Random Variables

EECS 126 at UC Berkeley

Spring 2022

1 Definition*

* This section explains the notation of $X(\omega)$, but it is supplemental and mostly optional.

A random variable X should represent "something that takes on certain values x, each with some associated probability $p_X(x)$." We say that $X = x$ is a particular *realization* of X, and we call the set of possible values X can take on its *range* or *support* $\mathcal{X}$. We may be familiar with discrete random variables, those with finite or countably infinite support, which satisfy

$$p_X(x) \geq 0; \quad \sum_{x \in \mathcal{X}} p_X(x) = 1; \quad p_X(x) = \mathbb{P}(X = x).$$

We may further be familiar with well-behaved continuous random variables and ways to describe them. But here we will present a more general definition of random variables that can take on any type of values — vectors, matrices, graphs, sets, etc. As with probability spaces, this definition of random variables is more formal, but it should not go against the basic intuition.

Definition for random variable

Definition 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{X}$ be a set of values, and let Σ be a σ-algebra on $\mathcal{X}$. Then a **random variable** is a mapping $X : \Omega \to \mathcal{X}$, such that for every $B \in \Sigma$ (where $B \subseteq \mathcal{X}$), its preimage $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ is an event in $\mathcal{F}$. We write

$$\mathbb{P}(X \in B) := \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

$\omega \in \Omega : X(\omega) \in B$

Let's unwrap this definition. Σ is a collection of subsets of $\mathcal{X}$, so an example of a $B \in \Sigma$ might be $B = \{x\}$. Then $X \in B$ means $X = x$, and $X^{-1}(\{x\}) = \{\omega \in \Omega : X(\omega) = x\}$ is the set of outcomes in which X takes on the value x. Finally, the probability of the event $\{X = x\}$ is supplied by $\mathbb{P}$, the probability measure on Ω.

$\Omega \Rightarrow \mathcal{F} \rightarrow \{x\}$

Why might we want to go through the trouble of this seemingly more convoluted definition?

1. We can keep the idea of "randomness" separate from "what types of values X takes on." There should be something in common between the same uniform distribution on n random integers and on n random graphs — we can describe what is common by the shared probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

2. Perhaps more convincingly, we are actually already familiar with this definition! Given X, we can consider some random variable $f(X)$, e.g. X^2. How do we find the probabilities associated with $f(X)$?

$$\mathbb{P}(f(X) = y) = \mathbb{P}(X \in f^{-1}(y)) = \mathbb{P}(X \in \{x : f(x) = y\}).$$

For example, $\mathbb{P}(X^2 = 9) = \mathbb{P}(X \in \{-3, 3\})$. So this definition makes the composition of functions with random variables very, very natural.

3. Lastly, we are also familiar with notation of the form $X \in B$ for $B \in \Sigma$. For example, $a \leq X \leq b$ means the same as $X \in [a, b]$, or the complement $(X \leq b) \setminus (X < a)$. Describing $\{X \in B\}$ as events in a σ-algebra means we are able to work with the probabilities of their unions, intersections, and complements.

We won't work with this level of formality in this course, and you're not expected to know any parts of this definition. The aim was to provide clarification for when we come across the notation $X(\omega)$. The key takeaway is that we can evaluate X at different ωs to get different values or realizations $X(\omega) = x$. The view that X is a function on a probability space — **this** is how we mathematically describe the randomness of random variables. The probability space supplies the randomness.

A final note: we usually don't specify an explicit probability space Ω when working with random variables. Instead, we think of X giving a probability *distribution* on $\mathcal{X}$ itself, seemingly independent of any underlying Ω. We can formalize this natural idea with the following definition:

Definition 2. The **distribution** or law of X is the function $\mu = \mathbb{P} \circ X^{-1}$, which is a probability measure on $\mathcal{X}$. We think of $\mu(B)$ as $\mathbb{P}(X \in B)$, for example $\mu(\{x\}) = \mathbb{P}(X = x)$. μ is also called a *pushforward* measure: we "push" the probabilities in Ω forward, along the mapping $X : \Omega \to \mathcal{X}$, onto the space $\mathcal{X}$.

We sometimes use the terms random variable and distribution almost as if they are interchangeable, but keep in mind their difference: a random variable $X : \Omega \to \mathcal{X}$ is a function that maps onto values, while its distribution is a probability measure $\mu : \Sigma \to [0, 1]$ that assigns probabilities $\mu(B) = \mathbb{P}(X \in B)$ to subsets $B \subseteq \mathcal{X}$, $B \in \Sigma$.

In the next section, we describe the random variables we will work with most often: $\mathbb{N}$- or $\mathbb{Z}$-valued discrete random variables, and well-behaved $\mathbb{R}$-valued continuous random variables.

2 Characterizing random variables

We can describe the distributions of discrete random variables, uniquely, without much issue.

