\infty}^{\infty} f_{x,y}(u, a)du. \tag{12}$$

We refer to this calculation of cdfs and pdfs of x and y from the joint cdf and pdf of x, y as *marginalization*, and the cdfs and pdfs of x and y calculated from the joint cdf and pdf as the *marginal cdf*s and the *marginal pdf*s.

4.9 Independent random variables

Let x, y be joint random variables with x n-dim and y m-dim. Then, x and y are *independent random variables* if for every event $A \in \mathcal{R}^n$ and $B \in R^m$,

$$P\{A \times B\} = P\{A\} \times P\{B\}. \tag{13}$$

The probability on the left hand side is of this equation is calculated using the joint random variable x, y and the probabilities on the right hand are calculated using the random variables x and y individually.

It is easy to show that the random variables x and y are independent if and only if

$$\begin{aligned} F_{x,y}(a,b) &= F_x(a)F_y(b), \text{ and} \quad (14) \\ f_{x,y}(a,b) &= f_x(a)f_y(b) \quad (15) \end{aligned}$$

for all a, b. Thus we say that two random variables are independent if and only if their joint cdf (pdf) is a product of their marginal cdfs (pdfs).

4.10 Conditioning of random variables

If x and y are n-dim and m-dim random variables and $A \subset \mathcal{R}^n$ and $B \subset \mathcal{R}^m$ are events, then the conditional probability of A given B is $P\{A \mid B\}$ given by

$$P\{A \mid B\} = \frac{P\{A \times B\}}{P\{B\}}, \quad (16)$$

which is short form for

$$P\{x \in A \mid B\} = \frac{P\{(x,y) \in A \times B\}}{P\{y \in B\}}. \quad (17)$$

Recall some notation. In $\mathcal{R}^n$, the notation $u \leq w$ means that all n components of u are less than or equal to the corresponding components of w, i.e. $u_i \leq w_i$, for $i = 1, \cdots, n$. Setting $\Delta a = (\Delta a_1, \cdots, \Delta a_n)$, the inequality $a \leq x \leq a + \Delta a$ represents the set of inequalities $a_i \leq x_i \leq a_i + \Delta a_i$ which is a cube in $\mathcal{R}^n$, and thus an event in $\mathcal{R}^n$. Similarly, $b \leq y \leq b + \Delta b$ is an event in $\mathcal{R}^m$. The conditional probability of the event $a \leq x \leq a + \Delta a$ given the event $b \leq y \leq b + \Delta b$ is

$$\begin{aligned} P\{a \leq x \leq a + \Delta a \mid b \leq y \leq b + \Delta b\} &= \frac{P\{a \leq x \leq a + \Delta a, b \leq y \leq b + \Delta b\}}{P\{b \leq y \leq b + \Delta b\}} \\ &\approx \frac{f_{x,y}(a,b) \text{ vol } (\Delta a) \text{ vol } (\Delta b)}{f_y(b) \text{ vol } (\Delta b)}. \end{aligned}$$

Thus,

$$\frac{P\{a \leq x \leq a + \Delta a \mid b \leq y \leq b + \Delta b\}}{\text{vol } (\Delta a)} \approx \frac{f_{x,y}(a,b) \text{ vol } (\Delta a) \text{ vol } (\Delta b)}{f_y(b) \text{ vol } (\Delta a) \text{ vol } (\Delta b)},$$

and,

$$\lim_{\text{vol } (\Delta a) \to 0, \text{ vol } (\Delta b) \to 0} \frac{P\{a \leq x \leq a + \Delta a \mid b \leq y \leq b + \Delta b\}}{\Delta a} = \frac{f_{x,y}(a,b)}{f_y(b)}.$$

The quantity on the right-hand side of the above equation is called the *conditional pdf* of x given y and denoted $f_{x|y}(a \mid b)$, so that

$$f_{x|y}(a \mid b) = \frac{f_{x,y}(a,b)}{f_y(b)} = \frac{f_{x,y}(a,b)}{\int_{-\infty}^{\infty} f_{x,y}(u,b)du}. \quad (18)$$

