# PROBABILITY

NICHOLAS HAYEK

*Based on lectures by Prof. Louigi Addario-Berry*

CONTENTS

# I Introduction

## PROBABILITY SPACES

A probability space, denoted, $(\Omega, \mathcal{F}, \mathbb{P})$, describes the environment in which all of our case-studies will take place. It is equipped with a sample space, $\Omega$, which contains all possible outcomes of our case-study, an event space, $\mathcal{F}$, to describe what *subsets* of $\Omega$ ("events") are relevant to our study, and finally the probability measure, $\mathbb{P}$, which assigns a probability to each event $A \subseteq \mathcal{F}$.

### *Kolmogorov's Axioms*

**1.1 Kolmogorov's Axioms**

1. $0 \leq P(A) \leq 1$
2. $P(\emptyset) = 0$ and $P(\Omega) = 1$
3. If $(A_n, n \geq 1)$ are disjoint events, then $\mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$

A bit of notation:
$[n] = \{1, 2, ...n\}$
$AB = A \cap B$
$(n)_k = n\dot{(}n-1)\cdot...\cdot(n-k+1)$

These axioms imply that:

$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$

where $A^c \cup A$ is by definition equal to the sample space $\Omega$.

Consider a sufficiently random number generator which spits out values from 1 to 50, from which we pick 3 numbers in succession. The following are common models and their resulting probabilities.

| Experiment Type | Probability Model | $\mathbb{P}$ |
|---|---|---|
| Sampling with replacement | $\Omega = [50]^3 = [50] \times [50] \times [50]$ | $\mathbb{P}(a,b,c) = \frac{1}{50^3}$ |
| Sampling without replacement | $\Omega = \{(a,b,c) : a \neq b \neq c \neq a\}$ | $\mathbb{P}(a,b,c) = \frac{1}{50 \cdot 49 \cdot 48}$ |
| Unordered with replacement | $\Omega = \{\text{permutations of } (a,b,c)\}$ | $\mathbb{P}(a,b,c) = \frac{1}{50^3 \cdot 3!}$ |
| Unordered without replacement | $\Omega = \{\text{permutations of } (a,b,c) : a \neq b \neq c \neq a\}$ | $\frac{1}{50 \cdot 49 \cdot 48 \cdot 3!} = \binom{50}{3}$ |

**1.2 Monotonicity of Probability**
If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

PROOF.

$A \subseteq B$, $B = A \cup A^c B \implies \mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c B) \geq \mathbb{P}(A)$, so $\mathbb{P}(A) \leq \mathbb{P}(B)$ $\square$

### *Infinite Sampling*

In the previous examples we considered events that were defined finitely. Let's now take a fair coin toss, but repeat it an arbitrary number of times, $n$. We can model a probability space which defines the odds of seeing heads for the first time after so many tosses:

$$\Omega = \mathbb{N} \quad \text{with} \quad \mathbb{P}(n) = \mathbb{P}(\underbrace{T, T, T, ..., T}_{n-1 \text{ times}}, H) = 1/2^n$$

Note that $\sum_{n\geq 1} 1/2^n = 1$, which satisfies the axiom $\mathbb{P}(\Omega) = 1$.

When we give the coin a bias, $p$, toward heads, $\mathbb{P}(n) = (1-p)^n \cdot p$ instead. Even still, $\mathbb{P}(\Omega) = 1$, which you can show for yourself.

Funky stuff can happen when you deal with infinities. Take a dartboard $D$ with area 1, and notate an arbitrary point on the dartboard $p$. Then, we can show that $\forall p \in$ dartboard, the odds of you hitting $p$ are 0.

PROOF.

We claim that $\forall p, \mathbb{P}(p) = 0$. Take $p$ to be arbitrary. Let $\delta := d(p, \partial D)$. Consider $B(p, \varepsilon)$, the open ball around $p$, with $\delta > \varepsilon > 0$. $\mathbb{P}(B(p, \varepsilon)) = \pi\varepsilon^2$

$\implies \mathbb{P}(p) < \pi\varepsilon^2 \implies \mathbb{P}(p) = 0$, since $\varepsilon$ arbitrary. □

### *Inclusion-Exclusion Formula*

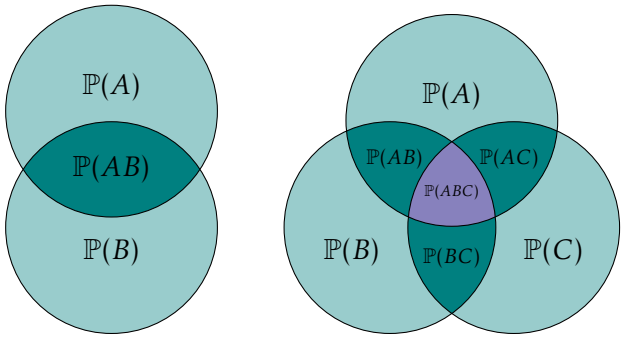
One of the Kolmogorov Axioms is that $\mathbb{P}(\bigcup_{n\geq 1} A_n) = \sum_{n\geq 1} \mathbb{P}(A_n)$ for pairwise disjoint events. But what about events for which $A_i \cap A_j \neq \emptyset$? We can work out the first few iterations:

(1) $\mathbb{P}(A) = (A)$

(2) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(AB)$

(3) $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(AB) - \mathbb{P}(AC) - \mathbb{P}(BC) + \mathbb{P}(ABC)$

As you can see, the expansions get progressively more unruly as the number of non-disjoint events increase. We can think of these terms as all the possible intersections and disjoint areas of overlapping sets.



These equations can be generalized. Define the union of events $\{A_1, ..., A_n\}$ for which $A_i \cap A_j \neq \emptyset$ may be true.

**1.3 Inclusion-Exclusion Formula**

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \mathbb{P}(A_i) \\ &- \sum_{1\le i_1<i_2\le n} \mathbb{P}(A_{i_1} \cap A_{i_2}) \\ &+ \sum_{1\le i_1<i_2<i_3\le n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &-... + (-1)^{n-1}\mathbb{P}(A_1A_2...A_n) \\ &\downarrow \\ &= \sum_{k=1}^{n} (-1)^{k-1} \underbrace{\sum \mathbb{P}(A_{i_1}A_{i_2}...A_{i_k})}_{1\le i_1<i_2<...<i_k\le n}\end{aligned}$$

## *Random Variables*

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we consider a *random variable* $X$ to be a function $f : \Omega \to \mathbb{R}$.

For example, consider rolling of two die, with $\Omega = [6] \times [6]$, as shown before. Let $X$ be the sum of the two rolls. This can be thought of as a function from the sample space to $\mathbb{R}$, $X((i, j)) = i + j$, and is random. We have:

$$\mathbb{P}(X = 5) = \mathbb{P}(\omega \in \Omega : X(\omega) = 5) = \mathbb{P}(X^{-1}(5))$$

More experiments: You roll a die 10 times. Define $N : \Omega \to \mathbb{R}$ to be the number of 6's that appear. What is $\mathbb{P}(N = 4)$?

Let $\Omega = [0, 1]$ and $A(\omega) = \tan(\frac{\pi}{2}\omega)$. What is $\mathbb{P}(A \ge 1)$? What is $A(1)$?

*Discrete* We'll consider 3 classes of random variables. Let $X$ be a random variable. Then $X$ is called discrete if there is a subset $S$ of $\mathbb{R}$, finite or countable, such that $\mathbb{P}(X \in S) = 1$.

Let $S_X := \{x \in \mathbb{R} : \mathbb{P}(X = x) > 0\}$ be the set of events for which a positive probability is assigned. Then the *probability mass function* of $X$ is

$$p_X(x) = \mathbb{P}(X = x) \text{ for } x \in S_X$$

*Ex.* The sum of two die rolls, denoted $X(i, j)$, can take on values $S = \{1, 2, ..., 12\}$. Since $S$ is a finite subset of the reals, and $\mathbb{P}(X \in S) = 1$, $X$ is a discrete random variable.

*Continuous* $X$ is called continuous if $\exists f : \mathbb{R} \to [0, \infty)$ such that $\forall a, b : -\infty \le a < b \le \infty$ we have

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(x)dx$$

This is called the *probability density function*, or the distribution of $X$.

*Degenerate* $X$ is called degenerate if $\exists \omega : \mathbb{P}(X = \omega) = 1$

# II Conditional Probability

## Conditioning

Given events $A$ and $B \in \Omega$, the *conditional probability* $\mathbb{P}(A|B)$ is, in English, the "probability of $A$ given that $B$ occurs." When $\mathbb{P}(A) > 0$, we define this probability as the following:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$$

---

**Examples:**

1. Toss a coin 2 times. What is $\mathbb{P}$(Two heads | First is heads)?

$$\mathbb{P} = \frac{1/4}{1/2} = \frac{1}{2} \text{ from above}$$

2. Sample 3 balls from an urn. What is $\mathbb{P}$(2 yellow | At least 1 yellow)? Without knowing the parameters of the urn, we have:

$$\frac{\mathbb{P}(\text{2 yellow and at least 1 yellow})}{\mathbb{P}(\text{At least 1 yellow})} = \frac{\mathbb{P}(2Y)}{\mathbb{P}(\geq 1\ Y)}$$

---

We can derive from the definition of conditional probability the following identity:

$$\mathbb{P}(ABC) = \mathbb{P}(A)\mathbb{P}(BC|A) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|AB)$$

which generalizes easily to

$$\mathbb{P}(A_1A_2...A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)...\mathbb{P}(A_n|A_1A_2...A_{n-1})$$

where $A_1, A_2, ..., A_n$ have $\mathbb{P} > 0$.

**2.1 Law of Total Probability**
Given that $A_1, B_1, B_2, ..., B_n \in \Omega$, that $B_1, B_2, ..., B_n$ partition $\Omega$, and that $P(B_i) > 0$, we have

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(AB_i) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

A further manipulation of our definition of conditional probability brings us the following:

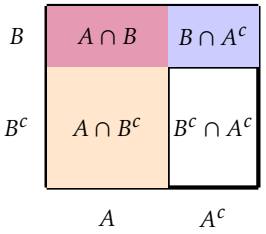
**2.2 Bayes' Formula**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \quad \text{given that all are well-defined}$$

$\mathbb{P}(B)$ can further be expanded as $\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$, which makes for a slightly more complex variation of Bayes' Formula.