Definition 3. The **probability mass function** (pmf) of a discrete random variable X is the unique function $p_X : \mathcal{X} \to [0, 1]$, given by $p_X(x) = \mathbb{P}(X = x)$, that satisfies

1. Nonnegativity: $p_X(x) \geq 0$, and
2. Normalization: $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Conversely, any function $p : \mathcal{X} \to [0, 1]$ that satisfies the above properties corresponds to a unique discrete random variable $X : \mathbb{P}(X = x) = p(x)$. Thus $X \mapsto p_X$ is a bijective correspondence.

Furthermore, if $A \subseteq \mathcal{X}$ is a subset of values, then we can usually say $\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$. Less commonly, we also have the unique **cumulative mass function** (cmf),

$$\mathbb{P}(X \leq x) = \sum_{n=-\infty}^{x} p_X(n).$$

Some common discrete distributions you will want to become familiar with are the degenerate distribution or point mass $X = x_0$, $\mathsf{Bernoulli}(p)$, Rademacher, $\mathsf{Uniform}([n])$, $\mathsf{Binomial}(n, p)$, $\mathsf{Geom}(p)$, and $\mathsf{Poisson}(\lambda)$. Many other resources cover these distributions quite well, so we will elect to only give an example here.

Example. The pmf and cmf of a geometric random variable $X \sim \mathsf{Geom}(p)$ are

$$\mathbb{P}(X = k) = p(1-p)^{k-1} \quad \text{and} \quad \mathbb{P}(X \leq k) = 1 - (1-p)^k.$$

We choose the convention that the support of X is the positive integers $\mathbb{Z}^+$, not the nonnegative integers $\mathbb{N}$.

On the other hand, continuous $\mathbb{R}$-valued random variables can be complicated in general, so we assume a condition of well-behavedness to simplify our lives: $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$. [The condition is formally called *absolute continuity* (with respect to the Lebesgue measure).] We can then make statements like $\mathbb{P}(X > x) = \mathbb{P}(X \geq x)$, or use integration, which also does not "distinguish" $[x, y]$ and $(x, y]$.

This condition does present a small problem: the probability of the event $\{X = x\}$ becomes meaningless. If a random variable is not uniform, then we expect that certain values are more likely than others, yet $\mathbb{P}(X = x) = \mathbb{P}(X = x') = 0$ for any values $x, x' \in \mathbb{R}$. If we cannot find the probability of X at a single point, what about the probability that X lies in a continuous interval?

Definition 4. The **cumulative distribution function** (cdf) of a continuous random variable is

$$F_X(x) := \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x]).$$

Theorem. *Every cumulative distribution function $F_X : \mathbb{R} \to [0, 1]$ satisfies the following properties.*

1. *Nondecreasing: for any $x, y \in \mathbb{R}$, $x \leq y \implies F_X(x) \leq F_X(y)$,*
2. *Right-continuity: for every $x \in \mathbb{R}$, $\lim_{y \downarrow x} F_X(y) = F_X(x)$, and*
3. *Normalization: $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.*

Conversely, every function $F : \mathbb{R} \to [0, 1]$ that satisfies the above properties is the cdf of some random variable. Furthermore, there is a bijective correspondence between continuous distributions and cumulative distribution functions, given by $\mu \mapsto F : F(x) = \mu((-\infty, x])$.

The above result of uniqueness justifies the definition of the cdf. We also have the identity

$$\mathbb{P}(X \in [a, b]) = F_X(b) - F_X(a).$$

However, we still lack an analogue of $\mathbb{P}(X = x)$ for continuous distributions. Very informally, we may consider $d\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x + dx) - \mathbb{P}(X \leq x) = \mathbb{P}(X \in [x, x + dx])$ as the analogue of the quantity "$\mathbb{P}(X = x)\, dx$."

Formally, we will outsource an fact from analysis: absolute continuity implies almost-sure differentiability. Thus the condition of well-behavedness we assumed in the beginning turns out to be precisely equivalent to the existence of the probability density function.

Definition 5. The **probability density function** (pdf) of a continuous random variable X is

$$f_X(x) = \frac{d}{dx} F_X(x).$$

Theorem. *Every probability density function $f_X : \mathbb{R} \to [0, \infty)$ satisfies the following properties. Note that the density does not have to be a probability value in $[0, 1]$.*

1. *Nonnegativity: $f_X(x) \geq 0$. (This follows from the nondecreasingness of the cdf.)*
2. *Normalization: $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$. (This follows from the normalization of the cdf.)*

Conversely, every function $f : \mathbb{R} \to [0, \infty)$ that satisfies the above properties is the pdf of some random variable. The pdf is unique up to almost-sure equivalence: $f_X(x) = f_Y(x)$ if and only if $X \overset{\text{a.s.}}{=} Y$, i.e. $\mathbb{P}(X = Y) = 1$.