If A is any event in $\mathcal{R}^n$, then the probability that x takes values in A given that $y = b$ is

$$P\{A \mid \{b\}\} = \int_A f_{x|y}(u \mid b)du. \tag{19}$$

Thus the conditional cdf $F_{x|y}$ is defined by

$$F_{x|y}(a \mid b) = \int_{H_a} f_{x|b}(u, b)du. \tag{20}$$

Using equation (18) we obtain:

$$f_{y|x}(b \mid a) = f_{x|y}(a \mid b) \times \frac{f_y(b)}{f_x(a)}. \tag{21}$$

This formula is called *Bayes rule* and is probably the most important formula in this course.

One final comment. If x and y are independent random variables, then

$$f_{x|y}(a \mid b) = \frac{f_{x,y}(a, b)}{f_y(b)} = \frac{f_x(a)f_y(b)}{f_y(b)} = f_x(a). \tag{22}$$

Similarly, $f_{y|x}(b \mid a) = f_y(b)$.

Example: Suppose x and y are n-dim random variables which are observed simultaneously in an experiment. Let $z = x + y$. We want to calculate the pdf of z.

First, we will calculate the conditional pdf $f_{z|y}$ from $f_{x|y}$. Suppose y has a fixed value $y = c$, then $z = x + c$. We take this equation to mean that z is a function of x, i.e. $z = f(x) = x + c$. For this function, the Jacobian matrix is identity for all x, hence its determinant is 1 for all x. And $f^{-1}(b) = b - c$, so that

$$f_{z|y}(b \mid c) = f_{x|y}(b - c \mid c).$$

Therefore,

$$\begin{aligned} f_z(b) &= \int_{-\infty}^{\infty} f_{z|y}(b \mid c)f_y(c)dc \\ &= \int_{-\infty}^{\infty} f_{x|y}(b - c \mid c)f_y(c)dc \\ &= \int_{-\infty}^{\infty} f_{x,y}(b - c, c)dc. \end{aligned}$$

An interesting special case is when x and y are independent random variables. In that case, $f_{x,y}(b - c, c) = f_x(b - c)f_y(c)$, so that

$$f_z(b) = \int f_x(b - c)f_y(c)dc.$$

That is $f_z = f_x \otimes f_y$, where $\otimes$ is the convolution operator. This special case is important in many applications. **End of example.**

The technique used in the above example is often used to calculate the pdf of a random variable z which is a function of two random variables x, y.

Example: Suppose x, y are two joint 1-dim random variables. Let $r = \sqrt{x^2 + y^2}$. We want to calculate the pdf of r from the pdfs of x and y.

There are many ways of solving this problem. One interesting way is to introduce another random variable $\theta = \tan^{-1}(y/x)$ such that the function $F : \mathcal{R}^2 \to \mathcal{R}^2$ given by

$$F(x, y) = (\sqrt{x^2 + y^2}, \tan^{-1}(y/x)) = (r, \theta)$$

is one-to-one with a well defined Jacobian matrix. Then, we consider (r, θ) as a joint random variable, and use the Jacobian result to get

$$f_{r,\theta}(r, \theta) = \frac{f_{x,y}(F^{-1}(r, \theta))}{\mid \det J_F(F^{-1}(r, \theta)) \mid}.$$

Finally, we marginalize to get f_r:

$$f_r(r) = \int f_{r,\theta}(r, \theta) d\theta.$$

I encourage you to work out in detail the above formulae for the sample case where x and y are independent random variables, i.e. $f_{x,y} = f_x f_y$.

4.11 Summary of Recipes

So far we used several tricks in the examples to calculate the pdfs of functions of random variables. These tricks are actually general procedures which can often be used in solving real-life problems. I list these procedures below as general recipes:

Recipe 1 (cdf by definition):

This is the mother of all recipes and always works. Given $y = f(x)$, calculate the cdf of y as $F_y(a) = F_x(f^{-1}(H_a)) = \int_{f^{-1}(H_a)} f_x(u)du$. This may or is not always computationally tractable.