## INDEPENDENCE

Two events $A, B$ are called *independent* if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. We have the implication, then, that two independent events $A$ and $B$ satisfy $\mathbb{P}(A|B) = \mathbb{P}(A)$ and vice-versa. We can generalize the definition of independence to the following:




Events $A_1, A_2, ..., A_n$ are independent IFF

$$\forall I \in [n] \quad \mathbb{P}\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} \mathbb{P}(A_i)$$

We can also extend our concept of independence to variables. Let $X_1, X_2, ..., X_n$ be random variables. They are mutually independent if

$$\forall B_1, B_2, ..., B_n \subseteq \mathbb{R} \quad \mathbb{P}(\cap_{1\le i\le n} X_i \in B_i) = \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i)$$

We call discrete random variables $X_1, ..., X_n$ independent if

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$$

where $\mathbb{P}(X_i = x_i) > 0$ (or else the condition would be trivial). Note that this definition is derived from our broader definition of random variables, and thus necessitates a proof:

PROOF.

Let $S_i := \{x_i : \mathbb{P}(X_i = x_i) > 0\}$. We'll first show that general independence implies our definition for discrete variables, and then the converse.

$(\implies)$ Fix $x_1, x_2, ..., x_n$ such that $x_i \in S_i \ \forall 1 \le i \le n$. Then we can take $B_i := \{x_i\}$, and we are done.

$(\impliedby)$ Fix $B_1, B_2, ..., B_n \in \mathbb{R}$. Then: $\{X_i \in B_i\} = \{X_i \in B_i S_i\} \cup \{X_i \in B_i S_i^c\}$. Based on our understanding of discrete random variables, though, $\mathbb{P}(X_i \in B_i S_i^c) = 0 \implies \{X_i \in B_i\} = \{X_i \in B_i S_i\}$. $S_i$ contains disjoint events, and so

$B_iS_i$ does as well. We can then say $\mathbb{P}(X_i \in B_i) = \sum_{x_i \in B_iS_i} \mathbb{P}(X_i = x_i)$

$$\begin{aligned}
\implies \prod_{i=1}^{n} \mathbb{P}(X_i \in B_i) &= \prod_{i=1}^{n} \left[ \sum_{x_i \in B_iS_i} \mathbb{P}(X_i = x_i) \right] \\
&= \sum_{x_i \in B_iS_i} \left[ \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \right] \\
&= \sum_{x_i \in B_iS_i} \mathbb{P}(X_1 = x_1, ..., X_n = x_n) \quad \text{from hypothesis} \\
&= \mathbb{P}\left( \bigcup_{x_i \in B_iS_i} \{X_1 = x_1, ..., X_n = x_n\} \right) \\
&= \mathbb{P}(X_1 \in B_1S_1,\ X_2 \in B_2S_2, ...,\ X_n \in B_nS_n) \\
&= \mathbb{P}(X_1 \in B_1,\ X_2 \in B_2, ...,\ X_n \in B_n) \text{ since we know } X_i \in S_i
\end{aligned}$$

□

We also have a notion of independence for continuous random variables $X_1, ..., X_n$:

$$\mathbb{P}(X_i \in (a_i, b_i) \text{ for } 1 \leq i \leq n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in (a_i, b_i))$$

for any $-\infty \leq a_i < b_i \leq \infty$.

---

**Examples:** Suppose we sample $k$ elements from $[n] = \{1, 2, 3, ..., n\}$, with a random variable $X_i$ being our $i^{\text{th}}$ sample. With replacement, we have

$$\Omega = [n]^k, \mathbb{P}(X_j = i) = \frac{1}{n} \quad \text{and} \quad \mathbb{P}(X_1 = i_1, X_2 = i_2, ..., X_k = i_k) = \frac{1}{n^k} = \prod_{j=1}^{k} \frac{1}{n}$$

Without replacement, we have $\mathbb{P}(X_1 = i_1, X_2 = i_2, ..., X_k = i_k) = \frac{1}{(n)_k}$

$$\begin{aligned}
\mathbb{P}(X_j = i) &= \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdot ... \cdot \frac{1}{n-j-1} \cdot ... \cdot \frac{n-k+1}{n-k+1} \\
&= \frac{(n-1)(n-2)...(n-k+1)}{n(n-1)(n-2)...(n-k+1)} = \frac{1}{n}
\end{aligned}$$

Note that $X_1, ..., X_k$ are *not* independent when drawn without replacement.

---

We call events $B_1$ and $B_2$ *conditionally independent* if

$$\mathbb{P}(B_1B_2|A) = \mathbb{P}(B_1|A)\mathbb{P}(B_2|A)$$

for $A : \mathbb{P}(A) > 0$. In general, we have

$$\mathbb{P}\left( \bigcap_{i \in I} B_i | A \right) = \prod_{i \in I} \mathbb{P}(B_i|A)$$

## PROBABILITY DISTRIBUTIONS

Before we define a distribution, note some preliminaries:

1. (a) If we have a set of independent random variables, $X$ and $Y$, then $f(X)$ and $g(Y)$ are also independent for any $f, g : \mathbb{R} \to \mathbb{R}$.
   (b) Similarly, if we have independent random variables $X_1, X_2, ..., X_n, X_{n+1}, ..., X_m$ and functions $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$, then $f(X_1, ..., X_n)$ and $g(X_{n+1}, ..., X_m)$ are independent from one another.
   (c) Lastly, a set of independent random variables $X_1, X_2, ..., X_n$ remains independent for $f(X_1), f(X_2), ..., f(X_n)$ for $f$ bijective.

2. A discrete random variable $X$ can take possible values $S_X := \{x : \mathbb{P}(X = x) > 0\}$. Define the *range* of the variable $X$ as $\text{range}(X) := \{X(\omega) : \omega \in \Omega\}$. Note that, in general, $S \neq \text{range}(X)$.
   (a) For example, let $\Omega = \mathbb{R}$ and $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \frac{1}{3}$. Let the random variable $X(\omega) = \omega$. Then, $S = \{1, 2, 3\}$, while $\text{range}(X) = \Omega$ itself.

### *Binomial Distribution*

Flip a coin $n$ times in succession. We can express the sample space as $\Omega = \{0, 1\}^n$. The probability of a particular arrangement, $\mathbb{P}(\omega_1, \omega_2, ..., \omega_n)$ is $p^i(1 - p)^j$, where $i$ and $j$ tally the number of heads (1) or tales (0), respectively.

Let $H$ determine, for a sequence of $n$ flips, how many heads appear. $H$ is then a discrete random variable which takes on values $S = \{0, 1, ..., n\}$.

The probability that there are exactly $k$ heads, $\mathbb{P}(H = k)$, is equal to the probability expressed above for $i = k$, times the number of unique arrangements such that $i = k$.

$$\mathbb{P}(H = k) = \sum_{\text{arrangements}} p^k(1 - p)^{n-k} = \binom{n}{k} p^k(1 - p)^{n-k}$$

We call this probability the *binomial distribution*, or $X \sim \text{Bin}(n, p)$, where a discrete random variable $X$ takes on possible values $\{0, 1, ..., n\}$ and $\mathbb{P}(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$ for $k \in [0, n]$

### *Bernoulli Distribution*

We have $X \sim \text{Ber}(p)$ if $X$ takes on either 0 or 1, and $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. The above coin example contains both Bernoulli and binomial distributions.

Let $B_1, B_2, ..., B_n$ be $\text{Ber}(p)$ and independent. Then we have that $S = B_1 + B_2 + ... + B_n$ is $\text{Bin}(n, p)$.

PROOF.

Fix $k \in \{0, 1, ..., n\}$. Then

$$\begin{aligned}\mathbb{P}(S = k) &= \mathbb{P}\left[\bigcup_{i\in\{0,1\} \text{ for } \#(i=1)=k} \{B_1 = i_1, B_2 = i_2, ..., B_n = i_n\}\right] \\ &= \sum_{i\in\{0,1\} \text{ for } \#(i=1)=k} [\mathbb{P}(B_1 = i_1, B_2 = i_2, ..., B_n = i_n)] \\ &= \sum p^k(1-p)^{n-k} = \binom{n}{k} p^k(1-p)^{n-k} \quad \square\end{aligned}$$

### *Geometric Distribution*

**A bit of notation**: Define $\bigotimes_{i=1}^{n} \text{Ber}(p)$ to be the probability $\mathbb{P}_n$ of a Bernoulli-distributed arrangement $\omega \in \Omega$, with $\Omega = \{0, 1\}^n$

To define a geometric distribution, we'll first need to give structure to the idea of infinite coin flips (or trials).

For $\Omega = \{0, 1\}^n$, we have probability $\bigotimes_{i=1}^{n} \text{Ber}(p)$. Dropping the last coordinate and adding an additional coordinate gives

$$\Omega = \{0, 1\}^{n-1} \iff \bigotimes_{i=1}^{n-1} \text{Ber}(p) \quad \text{and} \quad \Omega = \{0, 1\}^{n+1} \iff \bigotimes_{i=1}^{n+1} \text{Ber}(p)$$

We can then interpolate a "projective limit," in which

$$\Omega = \{0, 1\}^{\mathbb{N}} \iff \bigotimes_{i\geq 1} \text{Ber}(p)$$

Now we have something of a structure for infinite trials.

Let $G(\omega) = \inf\{i \geq 1 : \omega_i = 1\}$, or the first successful coin flip in an arbitrary series of flips. Then we have that

There are $k - 1$ unsuccessful trials at probability $p - 1$, followed by one success with probability $p$.

$$\mathbb{P}(G = k) = (1-p)^{k-1}p$$

This distribution is called *geometric*, with $G \sim \text{Geom}(p)$, where $p$ is the probability associated with $\bigotimes_{i\geq 1} \text{Ber}(p)$.