(We might notice that another function satisfies nonnegativity and normalization: the pmf for discrete distributions, where the integral over $\mathbb{R}$ is the analogue to the sum over $\mathbb{Z}$.) We can now apply the probability density function, along with the fundamental theorem of calculus, as follows.

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt \quad \text{and} \quad \mathbb{P}(X \in [a, b]) = \int_a^b f_X(x)\, dx = F_X(b) - F_X(a).$$

Definition 6. We find it convenient to define the **complementary cumulative distribution function** (ccdf) or *survivor function* of a random variable X,

$$\bar{F}_X(x) := \mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x).$$

Definition 7. Finally, we may define the **inverse distribution function** or percentile-point function (ppf) of X, used commonly in statistical applications and *inverse transform sampling*. If F_X is strictly increasing, then it is invertible, and $F_X^{-1} : [0, 1] \to \mathcal{X}$ is the inverse distribution function. In general,

$$F_X^{-1}(p) := \min \{x \in \mathbb{R} : F_X(x) \geq p\}$$

has the property that $F_X^{-1}(F_X(x)) \overset{\text{a.s.}}{=} x$. The *median* value or 50th percentile of X is given by $F_X^{-1}(0.5)$.

Some common continuous random variables you will want to be familiar with are $\mathsf{Uniform}([a, b])$, $\mathsf{Exponential}(\lambda)$, the Gaussian or normal $\mathcal{N}(\mu, \sigma^2)$, and the jointly Gaussian or multivariate normal $\mathcal{N}(\vec{\mu}, \Sigma)$.

Example. If $Z \sim \mathcal{N}(0, 1)$ is distributed as the *standard normal* distribution, then its pdf is

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right).$$

Its cdf does not have a closed-form expression (see the *error function*). In this course, we will use the convention that the second parameter in $\mathcal{N}(\mu, \sigma^2)$ is always variance, not standard deviation.

3 Multiple random variables

Definition 8. The **joint** cumulative distribution function of random variables X and Y is the function

$$F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y).$$

If X and Y are continuous, their joint probability density function is

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y);$$

If X and Y are discrete, their joint probability mass function is simpler:

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y).$$

For convenience, we may denote both discrete pmfs and continuous pdfs using the letter f. To find the discrete or continuous version, you can interchange $\sum_{n=-\infty}^{\infty} p_*(\cdot)$ with $\int_{-\infty}^{\infty} f_*(\cdot)\, dx$. The cmf and cdf are both defined as $F_X(x) = \mathbb{P}(X \le x)$, so the discrete–continuous distinction is less important for the cumulative functions.

Definition 9. The **marginal** probability density (or mass) function of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy.$$

Definition 10. The **conditional** probability density (or mass) function of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

Notice that for discrete probability mass functions, this is simply the definition of conditional probability. So for continuous distributions, the definition gives us the continuous analogue of Bayes' rule.

Graphically, finding the marginal density of X resembles a projection $\mathcal{X} \times \mathcal{Y}\mathcal{X}$. The new density at the *point* $X = x$ is the "accumulation" of the joint density along the vertical *line* $X = x$. Finding the conditional density resembles a slice: restricting ourselves to the line $X = x$ and dividing by $f_X(x)$ to re-normalize.

On that note, as probability density (or mass) functions, the joint, marginal, and conditional pdfs satisfy the properties of 1) nonnegativity and 2) normalization:

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx = \int_{-\infty}^{\infty} f_X(x)\, dx = \int_{-\infty}^{\infty} f_Y(y)\, dy = \int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\, dy.$$

We will generally assume that *Fubini's theorem* applies, so we can always exchange the order of integration in multiple integrals. If we want to find the probability of the event $\{X \in A, Y \in B\}$, for example the two-dimensional region $(X,Y) \in [a,b] \times [c,d]$, we have

$$\mathbb{P}(X \in A, Y \in B) = \int_a^b \int_c^d f_{X,Y}(x,y)\, dy\, dx = \int_c^d \int_a^b f_{X,Y}(x,y)\, dx\, dy.$$

Definition 11. Two random variables X and Y are **independent** if for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, the events $\{X \le x\}$ and $\{Y \le y\}$ are independent events:

$$\mathbb{P}(X \le x \cap Y \le y) = \mathbb{P}(X \le x) \cdot \mathbb{P}(Y \le y).$$

Equivalently, X and Y are independent iff their joint distribution splits as a product of marginal distributions, or iff the conditional distributions are equal to the marginal distributions. That is, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y); \quad f_{Y|X}(y \mid x) = f_Y(y); \quad f_{X|Y}(x \mid y) = f_X(x).$$

A finite collection of random variables $\{X_i\}_{i=1}^n$ are (*mutually*) independent if for every combination of $x_i \in \mathcal{X}_i$, the events $\{X_i \leq x_i\}$ are mutually independent:

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

So from the formal definition of random variables, we get the definition of independence of random variables for free, from the definition of independence of events, by treating $\{X \leq x\}$ as events in a σ-algebra.