Recipe 2 (pdf by definition):

This is the pdf version of the above recipe. Given $y = f(x)$, calculate the pdf of y by first calculating the probability of the event A_a (remember, this is a small cube centered at a). Thus $P\{A_a\} = P\{f^{-1}(A_a)\} = \int_{f^{-1}(A_a)} f_x(u)du$. Then,

$$f_y(a) = \lim_{\text{vol } A_a \to 0} \frac{P\{A_a\}}{\text{vol } A_a}.$$

The feasibility of this method depends on whether the integral $\int_{f^{-1}(A_a)} f_x(u)du$ and the limit is tractable.

Recipe 3 (pdf by Jacobian):

If $f : \mathcal{R}^n \to \mathcal{R}^n$ with a well defined Jacobian matrix and if $f^{-1}(a)$ is a finite set of points. Then the pdf of $y = f(x)$ is given by the formula of equation (8). This method requires the domain and co-domain of f to have the same dimension.

Recipe 4 (pdf by conditionals):

If $z = f(x, y)$ such that for a fixed y the function f is from $\mathcal{R}^n$ to $\mathcal{R}^n$ and satisfies the requirements of recipe 3. Then, fix y and use recipe 3 to calculate $f_{z|y}$ from $f_{x|y}$. From this, obtain the joint distribution $f_{z,y} = f_{z|y} f_y$, and marginalize this to get $f_z = \int f_{z,y} dy$.

Recipe 5 (pdf by auxiliary variables):

If $z = f(x, y)$, then introduce a new random variable $w = g(x, y)$ such that function F defined by $F(x, y) = (f(x, y), g(x, y)) = (z, w)$ is one-to-one and has a Jacobian. Use recipe 3 to get the pdf $f_{z,w}$ and marginalize to get the pdf $f_z = \int f_{z,w} dw$.

These recipes can be used often, but are, of course, not exhaustive.

4.12 Repeated trials

Because the measurement in an experiment is random, we often conduct repeated trials to get many measurements. To analyze this, we associate a random variable with each trial. For n trials, we have n random variables, $x_1, \ldots, x_n$. To analyze all of these random variables, we consider them as joint random variables that are independent, but identically distributed. What this means is the following: *Joint random variables:* this really means that we consider the n repeated trials of the experiment as a *single trial* of a super-experiment that yields a single $n-$tuple of measurements. *Independent random variables:* Taking $x_1, \ldots, x_n$ as independent random variables implies that any one of the n experiments has no influence on any other. *Identically distributed:* This means that the marginal probability densities of $x_1, \ldots, x_n$ are all equal (since they are repeated trials of the *same* experiment). We denote the marginal probability distribution by f_x.

The term *i.i.d.* is often used to describe independent identically distributed random variables. When $x_1, \ldots, x_n$ are i.i.d.,

$$\begin{aligned} f_{x_1,\ldots,x_n}(a_1 \ldots, a_n) &= f_{x_1}(a_1) \times \ldots \times f_{x_n}(a_n) \text{ (Independence)} \\ &= f_x(a_1) \times \ldots \times f_x(a_n). \text{ (Identically distributed)} \end{aligned} \tag{23}$$

5 Convergence of Random Variables

Convergence of random variables is probably the most advanced topic we will need. Convergence of random variables is important in estimation theory because we will often have the following situations:

1. We conduct n repeated trials and calculate a single number from the n measurements as an estimate of something. For example, we may calculate the average of the n measurements to estimate some underlying constant. We want to know what happens to our estimate as we increase n, the number of measurements (trials). In particular we want to know whether this estimate converges, and if it does, then does it converge to the true value of the underlying constant?

   We think of the situation in the following way: First, we imagine a single super(mega?)-experiment consisting of infinite trials of the basic experiment. Each trial of this super-experiment yields an infinite sequence of measurements. To this super-experiment, we attach

an infinite sequence of random variables $x_1, \ldots, x_n, \ldots$ where x_n corresponds to calculating a single number from the first n measurements. We then ask whether the sequence of random numbers converges to the desired constant. What we mean by this convergence is explained below.