These three distributions are the most fundamental to the remainder of these notes. The following two distributions will come up occasionally, but are generally extra-curricular.

### *Hypergeometric Distribution*

Suppose we are sampling without replacement from an urn of yellow and purple balls. Define the random variable $X$ as the number of yellow balls chosen out of $k$ draws, from an urn containing $N$ total balls. Let there be $m$ yellow and $N - m$ purple balls.

We then say that $X \sim \text{Hyp}(N, m, k)$ with

$$\mathbb{P}(X = j) = \frac{\binom{k}{j}\binom{N-m}{k-j}}{\binom{N}{k}}$$

### *Rayleigh Distribution*

We consider the "birthday paradox." Let there be $k$ people in a room. What is the probability that no 2 people share the same birthday?

$$\mathbb{P} = 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdot \ldots \cdot \frac{365 - (k-1)}{365} = \prod_{i=2}^{k} \left(1 - \frac{i-1}{365}\right)$$

The "paradox," which isn't one, occurs when we ask the question "how many people should there be in a room so that it is *likely* (i.e. $\mathbb{P} > \frac{1}{2}$) that at least two people share birthdays." Computing a few guesses using the equation, noting that we are looking for the compliment of the above probability, gives the answer: 23 people. Which seems way too small, hence the name.

# III Random Variables

## PDFS AND CDFS

### *Probability Density Functions*

Consider a continuous random variable $X$. We can model the probability that $X$ lies in an interval as follows:

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(x)dx$$

where $f : \mathbb{R} \to [0, \infty]$

A necessary condition here is that $\int_{\mathbb{R}} f(x)dx$, which integrates over the entire sample space, must be equal to 1. We call $f$ the *probability density function* of $X$, which we've seen before.

---

**Examples:**

1. A uniformly distributed random variable $X \sim \text{Unif}[a, b]$ has PDF

$$f(x) = \frac{1}{b-a}$$

   noting that the area of a rectangle with width b-a and height $\frac{1}{b-a}$ is 1, as required.

2. A variable is called *Wigner distributed* if its PDF is $f(x) = \frac{2}{\pi}\sqrt{1-x^2}$, defined for $|x| \leq 1$, a.k.a. a semicircle. Verify for yourself that the sample space has probability 1.

---

**3.1 Analysis of PDF** If $X$ is a continuous r.v. with PDF $f$, then we have the following:

$$\forall a \in \mathbb{R} \quad f(a) = \lim_{\varepsilon \to 0} \frac{\mathbb{P}(X \in [a, a+\varepsilon])}{\varepsilon}$$

PROOF.

From above, we have $\mathbb{P}(X \in [a, a+\varepsilon]) = \int_a^{a+\varepsilon} f(x)$. Thus:

$$\lim_{\varepsilon \to 0} \frac{\mathbb{P}(X \in [a, a+\varepsilon])}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \int_a^{a+\varepsilon} f(x) = f(a)$$

□

Even if $f$ has discontinuities, the proposition above still holds so long as there are finitely many of them.

### *Cumulative Density Functions*

Let $F : \mathbb{R} \to [0, 1]$ be a function defining

$$\mathbb{P}(X \leq x) = F(x)$$

We call this the *cumulative density function* of $X$.

---

**Examples:**

1. For a coin toss $X \sim \text{Ber}(p)$, the CDF looks like: $F = 0$ for all $x < 0$, $F = 1 - p$ for $x \in [0, 1)$, and $F = p$ for $x \geq 1$.

2. For $X \sim \text{Geom}(p)$, the first heads in an infinite series of coin tosses, we have

$$F(x) = \begin{cases} 0 \quad \text{for} \quad x < 1 \\ 1 - (1-p)^k \quad \text{for} \quad k \leq x < k+1 \\ \vdots \\ 1 - (1-p)^{\lfloor x \rfloor} \quad \text{for} \quad x \geq 0 \end{cases}$$

---

Thus, $\mathbb{P}(X \in (a, b]) = F(b) - F(a)$. If $X$ is continuous, we may further remove the point $b$, and conclude $\mathbb{P}(X \in (a, b)) = F(b) - F(a)$, or similarly $\mathbb{P}(X \in [a, b]) = F(b) - F(a)$.

We can also define a CDF for discrete variables. Let $\rho(s) : S \to [0, 1]$ be its probability mass function, with $s \in S_X$ being a possible value $X$ can take. Then $F(x)$ is

$$\mathbb{P}(X \leq x) = \sum_{s \in S \text{ with } s \leq x} p(s)$$

Given that $S$ has no limit points, we have that $F$ is piecewise constant. Conversely, we can say that both $S = \{\text{discontinuities of} f\}$ and $X$ is discrete if $X$ is piecewise constant.

Similarly, for a continuous variable with PDF $f(s)$, $F(x)$ is

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(s)ds$$

**3.2 CDF $\leftrightarrow$ PDF**
If a continuous random variable exists everywhere but finitely many points, then we can say that $F'(x) = f(x)$

**3.3 Properties of CDFs**

1. $F$ is non-decreasing
2. $F$ is right continuous
3. $\lim_{x\to-\infty} = 0$ and $\lim_{x\to\infty} = 1$

The proofs for these are good exercise. For (2) and (3), define

$$E_n \uparrow E \text{ if } E_1 \subseteq E_2 \subseteq ... \subseteq E_n \subseteq ... \text{ and } E = \bigcup_{n\geq 1} E_n \quad \text{and}$$

$$E_n \downarrow E \text{ if } E_1 \supseteq E_2 \supseteq ... \supseteq E_n \supseteq ... \text{ and } E = \bigcap_{n\geq 1} E_n$$

for a sequence of events $\{E_n\}$. One can characterize limits and continuity using sequences, per analysis.

## EXPECTATIONS

For a discrete random variable with possible values $S$, define the *expectation* of $X$ as

$$\mathbb{E}X = \sum_{k\in S} k\mathbb{P}(X = k)$$

For a continuous r.v. with PDF $f$, its expectation is

$$\mathbb{E}X = \int_{\mathbb{R}} xf(x)dx$$

One can think of expectations as a statistical average of events over very long periods of time. In future sections, we'll define other qualities of random variables using expectations. Note the following property:

**3.4 Linearity of Expectation**
For random variables $X$ and $Y$ and a constant $k$, $\mathbb{E}[X+Y+k] = \mathbb{E}X+\mathbb{E}Y+k$

Similarly, for independent variables $X$ and $Y$, we can say $\mathbb{E}[XY] = \mathbb{E}X \cdot \mathbb{E}Y$. The proofs for both these results will come up when we study multivariate distributions.

---

**Examples:**

1. Consider $X \sim \text{Bin}(n, p)$. Then we have

$$\begin{aligned}\mathbb{E}X &= \sum_{k=0}^{n} k\mathbb{P}(X = k) = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} = np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} = np\end{aligned}$$

the last step borrowing from the binomial theorem.

2. The expectation of $X \sim \text{Bin}(n, p)$ can be figured using linearity of expectation:
$$X = H_1 + H_2 + ... + H_n$$
where $H_i$ is the indicator function for a coin flip. Then
$$\mathbb{E}X = \mathbb{E}H_1 + \mathbb{E}H_2 + ... + \mathbb{E}H_n = np$$
since $\mathbb{E}H_i$ always equals $p$ (you can show this for yourself easily).

3. The expectation of $X \sim \text{Geom}(p)$ is $\frac{1}{p}$. The proof for this requires a clever rearrangement of summations (a grid to count occurrences of $(1-p)^i$ may help).

4. Throwing darts at a circular board of radius $r_0$, we have PDF $f(t) = \frac{2t}{r_0^2}$ (a proof for this is needed, but it's not complicated). We can calculate $\mathbb{E}X$ as
$$\mathbb{E}X = \int_0^{r_0} \frac{2t}{r_0^2} = \frac{2t^3}{3r_0^2}\Big|_0^{r_0} = \frac{2}{3}r_0$$

---

We can take $X$, itself a function from $\Omega \to \mathbb{R}$, and map it to anything we'd like. Define $g : \mathbb{R} \to \mathbb{R}$ and have $g(X)$. This is another random variable with $\Omega \to g(X(\omega))$.

If $X$ is discrete with possible values $S$, then
$$\mathbb{E}g(X) = \sum_{x \in S} g(x)\mathbb{P}(X = x)$$

This is sometimes called the "law of the unconscious statistician" (LOTUS for short). Similarly, for a continuous variable with PDF $f$, we have
$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)f(x)dx$$

---

**Examples:**

It's Thanksgiving, and you're breaking the wish-bone with your sibling. Suppose this wishbone has length 1, is perfectly straight, and has a uniformly random breaking point $U \sim \text{Unif}[0, 1]$. We want to find the expected length of the winner's side, given that the person who breaks off a larger piece wins. We can express this as
$$\mathbb{E}\max(U, 1-U) = \int_0^1 \max(x, 1-x)dx$$
since the PDF of $U$ is just 1.

For $0 \le x \le \frac{1}{2}$, $\int_0^{1/2} \max(x, 1-x)dx = \int_0^{1/2} 1 - xdx = \frac{3}{8}$. Similarly, for $\frac{1}{2} \le x \le 1$, we have $\int_{1/2}^{1} \max(x, 1-x)dx = \int_{1/2}^{1} xdx = \frac{3}{8}$. Thus, $\mathbb{E}\max(U, 1-U) = \int_0^1 \max(x, 1-x)dx = \frac{3}{4}$. One can generalize this for arbitrary length $l$.

---

*Characterizations of Variables Using Expectation*

Define the $n^{th}$ *moment* of a random variable $X$ to be

$$\mathbb{E}[X^n]$$

We say that the first moment is the *mean*, as described above, and sometimes notated as $\mu$. The $n^{\text{th}}$ *central moment* is defined to be $\mathbb{E}[(X-\mu)^n]$.