Definition 12. Two random variables X, Y are often (not always!) dependent if one is a *function* of the other, $Y = g(X)$, by which we mean

$$\mathbb{P}(g(X) = y) = \mathbb{P}(X \in g^{-1}(y)) = \mathbb{P}(\{\omega \in \Omega : g(X(\omega)) = y\}).$$

If X and Y are discrete, we can find the pmf of Y as

$$\mathbb{P}(g(X) = y) = \sum_{\{x \in \mathcal{X} : g(x) = y\}} p_X(x).$$

If g is invertible and $x = g^{-1}(y)$, then by the chain rule for inverse functions, we can find the pdf of Y as

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X \leq x) \implies f_Y(y) = \frac{d}{dy}\mathbb{P}(X \leq x) = \frac{d}{dy}F_X(g^{-1}(y)) = \frac{1}{|g'(x)|}f_X(x).$$

We conclude this section by discussing independent sums and ordered comparisons of multiple random variables.

Definition 13. If X and Y are independent random variables, then the distribution of their sum $X + Y$ is the **convolution** of the individual distributions:

$$\mathbb{P}(X + Y = n) = \sum_{k=-\infty}^{\infty} \mathbb{P}(X = k) \cdot \mathbb{P}(Y = n - k)$$

if X and Y are discrete; if they are continuous, then

$$f_{X+Y}(z) = (f_X * f_Y)(z) := \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z - x)\, dx.$$

Definition 14. Let $X_1, \ldots, X_n$ be i.i.d. random variables with common cdf $F_X(x)$. Then the **kth order statistic** $X_{(k)}$ is the random variable defined pointwise as the kth smallest value. Its cdf is

$$F_{X_{(k)}}(x) = \sum_{m=k}^{n} \binom{n}{m} \left[F_X(x)^m \cdot (1 - F_X(x))^{n-m}\right] = \sum_{m=k}^{n} \binom{n}{m} \left[\mathbb{P}(X \leq x)^m \cdot \mathbb{P}(X > x)^{n-m}\right].$$

A brief derivation: if $X_{(k)} \leq x$, then this is the same as for each $k \leq m \leq n$, choosing m variables to be $\leq x$ and the remaining $n - m$ variables to be $> x$. In particular, for $X_{\min} = X_{(1)}$ and $X_{\max} = X_{(n)}$,

$$F_{X_{\min}}(x) = 1 - (1 - F_X(x))^n \text{ and } F_{X_{\max}}(x) = F_X(x)^n.$$

Further topics to explore beyond this note include the distribution of products and powers of random variables; the algebra of random variables; families of distributions invariant under some transformation, such as Gaussian distributions under affine transformations; and many more.

In the next few notes, we will consider *expectation* and related functions, which further characterize certain features of the distributions of random variables; *concentration inequalities*, probabilistic bounds on the deviations of a random variable; and the *modes of convergence* of infinite sequences of random variables.

■

Random Variables Summary

① Definitions:

a) random variables is a mapping $X: \Omega \to \mathcal{X}$

b) probability mass function (pmf): (discrete random variable X)

and $P_X(x) = P(X=x)$

i) Nonnegativity: $P_X(x) \geq 0$

ii) Normalization: $\sum_{x \in \mathcal{X}} P_X(x) = 1$

iii) cumulative mass function (cmf): $P(X \leq x) = \sum_{n=-\infty}^{x} P_X(x)$

iv) cumulative distribution function (cdf): (continuous random variable)

$F_X(x) := P(X \leq x) = P(X \in (-\infty, x])$

while a single point x is meaningless $\Leftrightarrow$ $P((-\infty, x)) = P((-\infty, x])$

v) probability density function (pdf): $f_X(x) = \frac{d}{dx} F_X(x)$

vi) complementary cumulative distribution function (ccdf)

$\bar{F}_X(x) := P(X > x) = 1 - P(X \leq x)$

vii) inverse distribution function / percentile-point function (ppf)

$F_X^{\neg}: [0,1] \to \mathcal{X} \Rightarrow F_X^{\neg}(p) := \min\{x \in \mathbb{R}: F_X(x) \geq p\}$

② Multiple random variables

1) $F_{X,Y}(x,y) = P(X \le x, Y \le y)$ (joint cmf)

2) $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$ (joint pdf)

3) $P_{X,Y}(x,y) = P(X=x, Y=y)$ (joint pmf)

4) $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ (marginal pdf)

5) $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ (conditional pdf)