2. Once we know what it means for a sequence of random variables to converge to a constant, we can immediately generalize the idea to a sequence of random variables converging to another random variable. Suppose, as above, that $x_1, \ldots, x_n, \ldots$ is a sequence of random variables defined on a super-experiment. Further suppose that x is another random variable associated with the super-experiment. Then we say that the sequence of random variables $x_1, \ldots, x_n, \ldots$ converges to x if the sequence of random variables $x_1 - x, \ldots, x_n - x, \ldots$, converges to 0.

Although we used the idea of a super-experiment to motivate convergence of random variables, the definition (given below) only requires that there be some experiment in which we have a sequence of random variables $x_1, \ldots, x_n, \ldots$ plus another random variable x that the sequence may converge to.

There any many different (non-equivalent) notions of convergence of random variables. We will use two notions called *convergence in probability* and *convergence in distribution*.

Definition: The sequence of random variables $x_1, \ldots, x_n, \ldots$, converges in probability to a constant c if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P\{| \, x_n - c \, | > \epsilon\} = 0. \tag{24}$$

Convergence in probability is written as $x_n \xrightarrow{P} c$.

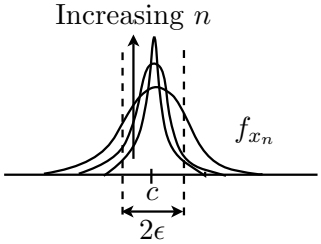


Figure 10: Convergence in Probability.

The definition is illustrated in Figure 10 (page 20) for $\mathcal{R}^1$ valued random variables. We pick any positive number ϵ which defines the half width of an interval around c. Then, we calculate the probability $P\{| \, x_n - c \, | > \epsilon\} = 1 - \int_{\epsilon}^{\epsilon} f_{x_n - c}(u) du$ and check whether it goes to zero as n goes to ∞. If it does go to zero for *any* ϵ, then we say that $x_n \xrightarrow{P} c$.

Definition: The sequence of random variables $x_1, \ldots, x_n, \ldots$, converges in probability to a random variable x, if $x_n - x \xrightarrow{P} 0$.

Convergence in probability is a very useful notion in estimation theory. *Convergence in distribution* is a slightly weaker notion, but is also useful:

Definition: The sequence of random variables $x_1, \ldots, x_n, \ldots$, *converges in distribution* to x a random variable x if the *cdf* F_{x_n} converges to the cdf F_x (this convergence is the ordinary convergence of functions). Convergence in distribution is denoted as $F_{x_n} \to F_x$.

Some comments on the difference between convergence in probability and convergence in distribution:

1. Convergence in distribution is really a convergence of the cdf as a deterministic function. It does not address the underlying "randomness" in the experiment. Convergence in probability does.

2. If a sequence of random variables converge in probability to another random variable, then they also converge in distribution. But the converse is not true. This is what we mean by "convergence in distribution is a weaker notion than convergence in probability".

3. The definition of convergence in distribution does not require the random variable sequence $x_1, \ldots, x_n, \ldots$ and the random variable x to be defined in the same experiment. All that is required is that their cdfs exist. In contrast, convergence in probability requires that $x_1, \ldots, x_n, \ldots$ and x be defined on the same experiment.

4. You will not be required to be prove that a sequence of random variables converges to another random variable. However, you will encounter theorems which say that a certain sequence of random variables converges to another random variable in probability or in distribution and you <u>are</u> required to understand what this means.

There are two important theorems in probability that use the notion of convergence in probability and convergence in distribution. The theorems are called the (weak) law of large numbers and the central limit theorem

Theorem: (*The law of large numbers*) If $x_1, \ldots, x_n, \ldots$ is a sequence of i.i.d. random variables with mean μ, then the sequence of random variables defined by $\hat{x}_n = \frac{1}{n}(x_1 + \ldots + x_n)$ converges in probability to μ, i.e. $\hat{x}_n \xrightarrow{P} \mu$.