It is a fact, though not important yet, that a bounded random variable's probability distribution ("bounded" in the sense that $X$ is equipped with some $r$ such that $\mathbb{P}(|X| < r) = 1$) can be uniquely determined by considering $\mathbb{E}[X^n]\ \forall n \ge 1$. We'll revisit this in Part V.

Define the *variance* of a random variable to be $\mathbb{E}[(X-\mu)^2]$, usually denoted $\text{Var}(X)$. Furthermore, $\sqrt{\text{Var}(X)} := \sigma$, the *standard deviation* of $X$. A useful alternate form for variance is:

$$\text{Var}(X) = \mathbb{E}[X^2] - [\mathbb{E}X]^2$$

---

**Examples:**

1. Let $X \sim \text{Ber}(p)$. Then $\mathbb{E}X = \sum_{k=0,1} k\mathbb{P}(X=k) = p$ and $\mathbb{E}[X^2] = \sum_{k=0,1} k^2\mathbb{P}(X=k) = p \implies \text{Var}(X) = p - p^2 = p(1-p) = pq$
2. Let $X \sim \text{Bin}(n,p)$. From above, $\mathbb{E}X = np$. A similar proof will yield $\mathbb{E}[X^2] = n(n-1)p^2 + np \implies \text{Var}(X) = np(1-p) = npq$
3. Let $X \sim \text{Unif}[a,b]$. Then $\mathbb{E}X = \int_{\mathbb{R}} \frac{x}{b-a}$, but since $f$ only defined for $a \le x \le b$, this is just $\int_a^b \frac{x}{b-a} = \frac{a+b}{2}$. Similarly, one can find that $\mathbb{E}[X^2] = \frac{b^2+ab+a^2}{3}$

---

Before the next section, here's a quick look at medians and quantiles. Define the *median* of a r.v. $X$ to be a real value $r$ such that

$$\mathbb{P}(X \le r) \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}(X \ge r) \ge \frac{1}{2}$$

This is a special case of the $\text{n}^{\text{th}}$ *quantile* of $X$, which is similarly

$$\mathbb{P}(X \le r) \ge p \quad \text{and} \quad \mathbb{P}(X \ge r) \ge 1-p$$

# IV Inter-Distributive Approximations

## THE GAUSSIAN

A general, awesome quality of probability is that—across an infinite number of arrangements in which one can observe probabilistic events—the distribution of real-life randomness almost always looks the same. We can characterize the eventual similarity of different rigorous distributions in a rigorous way.

As a precursor to this chapter, let's define a continuous random variable $X$ to be *Gaussian*, or *Normal* distributed, usually denoted $\mathcal{N}(0,1)$, if it has PDF

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Unlike previous distributions we've looked at, this one is continuous. We can state it's mean and variance as follows:

$$\mathbb{E}X = \int_{\mathbb{R}} x\varphi(x)dx = 0 \qquad \mathbb{E}[X^2] = \int_{\mathbb{R}} x^2\varphi(x)dx = 1 \qquad \implies \operatorname{Var}(X) = 1$$

Notice that, in our notation $\mathcal{N}(0,1)$, the first parameter equals our mean, and the second our variance. These are, in actuality, the parameters we use [i.e. $\operatorname{Var}(\mu, \sigma^2)$]

**4.1 Modifying $\mu$ and $\sigma^2$**
Let $X \sim \mathcal{N}(0,1)$. Then the random variable $Y = \sigma X + \mu$ is $\mathcal{N}(\mu, \sigma^2)$

The CDF of $X \sim \mathcal{N}(0,1)$ is notated

$$\mathbb{P}(X \leq x) = \Phi(x) = \int_{-\infty}^{x} \varphi(s)ds$$

With generality, the CDF of $Y \sim \mathcal{N}(\mu, \sigma^2)$ is

$$\mathbb{P}(Y \leq x) = \mathbb{P}\left(X \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Since we have $f(x) = F'(x)$, the PDF of $Y$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

## DE MOIVRE-LAPLACE THEOREM

Let $S_n \sim \text{Bin}(n,p)$ be a binomially-distributed random variable. From previous sections, note that $\mathbb{E}S_n = np$ and $\operatorname{Var}(S_n) = npq$

Define $S_n^* := \frac{S_n - np}{\sqrt{npq}}$ (this is something along the lines of $S_n$'s fluctuation about its mean). Then we have

**4.2 Binomial Central Limit Theorem**
Let $-\infty < a < b < \infty$. Then we have

$$\mathbb{P}(S_n^* \in [a,b]) \to \int_a^b \varphi(x)dx \quad \text{as } n \to \infty$$

Similarly, if $np(1-p) > 10$, as a rule of thumb,

$$\mathbb{P}(S_n^* \in [a,b]) \text{ is close to } \Phi(b) - \Phi(a)$$

If we are merely considering the probability that $S_n = k$ (without range), the approximation $\Phi(b) - \Phi(a)$ will fail, since $\Phi(k) - \Phi(k) = 0$. Thus, we preform a "continuity correction," in which we note that $\mathbb{P}(S_n = k) = \mathbb{P}(S_n \in [k - \frac{1}{2}, k + \frac{1}{2}])$, and compute the Gaussian as usual.

A corollary of the above theorem is the law of large numbers, which is

**4.3 Binomial Law of Large Numbers**

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1$$

where $S_n$ is a binomially-distributed variable.

## POISSON APPROXIMATIONS

Let $X$ be a discrete random variable. We say it is *poisson* distributed, or $X \sim \text{Poi}(\lambda)$, if

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Its mean and variance, both of which aren't too hard to verify, is $\mathbb{E}X = \lambda$ and $\text{Var}(X) = \lambda$.

**4.4 Poisson Approximating**
Let $S_n$ be $S_n \sim \text{Bin}(n,p)$ and $\lambda = np$. Then we have

$$\lim_{n\to\infty} \mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

This approximation is has an error bound
$|\mathbb{P}(\text{Bin}(n,p) \in A) - \mathbb{P}(\text{Poi}(\lambda) \in A)| \le np^2 \quad \forall A \subseteq \mathbb{N}$ and thus, if $np^2 \ll 1$, then a poisson approximation should be good.

Qualitatively, poisson distributions are appropriate in settings where

(a) One counts the number of rare events

(b) There are many possible events

(c) Each event has approximate independence

For example, we could consider the number of customers in the 2nd check-out line at the Mont-Royal Provigo (the probability that you or I are in this particular line is quite low, with approximate independence, and there are many millions of possible arrangements of Montrealers being or not being in this line at various times).

## EXPONENTIAL DISTRIBUTIONS

Let $\lambda > 0$ be a fixed real value. We say that a random variable $X$ is *exponentially distributed*, or $X \sim \text{Exp}(\lambda)$, if

$$\mathbb{P}(X > t) = e^{-\lambda t} \ \forall t \geq 0$$

We can easily derive its CDF, $\mathbb{P}(X \leq t) = 1 - \mathbb{P}(X > t) = 1 - e^{-\lambda t}$. Furthermore, we can see that its PDF is

$$F'(t) = f(t) = \begin{cases} 0 & \text{for } \ t < 0 \\ \lambda e^{-\lambda t} & \text{for } \ t \geq 0 \end{cases}$$

Employing some calculus, one can find that $\mathbb{E}X = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$. The exponential distribution has some surprising properties, the first of which is that it acts the same along shifted time-frames, as follows:

**4.5 Memoryless Property of Exponentials**
If $X \sim \text{Exp}(\lambda)$, then $\forall s, t > 0$ we have

$$\mathbb{P}(X > t + s \mid X > t) = \mathbb{P}(X > s)$$

PROOF.

$$\mathbb{P}(X > t + s \mid X > t) = \frac{\mathbb{P}(X > t + s, X > t)}{\mathbb{P}(X > t)} \quad (\mathbb{P}(X > t+s))$$

$$= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(X > s)$$

□

There also exists a converse, which states that *if* a random variable has the memoryless property as given above, then $\exists \lambda^*$ such that $X \sim \text{Exp}(\lambda^*)$

As with other distributions we've looked at, the exponential approximates, at its limit, a geometric distribution. Here are the formal conditions

**4.6 Exponential Approximation of the Geometric Distribution**
Let $X_i$ be independent Bernoulli variables, Ber($\frac{\lambda}{n}$), with $\frac{\lambda}{n} < 1$. Then define $G := \inf\{i : X_i = 1\}$, and note that this is a Geom($\frac{\lambda}{n}$). For $t > 0$, we have that

$$\mathbb{P}(G \geq tn) = \mathbb{P}(\text{Exp}(\lambda) \geq t) \quad \text{as} \quad n \to \infty$$

## POISSON PROCESSES

A Poisson process makes the modeling of temporal-spacial events extremely easy. Define a Poisson point process with rate $\lambda > 0$, or PPP($\lambda$), as a collection of randomly selected points $P$ on the line $I \in \mathbb{R}$, almost always $[0, \infty)$, where the time-space average of points appearing is $\lambda$ and the following conditions are satisfied:

(a) For any $r \in [0, \infty)$, there exists at *most* 1 point

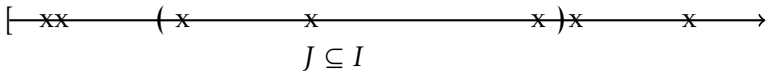
(b) For any subset $J \subseteq I$, $|P \cap J$ is a random variable Poi($\lambda|J|$), where $|P \cap J|$ is a measure of shared points between $P$ and $J$, and $|J|$ is simply the size of $J$.

For instance, one might consider

$J := [a, b]$. Then $|P \cap J| \sim \text{Poi}(\lambda(b - a))$

(c) For disjoint subsets $J_1, J_2, ..., J_n \subseteq I$, we have that $|P \cap J_1|, |P \cap J_2|, ..., |P \cap J_n|$ are mutually independent variables.

To note, mutual indendence implies that $A_1$ is pairwise independent with every other $A_i$, as usual, but also independent of any sets made up of $A_i$'s.