Theorem: (*The central limit theorem*) Let $x_1, \ldots, x_n, \ldots$ be a sequence of i.i.d. random variables with mean μ and a finite variance-covariance matrix Σ. If $\hat{x}_n = \frac{1}{n}(x_1 + \ldots + x_n)$, then the sequence $y_n = \sqrt{n}(\hat{x}_n - \mu)$ converges in distribution to $\mathcal{N}(0, \Sigma)$ (the notation $\mathcal{N}(0, \Sigma)$ will be given meaning below).

6 Expected Values and Moments

Definition: If x is an n-dim random variable, and $g : \mathcal{R}^n \to \mathcal{R}^m$ is a function, then the *expected value* of g, denoted $E(g)$ or $E[g]$, is

$$E(g) = \int_{-\infty}^{\infty} g(u) f_x(u) du. \tag{25}$$

The notation for expected values is a little peculiar. When we have a formula for $g(x)$, then we write $E(\text{formula for } g(x))$ instead of $E(g)$, e.g. if $g(x) = x^T x$, then we write $E(x^T x)$ for $E(g)$. Also it is common to use x as the variable of integration so that $E(g) = \int g(x) f_x(x) dx$. When written

this way, the subscript in the probability density f_x should not be viewed as the same variable as the one used in the integration.

You are probably familiar with $E(x) = \int x f_x(x)dx$ as the *mean* of the random variable x. The mean of the random variable x is also written as $\bar{x}$.

The notation for E suggests that we think of E as a function - it accepts a function g as an input and produces a vector $E(g)$ as an output. When viewed this way, we call E the *expectation operator*. A very important property of the expectation operator is that it is linear:

Theorem: If $g, h : \mathcal{R}^n \to \mathcal{R}^m$, then for all real numbers α, β,

$$E(\alpha g + \beta h) = \alpha E(g) + \beta E(h).$$

Proof:

$$\begin{aligned} E(\alpha g + \beta h) &= \int (\alpha g(u) + \beta h(u)) f_x(u)du \\ &= \alpha \int g(u) f_x(u)du + \beta \int h(u) f_x(u)du \\ &= \alpha E(g) + \beta E(h). \end{aligned}$$

Later, we will use the following properties of expected values:

1. If c is a constant, then $E[c] = c$. The proof is simple: $E[c] = \int c f_x(u)du = c \int f_x(u)du = c$.
2. If x is an n-dim random variable with $\mu = E[x]$ and $c \in \mathcal{R}^n$ is a constant, then $E[x^T c] = \mu^T c$. For a proof, begin by noting that
$$\mu = \begin{bmatrix} \mu_1 \\ \cdots \\ \mu_n \end{bmatrix} = \int \begin{bmatrix} x_1 \\ \cdots \\ x_n \end{bmatrix} f_x(u_1, \ldots, u_n)du_1 \ldots du_n.$$
so that $\mu_k = \int x_k f_x(u_1, \ldots, u_n)du_1 \ldots du_n = E[x_k]$ for $k = 1, \ldots, n$.
Thus, $E[x^T c] = E[\sum_{k=1}^n x_k c_k] = \sum_{k=1}^n E[x_k c_k] = \sum_{k=1}^n \mu_k c_k = \mu^T c$.
3. If x is an n-dim random variable with $\mu = E[x]$ and $c \in \mathcal{R}^n$ is a constant, then
$$E[\|x - c\|^2] = \|\mu - c\|^2 + E[(x - \mu)^T (x - \mu)].$$
The proof is as follows:
$$\begin{aligned} E[\|x - c\|^2] &= E[(x - c)^T (x - c)] \\ &= E[(x - \mu + \mu - c)^T (x - \mu + \mu - c)] \\ &= E[(x - \mu)^T (x - \mu) + 2(x - \mu)^T (\mu - c) + (\mu - c)^T (\mu - c)] \\ &= E[(x - \mu)^T (x - \mu)] + 2E[(x - \mu)^T (\mu - c)] + E[(\mu - c)^T (\mu - c)]. \end{aligned}$$
Consider the second the third term:
$$\begin{aligned} E[(\mu - c)^T (\mu - c)] &= (\mu - c)^T (\mu - c) \text{ (since } (\mu - c)^T (\mu - c) \text{ is a constant.)} \\ E[(x - \mu)^T (\mu - c)] &= E[x^T (\mu - c) - \mu^T (\mu - c)] \\ &= E[x^T (\mu - c)] - \mu^T (\mu - c)] \\ &= E[x^T (\mu - c)] - E[\mu^T (\mu - c)] \\ &= \mu^T (\mu - c) - \mu^T (\mu - c) = 0. \end{aligned}$$

6.1 Moments

Definition: If x is a 1-dim random variable, then $E(x^k)$ is called the k^{th} *moment* of x and $E((x-\bar{x})^k)$ is called the k^{th} *central moment* of x.

The first central moment is always zero. The second central moment is called the *variance*. The square root of the second central moment is the *standard deviation*.

Similar moments and central moments can be defined for n-dim random variables. The following are especially important:

$$\hat{x} = E(x)(\text{Mean}) \tag{26}$$

$$C_{xx} = E((x-\bar{x})(x-\bar{x})^T)(\text{Variance-covariance matrix}) \tag{27}$$

If x is n-dim, then C_{xx} is an $n \times n$ matrix whose i, j^{th} entry is C_{xx} is $E((x_i - \bar{x}_i)(x_j - \bar{x}_j))$. Thus, C_{xx} contains variances of the components of x along the diagonal and the covariances of the components in other locations. The matrix C_{xx} is symmetric.

Given two n-dim random variables x and y the *cross-covariance matrix* is $C_{x,y} = E((x-\hat{x})(y-\hat{y})^T)$. If x and y are independent random variables, then the cross-covariance matrix is 0.

7 Discrete Random Variables

So far all our examples and illustrations have implicitly assumed that the pdf of random variables is a smooth function of its arguments. There is an important class of random variables called *discrete variables* whose pdf is a weighted sum of delta functions centered at $a_1, \ldots$:

$$f_x(a) = \sum_i \alpha_i \delta(a - a_i), \tag{28}$$

where, all $\alpha_i \geq 0$, and $\sum_i \alpha_i = 1$, and sum may have finite or infinite terms.

Usually we do not talk about the pdf of a discrete random variable. Instead we talk of the probability distribution p_x of the random variable:

$$p_x(a) = \begin{cases} \alpha_i & \text{if } a = a_i \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

Dealing with discrete random variables sometimes requires care. We will deal with this as the situation arises.

8 Some commonly-used probability density functions

We will need a number of models of probability density functions for solving estimation theoretic questions. There is an entire discussion of this to come. For now, we will simply list a few distributions so that we can construct some simple illustrative examples:

8.1 Scalar (univariate) densities

The univariate densities discussed below are illustrated in figure 11 on page 25.

Normal: This pdf is given by:

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2}), \tag{30}$$

where μ and σ^2 are the mean and the variance of the density. The normal density is also called the *Gaussian density*. The normal density is also denoted as $\mathcal{N}(\mu, \sigma)$.

The normal density has a number of interesting properties which are stated below in the discussion of the multivariate normal density.

Uniform: The *uniform density* on the interval $[a, b]$, is

$$f_x(x) = \begin{cases} \frac{1}{(b-a)} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

This density is denoted $\mathcal{U}(a, b)$. Note that there is no uniform density over an infinite or a semi-infinite interval.

Gamma: The *Gamma density* is defined for non-negative values of its argument:

$$f_x(x) = x^{k-1} \frac{e^{-x/\theta}}{\theta^k \Gamma(k)},$$

where $k > 0$ is called *shape parameter* and θ is called the *scale parameter*, and Γ is the Gamma function of calculus. The mean and variance of a Gamma distribution are $k\theta$ and $k\theta^2$ respectively.

Notice the profound change in shape of this density as k increases from 0 to 1 and beyond.