$J \subseteq I$

It is difficult to express the usefulness of Poisson point processes in theoretical language, so let's construct an example: we have a store in which, on average, 5 customers per hours buy something. We want to model the probability that we have exactly 2 sales between the hours of 9-10am, 3 between 10 and 10:30 am, and 5 between 10:30 and 1pm. This situation can be modeled as PPP(5).

$$\begin{aligned}
&\mathbb{P}(2 \in [9-10], 3 \in [10, 10.5], 5 \in [10.5, 13]) \\
&= \mathbb{P}\left(|P \cap [9, 10]| = 2, |P \cap [10, 10.5]| = 3, |P \cap [10.5, 13]| = 5\right) \\
&= \mathbb{P}(|P \cap [9, 10]| = 2) \cdot \mathbb{P}(|P \cap [10, 10.5]| = 3) \cdot \mathbb{P}(|P \cap [10.5, 13]| = 5) \\
&= \mathbb{P}[\text{Poi}(5) = 2]\mathbb{P}[\text{Poi}(\frac{5}{2}) = 3]\mathbb{P}[\text{Poi}((2.5)^3) = 5] \\
&= \frac{5^2(2.5)^3(12.5)^3 e^{-5} e^{-5} e^{-12.5}}{2(3!)(5!)}
\end{aligned}$$

If one orders points the points of a Poisson process, $p_1, p_2, ...$, with $p_1 < p_2 < ...$ on the interval $[0, \infty)$, a group of CDF and PDFs can be derived to characterize the probability that, before or at time $t$, one sees a particular $p_n$:

**4.7 Distribution of Points in a PPP**

For a time $t \in [0, \infty)$ and $p_n$ in a point process $\text{PPP}(\lambda)$, we have

$$\mathbb{P}(p_n \leq t) = 1 - \sum_{i=0}^{n-1} \frac{(\lambda t)^i e^{-\lambda t}}{i!} \quad \text{and} \quad \mathbb{P}(p_n = t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}$$

*Relation to the Gamma Distribution*

Define the Gamma function function $\Gamma(r) : \mathbb{R}_+ \to \mathbb{R}_+ = \int_0^\infty x^{r-1} e^{-x} dx$. This function is equivalent to $(r-1)!$ We say that a random variable $X$ is *Gamma distributed* with parameters $r$ and $\lambda$, or $X \sim \text{Gamma}(r, \lambda)$, if

$$\mathbb{P}(X = x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} \; \forall x > 0$$

One can see that the PDF of a point $p_r$ in a Poisson process with rate $\lambda$ is precisely $\text{Gamma}(r, \lambda)$.

# V Transformations &c.

## MOMENT GENERATING FUNCTIONS

For a random variable $X$, define its *moment generating function* $M : \mathbb{R} \to \mathbb{R}_+$ as

$$M_X(t) = \mathbb{E}[e^{tX}]$$

---

**Examples:**

Bernoulli We're flipping coins: $M(t) = \mathbb{E}[e^{tX}] = \sum_{k=\{0,1\}} e^{tk}\mathbb{P}(X = k) = (1-p) + e^t(p)$

Expodential Recall $f(x) = \lambda e^{-\lambda x}$. Then $M(t) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx = \begin{cases} \infty & \text{if } t \geq \lambda \\ \frac{\lambda}{\lambda - t} & \text{if } t < \lambda \end{cases}$

Geometric Recall $\mathbb{P}(X = k) = (1-p)^{k-1}p$. We have that $M(t) = \sum_{k\geq 1} e^{tk}(1-p)^{k-1}p =$

$$pe^t \sum_{k\geq 1} [e^t(1-p)]^k = \begin{cases} \infty & \text{if } e^t(1-p) \geq 1 \\ \frac{pe^t}{1-(1-p)e^t} & \text{if } e^t(1-p) < 1 \end{cases}$$

---

You may be wondering why $M_X(t)$ is called a "moment generating" function: by taking derivatives of $M$, one can actually extract the $n^{\text{th}}$ moments of $X$. Consider the discrete case:

With $S_X := \{s_1, s_2, ..., s_n\}$, the moment generating function is given by $M(t) = \sum_{i=1}^n e^{ts_i}\mathbb{P}(X = s_i)$. Differentiating, we get $M'(t) = \sum_{i=1}^n e^{ts_i}s_i\mathbb{P}(X = s_i)$, and can see that $M'(0)$ is the mean! Differentiating once more yields $M''(t) = \sum_{i=1}^n e^{ts_i}s_i^2\mathbb{P}(X = s_i) \implies M''(0) = \mathbb{E}[X^2]$, which is the *second* moment.

We can generalize to the following:

**5.1 Deriving Moments from MGFs**
Let $M_X(t) = \mathbb{E}[e^{tX}]$. When $X$ is discrete or $M(t)$ is finite close to 0, we have that

$$M_X^{(n)}(0) = \mathbb{E}[X^n]$$

### *Equal Distributions*

Let $X$ and $Y$ be two random variables, not necessarily defined identically. We say that $X$ and $Y$ are *equal in distribution*, denoted $X \overset{d}{=} Y$, if

$$\mathbb{P}(X = B) = \mathbb{P}(Y = B) \text{ for all subsets } B \subseteq \mathbb{R}$$

Though a proof won't be provided, $X$ and $Y$ have the same CDF or PDF if and only if $X \overset{d}{=} Y$. However, even if $M_X(t) = M_Y(t)$, it may be that $X \overset{d}{\neq} Y$.

**5.2 Inversion Theorem**
Suppose $X$ and $Y$ are such that $M_X^{(n)}(0) = \mathbb{E}[X^n]$ and $M_Y^{(n)}(0) = \mathbb{E}[Y^n]$, i.e. $M_X$ and $M_Y$ are both finite about a neighborhood of 0. Then we have that

$$M_X(t) = M_Y(t) \implies X \overset{d}{=} Y$$

### *Characteristic Functions*

$M_X(t) = \mathbb{E}[e^{tX}]$ may be modified slightly to yield the *characteristic function* of $X$. Define $\varphi_X(t) : \mathbb{R} \to \mathbb{C} = \mathbb{E}[e^{itX}] = M(it)$.

Decomposing, one writes $\mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX) + i\sin(tX)] = \mathbb{E}[\cos(tX)] + i\mathbb{E}[\sin(tX)]$. A significant advantage of $\varphi_X(t)$ is that, for any variable $X$ and any value $t$, $\varphi$ will *always* be bounded (see the trig substitution). If the $n^{\text{th}}$ moment of $X$ exists at all, we have $\varphi_X^{(n)}(0) = i^n\mathbb{E}[X^n]$

## FUNCTIONS OF RANDOM VARIABLES

Suppose $X \sim \text{Unif}\{-1, 0, 1, 2\}$ and $Y := X^2$. What is the distribution of $Y$? We can consider this question term-by-term:

$$\mathbb{P}(Y = 4) = \mathbb{P}(X^2 = 4) = \mathbb{P}(X \in \{2, -2\}) = \mathbb{P}(X = 2) = \tfrac{1}{4}$$
$$\mathbb{P}(Y = 1) = \mathbb{P}(X \in \{-1, 1\}) = \tfrac{1}{2}$$
$$\mathbb{P}(Y = 0) = \mathbb{P}(X = 0) = \tfrac{1}{4}$$

With $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$, we have found all $y \in S_Y := \{0, 1, 4\}$. Thus, the PMF of $Y$ is

$$\rho_Y(y) = \begin{cases} \frac{1}{4} & y = 0 \\ \frac{1}{2} & y = 1 \\ \frac{1}{4} & y = 4 \end{cases}$$

We can fomalize the definitions of $S_Y$ and $\rho_Y(y)$ as follows:

$$S_Y := \{y \in \mathbb{R} : \exists x \in S_X \text{ with } g(x) = y\} = \{g(x) : x \in S_X\}, \text{ where } Y = g(X)$$

$$\rho_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(X = \{x \in S_X : g(x) = y\}) = \sum_{x \in S_X \text{ s.t. } g(x) = y} \mathbb{P}(X = x)$$

**5.3 Character of a Transformation**
If $X$ is discrete, any transformation $Y = g(X)$ will also be discrete. If $X$ is continuous, however, $Y = g(X)$ may continuous, discrete, or both.

**Examples:**

1. Let $X \sim \text{Unif}[-1, 1]$, a continuous random variable, and $Y = g(X) = \mathbb{1}_{x \geq 0}$

We then have that $\mathbb{P}(Y=0)=\mathbb{P}(X<0)=\frac{1}{2}$ and $\mathbb{P}(Y=1)=\mathbb{P}(X\geq 0)=\frac{1}{2}$. Thus, $Y$ is a discrete variable with $S_Y=\{0,1\}$.

Now suppose that $Y=X^2$, with $X$ defined as above. Then we have that $\mathbb{P}(Y\leq x)=\begin{cases}0 & x<0\\ 1 & x\geq 1\end{cases}$ since $Y$ is non-negative. Further, $\mathbb{P}(Y\leq x)=\mathbb{P}(X\in[-\sqrt{x},\sqrt{x}])=\sqrt{x}$ for $x\in[0,1)$. We observe both a discrete and continuous "character" for $Y$.

2. Let $X\sim[0,1]$ and $Y=-\ln(1-X)$. Then

$$\begin{aligned}\mathbb{P}(Y\leq t)&=\mathbb{P}(-ln(1-X)\leq t)=\mathbb{P}(\ln(1-X)\geq -t)\\ &=\mathbb{P}(1-X\geq e^{-t})=\mathbb{P}(X\geq 1-e^{-t})\\ &\implies Y\sim \text{Exp}(1)\end{aligned}$$

One can also conclude that $Y=-\frac{1}{\lambda}\ln(1-X)$, where $X\sim\text{Unif}[0,1]$, is $\text{Exp}(\lambda)$. Generally speaking, one can extract any distribution one likes from a uniform random variable (or really any continuous variable).

---

### *Affine Transformations*

Let $X$ have a PDF/PMF $f_X(x)$, and let $Y:=aX+b$, where $a$ and $b$ are real-valued. One calls this an "affine transformation" (not just in probability!).