Poisson: This is a discrete distribution. Its probability function is given by:

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{32}$$

for x integer ≥ 0, where $E(x) = \lambda$. It turns out that $E((x-\lambda)^2) = \lambda$. That is the mean and variance of a Poisson distribution are equal. The number λ is called the *intensity* of the Poisson distribution. The Poisson distribution is denoted $Poisson(\lambda)$.

The Poisson distribution is used to model the number of photons arriving at a detector in a finite time interval.

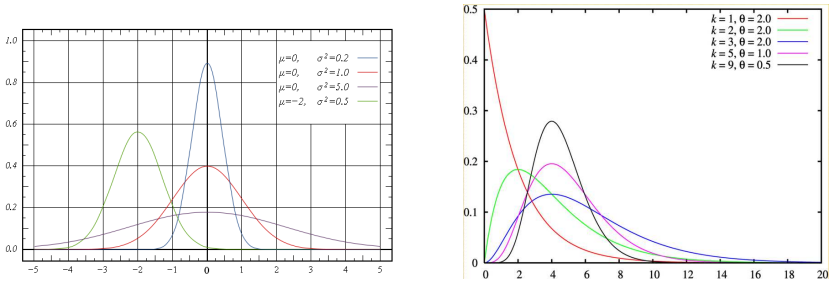
The Poisson distribution has the following property: If independent random variables $x_1, x_2, \ldots, x_n$ are distributed according to $Poisson(\lambda_1), Poisson(\lambda_2), \ldots, Poisson(\lambda_n)$, then the random variable $x_1 + x_2 + \ldots + x_n$ is distributed according to $Poisson(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$.

When the intensity λ is large, the Poisson distribution can be approximated by a normal distribution with mean λ and variance equal to $\sqrt{\lambda}$.

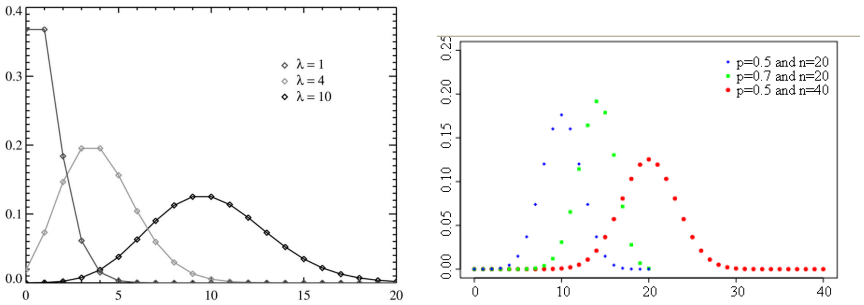
Binomial: This is also a discrete distribution. If the probability of one event in a trial is p, then the probability of observing k events in n trials is given by the Binomial distribution:

$$f(k) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k}. \tag{33}$$

The mean and variance of the Binomial distribution are np and $np(1-p)$ respectively. We will denote the Binomial distribution as $\mathcal{B}(n, p)$.



(a) Univariate normal density (b) Gamma density



(c) Poisson distribution (d) Binomial distribution

Figure 11: Some univariate densities and distributions (from Wikimedia.org)

8.2 Multivariate densities

Multivariate Normal (or multivariate Gaussian): If x is an n-dim random variable then it has a multivariate normal density $\mathcal{N}(\mu, \Sigma)$ if

$$f(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{34}$$

and μ is the mean and $\Sigma = C_{xx}$ is the variance-covariance matrix of the random variable. The matrix Σ is required to be positive definite and symmetric.

Now that you understand the notation $\mathcal{N}(\mu, \Sigma)$, read the central limit theorem again.

The multivariate normal has the following properties:

1. Because the matrix Σ is symmetric and positive definite, there exists an $n \times n$ orthonormal matrix U such that $\Sigma = U^T\Lambda U$, where Λ is a diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$. The transposed rows of the U matrix form an orthonormal basis of $\mathcal{R}^n$, (i.e. $U^TU = UU^T = I$). The λ_k and the (transpose of) the k^{th} row of U form an eigensystem (an eigenvalue + eigenvector) of Σ.