Let $a>0$. Then $\mathbb{P}(Y\leq x)=\mathbb{P}(aX+b\leq x)=\mathbb{P}\left(X\leq\frac{x-b}{a}\right)$. Thus, $f_Y(x)=F_X'\left(\frac{x-b}{a}\right)=\frac{1}{a}f_X\left(\frac{x-b}{a}\right)$. Similarly, when $a<0$, we have that $f_Y(x)=-\frac{1}{a}f_X\left(\frac{x-b}{a}\right)$. Generalizing, one yields

**5.4 Distribution of an Affine Transformation**
$f_Y(x)=\frac{1}{|a|}f_X\left(\frac{x-b}{a}\right)$, where $X$ is a random variable and $Y:=aX+b$.

One can thus show that, for $X\sim\mathcal{N}(0,1)$, $Y=\sigma X+\mu$ is $\mathcal{N}(\mu,\sigma^2)$, and further that $aY+b$ is $\mathcal{N}[a\mu+b,(a\sigma)^2]$

The following is applicable to a much broader set of transformations:

**5.5 General PDF of a Transformation**
Let $X$ be a random variable with density function $f_X$. Let $Y=g(X)$, with $g$ differentiable everywhere and the set of points $\{x:g'(x)=0\}$ finite. We then have that

$$f_Y(y)=\sum\frac{f_X(x)}{|g'(x)|}\text{, summing over all }\{x:g(x)=y\ \text{ with }\ g'(x)\neq 0\}$$

For example, let $Y = (X-1)^2$. Then $f_Y(y) = \sum_{x=1\pm\sqrt{y}} \frac{f_X(x)}{|g'(x)|} =$

$$\frac{f_X(1+\sqrt{y})}{|g'(1+\sqrt{y})|} + \frac{f_X(1-\sqrt{y})}{|g'(1-\sqrt{y})|} = \frac{f_X(1+\sqrt{y}) + f_X(1-\sqrt{y})}{2\sqrt{y}}$$

# VI Multivariate Distributions

## RANDOM VECTORS

### *Discrete Case*

Consider a vector $\vec{X} := (X_1, X_2, ..., X_n)$, where $X_i$ are all random variables. We can think of $\vec{X}$ as a random variable itself, with $\vec{X} : \Omega \to \mathbb{R}^n$. If we want to describe the probability of that particular vector takes on particular values $(x_1, ..., x_2)$, we call the appropriate function a *joint density function*.

If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X_1, ..., X_n : \Omega \to \mathbb{R}$ are discrete random variables, then the joint PMF is

$$\rho_{X_1,...,X_n}(x_1, ..., x_n) = \mathbb{P}(X_1 = x_1, ..., X_n = x_n)$$

As with the univariate case, the probability of all events must sum to 1:

$$\sum_{x \in S_X} \rho_{X_1,...,X_n}(x) \quad \text{with} \quad S_X := [S_{X_1}] \times ... \times [S_{X_n}]$$

Note that $[S_{X_1}] \times ... \times [S_{X_n}]$ may be larger than the possible values for $\vec{X}$, $x \in S_{\vec{X}}$, for which $\mathbb{P}(\vec{X} = x) > 0$. In this sense, the expression above is also true when summing over $x \in S_{\vec{X}}$.

Suppose $X_1, ..., X_n$ are random variables and $g : \mathbb{R}^n \to \mathbb{R}$ is a function which does not "blow up" at any input (such that the expression below will make sense). We have that the expectation

$$\mathbb{E}[g(X_1, ..., X_n)] = \sum_{(x_1,...,x_n) \in S_X} g(x_1, ..., x_n)\rho(x_1, ..., x_n) \quad \text{with } S_X \text{ defined as above}$$

One sees that when $g$ is the identity function, we get a plain expectation $\mathbb{E}[(X_1, ..., X_n)]$

Suppose we want to single out the probability of a particular coordinate $X_i \in \vec{X}$. The probability that $X_i = k$ can be derived from our joint PMF, where one sums over all possible values of $X_j$ *except* $X_i$. Define the *marginal probability function*:

$$\rho_{X_i}(k) = \sum_{(x_1,...,x_{i-1},k,x_{i+1},...,x_n) \in S_X} \rho(x_1, ..., x_{i-1}, k, x_{i+1}, ..., x_n)$$

One can similarly "single out" a whole range of coordinates within $\vec{X}$, say $(x_1, ..., x_m)$ where $m < n$, by fixing these values in a summation of $\rho$ across $S_X$.

### *Multinomial Distribution*

Suppose we are rolling an $r$-sided die, where the probability that one rolls side $i$ is $p_i$, with $0 \leq p_i \leq 1$. Thus, we require that $p_1 + ... + p_r = 1$. Let $X_i$ denote

the number of $i$-side rolls one sees in a series of $n$ rolls. We have that

$$\mathbb{P}(X_1 = k_1, ..., X_n = k_n) = \binom{n}{k_1, ..., k_r} p_1^{k_1} ... p_r^{k_r}$$

where $\binom{n}{k_1,...,k_r}$, the "multinomial" term, is defined to be $\frac{n!}{k_1!...k_r!}$.

In generality, if we are counting instances of $r$ events in $n$ trials such that $p_1 + ... + p_r = 1$ and the "tallies" $X_1 = k_1, ..., X_r = k_r$ sum to $n$, we have

$$(X_1, ..., X_r) \sim \text{Mult}(n, r, p_1, ..., p_r)$$

with the distribution defined above. Note that the binomial distribution is a particular case of the multinomial distribution with $r = 2$, $p_1 = p$, and $p_2 = 1 - p$.

### *Continuous Case*

As before, let $X_1, ..., X_n$ be random variables (continous this time). We say that a function $f : \mathbb{R}^n \to \mathbb{R}^+$ is a *joint density* of $(X_1, ..., X_n)$ if, for all $B \subseteq \mathbb{R}^n$

$$\mathbb{P}\left((x_1, x_2, ..., x_n \in B)\right) = \overbrace{\int_B \int_B ... \int_B}^{\text{n times}} f(x_1, x_2, ..., x_n)\, dx_1 ... dx_n$$

Note that, if $f$ is a valid density, we require that $\int_{\mathbb{R}} \int_{\mathbb{R}} ... \int_{\mathbb{R}} f(x_1, x_2, ..., x_n) = 1$

As we did in the discrete case, we can "single out" a coordinate in $\vec{X}$, and write

$$f_{X_i}(x) = \overbrace{\int_{\mathbb{R}} ... \int_{\mathbb{R}}}^{n-1 \text{ times}} f(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_n)$$

This, too, is called the marginal density function.

When working with continuous variables, one should be careful to ensure that a density function even *exists* for a set of random variables. This is not always the case. As an example, suppose $X = Y$. See that $\mathbb{P}(X = Y) = 1$. But then we have

$$1 = \int_{\{x=y\}} \int f(x, y) dx dy = \int_{\mathbb{R}} \int_x^x f(x, y) = 0 \quad \lightning$$

Generally, and informally, if $\mathbb{P}(X, Y \in A) = 1$ for some subset $A$, where the "area" or "measure" of $A = 0$, then no joint density exists. For the above example, see that $\mathbb{P}((X, Y) \in y = x) = 1$, where $x = y$ describes a line of inherent area 0. Conversely, if $X, Y$ *do* have a joint density, then $\mathbb{P}((X, Y) \in A) = 0$ if $A$ has area 0. In one final arrangement of words, if $\mathbb{P}((X, Y) \in A) > 0$, then the area of $A > 0$.

## INDEPENDENCE

Suppose variables $X_1, ..., X_n$, continuous or discrete, are defined on a common space (probability space). Their *joint CDF* is given by

$$F(x_1, ..., x_n) = \mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n)$$

as usual. Also as usual, the joint CDF always determines the distribution of $X_1, ..., X_n$.

**6.1 Determining Multivariate Independence**
$X_1, ..., X_n$ are independent IFF

$$F(x_1, ..., x_n) = F(x_1)...F(x_n)$$

Furthermore, for continuous variables,

$$f(x_1, ..., x_n) = f_{X_1}(x_1)...f_{X_n}(x_n)$$

implies independence. If $X_1, ..., X_n$ are already known to be independent, then their joint density always exists, and can be derived using the univariate densities as above.

For discrete variables $X_1, ..., X_n$ with PMFs $\rho_{X_i}$, $X_1, ..., X_n$ are independent IFF

$$\rho(k_1, ..., k_n) = \rho_{X_1}(k_1)...\rho_{X_n}(k_n)$$

**Examples:**

1. We'll consider our dart board example again, where $D := B(0, r_0) = \{x, y : x^2 + y^2 \leq r_0\}$. Define Leb$(A)$ for some $A \subseteq R^n$ to be the $n^{\text{th}}$-dimensional volume of $A$. Then $\mathbb{P}[(X, Y) \in A] = \frac{\text{Leb}(A)}{\pi r_0^2}$. Since we can write $\int\int_D \mathbb{1}_{(x,y)\in A} = \text{Leb}(A)$, we find that the PDF of $(X, Y)$ is $f(x, y) = \frac{1}{\pi r_0^2}\mathbb{1}_{(x,y)\in A}$.

   To calculate the marginal densities $f_X$ and $f_Y$, we have

$$\int_a^b f_X = \int\int_D \frac{1}{\pi r_0^2}\mathbb{1}_{X\in[a,b]} = \int_a^b \int_{-\sqrt{r_0^2-x^2}}^{\sqrt{r_0^2-x^2}} \frac{1}{\pi r_0^2} = \int_a^b \frac{2\sqrt{r_0^2 - x^2}}{\pi r_0^2} \implies f_X = \frac{2\sqrt{r_0^2 - x^2}}{\pi r_0^2}$$

   Similarly, we find that $f_y = \frac{2\sqrt{r_0^2-y^2}}{\pi r_0^2}$. Notice that $f_X f_Y \neq f_{X,Y}$.

"Independent and identically distributed"

2. Let $(X_i, i \geq 1)$ be a sequence of IID random variables, $X_i \sim \text{Unif}\{1, ..., 6\}$. Define the variables $S = X_1 + X_2$ and $I = \mathbb{1}_{X_1=1}$. One variable gives information on the sum of $X_1$ and $X_2$, and the other about whether or not $X_1$ was rolled a 1. It *may* be intuitive to think that these variables are independent of eachother, but see that $\mathbb{P}(S = 12, I = 1) = 0$ while $\mathbb{P}(S = 12)\mathbb{P}(I = 1) = \frac{1}{36}\frac{1}{6}$. We conclude that $S$ and $I$ are not independent.

With the same setup, define $N := \min(k : X_k + X_{k+1} \in \{2, 6\})$, i.e. the first pair of adjacent rolls that sum to *either* 2 or 6. The probability that $X_i + X_{i+1} \in \{2, 6\}$ is $\frac{1}{6}$ (for 2, this is 1/36, and for 6, this is 5/36). Thus, $N \sim \text{Geom}(\frac{1}{6})$.

Now consider $Y = S_N$, i.e. the first sum for which $X_k + X_{k+1} \in \{2, 6\}$ holds. Then $\mathbb{P}(Y = 2)$ or $\mathbb{P}(Y = 6)$. We'll look at the first case only, since the second follows similarly:

$$\mathbb{P}(Y = 2, N = k) = \sum_{k \geq 1} \mathbb{P}(Y = 2, N = k) = \sum_{k \geq 1} \left(\frac{5}{6}\right)^{k-1} \frac{1}{36} \implies \mathbb{P}(Y = 2) = \frac{1}{6}$$

We deduce, then, that $\mathbb{P}(Y = 2, N = k) = \mathbb{P}(Y = 2)\mathbb{P}(N = k)$, and similarly for 6. Thus, $N$ and $Y$ are independent.

---

Though a proof won't be provided, we have generally that, for IID variables $(Z_n, n \geq 1)$ where $\exists B \subseteq \mathbb{R} : \mathbb{P}(Z_i \in B) > 0$, the variables $N := \min(i : Z_i \in B)$ and $Y = Z_N$ are independent.

## CHANGE OF VARIABLES IN MULTIVARIATE SETTING

Suppose random variables $(X, Y)$ have a joint density $f_{X,Y}$. Consider the set $K \in \mathbb{R}^2$ such that $f = 0$ outside $K$, and let $L \in \mathbb{R}^2$ be arbitrary. Define a bijection $G : K \to L : (X, Y) \to (g(X, Y), h(X, Y))$, with $g(X, Y) = U$ and $h(X, Y) = V$. Since $G$ is a bijection, we also have $G^{-1} : L \to K : (U, V) \to (p(U, V), q(U, V))$.

We are looking for the distribution $f_{U,V}$, ultimately, and now see that $L \in \mathbb{R}^2$ is the set for which $f_{U,V} = 0$ outside of it.

**6.2 PDF with a Change of Variables**

If the partial derivatives contained in the Jacobian $\begin{bmatrix} \frac{\partial p}{\partial u} & \frac{\partial p}{\partial v} \\ \frac{\partial q}{\partial u} & \frac{\partial q}{\partial v} \end{bmatrix}$ exist and are continuous on $L$, and $\det(\text{Jac}(u, v)) \neq 0$ for *any* $u, v \in L$, then $(U, V)$ has a joint density which is given by

$$f_{U,V}(u, v) = f(p(u, v), q(u, v)) \cdot |\det(\text{Jac}(u, v))|$$

where $X \to U$, $Y \to V$, and $f_{X,Y}$ is the joint distribution of $X$ and $Y$.

Let's return to a dart board of radius 1 to demonstrate a reasonable change of variables. Let $(X, Y) \sim \text{Unif}(D)$. If $(R, \theta)$ represent a polarized version of these coordinates, with $(R, \theta) \sim \text{Unif}(D)$, we can find $F_{R,\theta}$ without too much trouble. Let $u, v$ be the variables for $R$ and $\theta$, respectively. Then

$$F_{R,\theta}(u, v) = \mathbb{P}(R \leq u, \theta \leq v) = \frac{uv^2}{2\pi} \implies f_{R,\theta} = \frac{\partial^2}{\partial u \partial v} \frac{uv^2}{2\pi} = \frac{u}{\pi}$$

since we are now measuring the $v$-degree "slice" of a circle with radius $u$. The marginal densities are then found by integrating over the opposite variable:

$$f_R(u) = \int_0^{2\pi} \frac{u}{\pi} = 2u \quad \text{and} \quad f_\theta(v) = \int_0^1 \frac{u}{\pi} = \frac{1}{2\pi}$$

This solution is perfectly valid, and is done from first principles. Using a change of variables, however, we can consider more complicated setups. Assume, instead of $(X, Y) \sim \text{Unif}(D)$, we have $X, Y \sim \mathcal{N}(0, 1)$ both independent normal.

This shouldn't change any statements about the distribution.

To satisfy the conditions of our theorem, we'll consider $(X, Y) \in \mathbb{R}^2 \setminus \{0\}$. Then let $G : (X, Y) \to (R, \theta)$. We are concerned with $p, q : (u, v) \to (x, y)$, where $u, v$ are the parameters of $R, \theta$, respectively, i.e. the functions that invert our transformation. This is easy, though, as $y = u \sin(v)$ and $x = u \cos(v)$.

Without performing the calculations here, we see that $\det(\text{Jac}(u, v)) = u$, which is positive everywhere (we removed the case where $R = 0$). Thus, one can write $f_{R,\theta} = f(p(u, v), q(u, v))u$. Remember that $X, Y$ were independent normals, so $f_{X,Y} = f_X f_Y = \varphi(x)\varphi(y) = \frac{1}{2\pi} e^{\frac{-(x^2+y^2)}{2}}$, and we write

$$f_{R,\theta} = \frac{1}{2\pi} \exp\left[\frac{-u^2(\cos^2(v) + \sin^2(v))}{2}\right] u = \frac{1}{2\pi} e^{\frac{-u^2}{2}} u \implies f_R = e^{\frac{-u^2}{2}}, f_\theta = \frac{1}{2\pi}$$

# VII Sums & Exchangeability

We've discussed univariate and multivariate densities, their relationship with one another via independence and marginal density functions, and the densities of some particular well-behaved transformations. Here, we'll consider sums of variables and some important symmetries one can take advantage of in problem-solving.

## SUMS OF VARIABLES

Let $X, Y$ be independent continuous variables with densities $f_X$ and $f_Y$. Then, the density of $X + Y$ is given by

$$f_{X+Y} = f_X * f_Y = \int_{\mathbb{R}} f_X(x) f_Y(z-x) dx$$

We call $f_X * f_Y$ the "convolution" of $f_X$ and $f_Y$.

**7.1 Density of Summed Normals**
Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2), Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then the variable $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. In fact, for any string of independent normals $X_i$ with mean $\mu_i$ and variance $\sigma_i^2$, then

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(a_1\mu_1 + ... + a_n\mu_n, a_1^2\sigma_1^2 + ... + a_n^2\sigma_n^2) \quad \text{where } a_i \text{ all constants}$$

**Example:**

1. Let $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\lambda)$ be independent. Then

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x) dx = \int_{\mathbb{R}} \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx = \lambda z e^{-\lambda z}$$

   We conclude that $X + Y \sim \text{Gamma}(z, \lambda)$.

2. Recall that a variable $X \sim \text{Gamma}(a, \lambda)$ if $f_X = \frac{\lambda^a x^{a-1} e^{-\lambda x}}{\Gamma(a)} \mathbb{1}_{x \geq 0}$. One can show, for two independent $X \sim \text{Gamma}(a, \lambda)$, $Y \sim \text{Gamma}(b, \lambda)$, the sum $X + Y \sim \text{Gamma}(a + b, \lambda)$.

   This exercise requires the change of variables $x = tz$.

3. Let $X, Y$ be IID Unif$[0, 1]$. Then

$$f_{X+Y} = \int_{\mathbb{R}} f_X(x) f_Y(z-x) = \int_0^1 dx \text{ for } z \in [0,1] \text{ and } \int_{z-1}^1 dx \text{ if } z \in [1,2]$$

   and 0 elsewhere

## EXCHANGEABILITY

Or "equal in distribution."

We've previously defined what it means for two variables to be identically distributed. The multivariate case is similar. We say two sequences $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n)$ are *identically distributed*, and write $(X_1, ..., X_n) \overset{d}{=} (Y_1, ..., Y_n)$, if

$$\mathbb{P}[(X_1, ..., X_n) \in B] = \mathbb{P}[(Y_1, ..., Y_n) \in B] \; \forall B \subseteq \mathbb{R}^n$$

This is identical to saying $F_{X_1,...,X_n}(k_1, ..., k_n) = F_{Y_1,...,Y_n}(k_1, ..., k_n) \; \forall k_i \in \mathbb{R}$.

We then say that a sequence $(X_1, ..., X_n)$ is *exchangeable* IFF, for any permutation $(\sigma(1), ..., \sigma(n))$, we have $(X_1, ..., X_n) \overset{d}{=} (X_\sigma(1), ..., X_\sigma(n))$. We have a few equivalent conditions for this

1. $F_{X_1,...,X_n}$ is a symmetric function IFF exch.
2. For continuous variables, $f_{X_1,...,X_n}$ is symmetric, IFF exch.
3. For discrete variables, $\rho_{X_1,...,X_n}$ is symmetric IFF exch.
4. If $X_1, ..., X_n$ are IID, then $(X_1, ..., X_n)$ are exchangeable. If they are exchangeable, they are identically distributed, but *may not be* independent.

PROOF OF (1).

Suppose that $F(x_1, ..., x_n)$, the joint CDF of $X_1, ..., X_n$, is a symmetric function, i.e. $F(x_1, ..., x_n) = F(x_{\sigma^{-1}(1)}, ..., x_{\sigma^{-1}(n)})$. We use $\sigma^{-1}$ instead of $\sigma$ for purely notational reasons. Then

$$F_{X_1,...,X_n}(x_{\sigma^{-1}(1)}, ..., x_{\sigma^{-1}(n)}) = F_{X_\sigma(1),...,X_{\sigma(n)}}(x_1, ..., x_n)$$

and we conclude that $(X_1, ..., X_n) \overset{d}{=} (X_{\sigma(1)}, ..., X_{\sigma(n)})$.