2. The contours $f(x) = c$ are ellipsoids in $\mathcal{R}^n$ with center at μ, the transposed rows of U as the directions of the axes whose lengths are proportional to $\sqrt{\lambda_k}$.

3. Suppose we let $\Lambda^{1/2}$ be a diagonal matrix with the k^{th} diagonal entry equal to $\sqrt{\lambda_k}$, then the matrix $A = U^T\Lambda^{1/2}U$ has the property that $AA = \Sigma$ (check this), and hence A is usually denoted $\Sigma^{1/2}$. If a random variable Z is distributed as $\mathcal{N}(0, I)$ and $x = \mu + \Sigma^{1/2}z$, then x is

distributed as $\mathcal{N}(\mu, \Sigma)$. This is often a useful way of thinking about a multivariate normal distribution.

4. If C is an $m \times n$ matrix with $m \leq n$, then the random variable Cx is distributed as $\mathcal{N}(C\mu, C\Sigma C^T)$. We refer to this by saying that "linear combinations preserve normality".

5. Suppose we partition x as $x = (x_a, x_b)$, and correspondingly μ as $\mu = (\mu_a, \mu_b)$ and Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \tag{35}$$

   then, the marginal density of x_a is $\mathcal{N}(\mu_a, \Sigma_{aa})$, and the conditional density of x_b given x_a is

$$\mathcal{N}(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}). \tag{36}$$

6. In the above, if x_a and x_b has the same dimension, then $x_a + x_b$ has the normal density $\mathcal{N}(\mu_a + \mu_b, \Sigma_{aa} + \Sigma_{ab} + \Sigma_{ba} + \Sigma_{bb})$.

Multinomial: This is a discrete probability distribution. The classic "urn" example describes it easily. Suppose that an urn has balls with k different colors and that the probability of choosing a ball with the i^{th} color is p_i. We draw n balls and use $x = (x_1, \ldots, x_k)$ to denote the number of balls of different colors that we have drawn (x_i is the number of balls of the i^{th} color). Then x is a random variable with a multinomial distribution:

$$f(x) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}. \tag{37}$$

The multinomial distribution has the property that the marginal distribution of x_i is Binomial with parameters n and p_i.

9 Parameterized Densities

All of the densities we considered above have parameters (e.g. the mean and the standard deviation are parameters of the univariate normal). Very often we will explicitly have to consider how the probability density varies as a function of the parameters. When we need to do this, we will write the probability density function as $f_x(u; \theta)$, where u is the value at which the density is evaluated, and θ is the parameter. If a probability density function contains more than one parameter, then we often group all the parameters into a single vector of parameters θ.

If a probability density function does not have parameters, then we can create a parameterized probability density from it by using a parameterized function. For example, suppose x has the probability density f_x, then setting $y = \theta x$, where $\theta > 0$ is a scalar parameter gives the probability density function $p_y(y; \theta) = \frac{1}{\theta} f_x(y/\theta)$.

Probability density functions which are related to each other by changes of parameter are said to belong to a *parametric family of densities*.

There is an alternate notation for parametric densities. When the random variable is understood from context, it is common to suppress the subscript and write the density simply as $f(x; \theta)$.

9.1 Parameters as random variables

Given a value of the parameter θ, the probability density function $f(x;\theta)$ describes the density of the random variable x. If we regard θ in turn as a random variable, then $f(x;\theta)$ can be thought of as the conditional probability density of x given θ. If θ has the marginal distribution $f_\theta(\theta)$, then the joint distribution of x and θ is $f_{x,\theta}(x,\theta) = f(x;\theta)f_\theta(\theta)$. From the marginal, we can calculate the marginal of x as $f_x(x) = \int f_{x,\theta}(x,\theta)d\theta = \int f(x;\theta)f_\theta(\theta)d\theta$.

10 Acknowledgements

Thanks to Nicha Dvornek for reading this manuscript closely and find typos.