Now let $(X_1, ..., X_n) \overset{d}{=} (X_{\sigma(1)}, ..., X_{\sigma(n)})$. Then we have

$$\begin{aligned} F(x_1, ..., x_n) &= \mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n) \\ &= \mathbb{P}(X_{\sigma(1)} \leq x_1, ..., X_{\sigma(n)} \leq x_n) = \mathbb{P}(X_1 \leq x_{\sigma^{-1}(1)}, ..., X_n \leq x_{\sigma^{-1}(n)}) \\ &= F(x_{\sigma^{-1}(1)}, ..., x_{\sigma^{-1}(n)}) \implies F \text{ symmetric} \quad \square \end{aligned}$$

PROOF OF (4).

Let $X_1, ..., X_n$ be IID, and fix $x_1, ..., x_n$. Then

$$\begin{aligned} F_{X_1,...,X_n}(x_1, ..., x_n) &= F_{X_1}(x_1)...F_{X_n}(x_n) \text{ by independence} \\ &= F_{X_1}(x_1)...F_{X_1}(x_n) = F_{X_{\sigma(1)}}(x_1)...F_{X_{\sigma(n)}}(x_n) \text{ by ID dist.} \\ &= F_{X_{\sigma(1)},...,X_{\sigma(n)}}(x_1, .., x_n) \implies (X_1, ..., X_n) \text{ exchangeable} \end{aligned}$$

Now let $(X_1, ...., X_n)$ be exchangeable. Then

$$\begin{aligned}\mathbb{P}(X_1 \leq x_1) &= \mathbb{P}\Big[(X_1, ..., X_n) \in (-\infty, x] \times \mathbb{R}^{n-1}\Big] \\ &= \mathbb{P}\Big[(X_i, X_1, ..., X_{i-1}, X_{i+1}, ..., X_n) \in (\infty, x] \times \mathbb{R}^{n-1}\Big] \\ &= \mathbb{P}(X_i \leq x)\ \forall i \in [n] \quad \square\end{aligned}$$

For a given sequence $X_1, ..., X_n$, our notions of exchangeability extend to subsets, $X_1, ..., X_k$ and $X_{i_1}, ..., X_{i_k}$. In detail, we have that if $X_1, ..., X_n$ are exchangeable, then:

The proof for this is similar to those previous.

$$\forall \text{ distinct } 1 \leq k \leq n \text{ and } i_1, ..., i_k \in [1, n], \quad (X_{i_1}, ..., X_{i_k}) \overset{d}{=} (X_1, ..., X_k)$$

---

**Example:**

1. Let $X_1, X_2, X_3$ be IID $\sim$ Unif$[0, 1]$. We are interested in the probability that $X_1 = \max(X_1, X_2, X_3)$. Normally, one would cook up an equation for the joint PDF $f_{X_1,X_2,X_3}$ and integrate over certain bounds, but we can use exchangeability! Since $X_1, X_2, X_3$ are IID, they are exchangeable, and thus $\mathbb{P}(X_1 = \max(X_1, X_2, X_3) = \mathbb{P}(X_2 = \max(X_1, X_2, X_3)) = \mathbb{P}(X_3 = \max(X_1, X_2, X_3))$. Since one of these variables *must* be the maximum, and since we can remove the point-cases where $X_i = X_j$ by continuity, $\mathbb{P}(X_1 = \max(X_1, X_2, X_3)) = \frac{1}{3}$.
2. Suppose we sample without replacement $1 \leq k \leq n$ times from a pool of $n$ choices. Then, $(X_1, ..., X_n)$ are exchangeable. The proof for this is as much as writing down the joint PMF and seeing that it is symmetric.

---

One final, unsurprising, result about exchangeability is that, for $X_1, ..., X_n$ exchangeable and $g : \mathbb{R} \to \mathbb{R}$ with $g(X_i) = Y_i$, $Y_1, ..., Y_n$ are also exchangeable.

PROOF.

Fix $B \subseteq \mathbb{R}^n$ and define $g^{-1}(B) = \{(z_1, ..., z_n) \in \mathbb{R}^n : \langle g(z_1), ..., g(z_n)\rangle \in B\}$. Then we have $(Y_1, ..., Y_n) \in B \iff (X_1, ..., X_n) \in g^{-1}(B)$, so

$$\begin{aligned}\mathbb{P}[(Y_1, ..., Y_n) \in B] &= \mathbb{P}[(X_1, ..., X_n) \in g^{-1}(B)] = \mathbb{P}[(X_{\sigma(1)}, ..., X_{\sigma(n)}) \in g^{-1}(B)] \\ &= \mathbb{P}\Big[g\left(X_{\sigma(1)}\right), ..., g\left(X_{\sigma(n)}\right) \in B\Big] \\ &= \mathbb{P}[(Y_{\sigma(1)}, ..., Y_{\sigma(n)}) \in B] \implies Y_1, ..., Y_n \text{ exchangeable} \quad \square\end{aligned}$$

# VIII Multivariate Expectation and Variance

## EXPECTATION

Previously, we've seen that, if $X$ and $Y$ are random variables defined on a common space, with $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, then $\mathbb{E}[X + Y] = \mathbb{E}X + \mathbb{E}Y$. This generalizes for $n$ variables:

**8.1 Linearity of Expectation**
Let $X_1, ..., X_n : \Omega \to R$ such that $E|X_i| < \infty$. Then

$$\mathbb{E}[X_1 + ... + X_n] = \mathbb{E}X_1 + ... + \mathbb{E}X_n$$

PROOF.

By induction: we have $\mathbb{E}[X_1 + ... + X_{n+1}] = \mathbb{E}[X_1 + ... + X_n] + \mathbb{E}X_{n+1}$, so long as $|X_1 + ... + X_n| < \infty$, but this can be verified using the triangle inequality, i.e. $|X_1 + ... + X_n| \leq X_1 + ... + X_n < \infty$. Then $\mathbb{E}[X_1 + ... + X_n] + \mathbb{E}X_{n+1} = \sum_{i=1}^{n+1} \mathbb{E}X_i$, and we are done. □

An immediate corollary of linearity is that, for functions $g_i : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}[g_i(X_i)] < \infty$, we have $\mathbb{E}[g_1(X_i) + ... + g_n(X_n)] = \sum_{i=1}^{n} \mathbb{E}[g_i(X_i)]$.

### *Indicator Method*

Suppose $X$ is an positive, integer-valued discrete variable, such as the sum of coin tosses. We can always write $X = \sum_{i=1}^{n} \mathbb{1}_{E_i}$ for events $E_1, ..., E_n$.

By multivariate linearity, then, we can write $\mathbb{E}X = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}_{E_i}] = \sum_{i=1}^{n} \mathbb{P}(E_i)$. We call this the *indicator method* for calculating expectation.

---

**Example:** Consider a "head run" in a series of coin tosses, i.e. a finite sequence in our flips where one only sees heads. Let $X = \{\#$ of length-5 head runs in 100 tosses$\}$. We can write this the sum of indicators which look for length-5 head runs. If we notate each flip $f_i \sim \text{Ber}(1/2)$ for $i \in [100]$, one writes

$$X = \left( \sum_{i=2}^{95} \mathbb{1}_{\{f_{i-1}=0, f_i=1, ..., f_{i+4}=1, f_{i+5}=0\}} \right) + \mathbb{1}_{\{f_1=1, ..., f_5=1, f_6=0\}} + \mathbb{1}_{\{f_{95}=0, f_{96}=1, ..., f_{100}=1\}}$$

By exchangeability, this is $95\mathbb{P}(f_i = 1, ..., f_{i+5} = 1) + \frac{1}{2^6} + \frac{1}{2^6} = \frac{95}{2^7} + \frac{2}{2^6} = \frac{98}{2^7}$.

---

### *Expectation of Products*

There isn't a general form for $\mathbb{E}[XY]$ in absolute terms, but we can consider some special cases to get a gist:

1. Let $X \sim \mathcal{N}(0,1)$ and $R := \begin{cases} 1 & p = 1/2 \\ -1 & p = 1/2 \end{cases}$, with $X$ and $R$ independent from eachother. One intuitively concludes that $\mathbb{E}[XR]$ must be $1/2\mathbb{E}X - 1/2\mathbb{E}X = 0$, which is indeed $\mathbb{E}X\mathbb{E}R$.
2. If we let $R \sim \text{Ber}(1/2)$ and $X$ defined as above, this expectation becomes $1/2\mathbb{E}X$, which is $\mathbb{E}R\mathbb{E}X$, again.
3. Now for an example with actual justification: let $X = Y$, where the distributions of $X$ and $Y$ remain unknown. $\mathbb{E}[XY] = \mathbb{E}[X^2]$. Also, $\mathbb{E}X\mathbb{E}Y = (\mathbb{E}X)^2$, so $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ only when $\text{Var}(X) = 0$, i.e. $X$ is degenerate.

With the gist gotten, we'll make the following unsurprising claim:

**8.2 Expectation of Independent Products**
If $X$ and $Y$ are independent random variables such that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, then $\mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y$. In fact, for any sequence of (finite expectation) variables $X_1, ..., X_n$, we have

$$\mathbb{E}[X_1 \cdot ... \cdot X_n] = \prod_{i=1}^{n} \mathbb{E}X_i$$

PROOF. Let $X, Y$ have densities $f_X$, $f_Y$, respectively. Then

$$\mathbb{E}[XY] = \iint_{\mathbb{R}} xyf_{X,Y}\,dydx = \iint_{\mathbb{R}} xf_X yf_Y\,dydx = \int_{\mathbb{R}} xf_X \int_{\mathbb{R}} yf_Y = \mathbb{E}X\mathbb{E}Y$$

Similarly, if $X$ and $Y$ are discrete, then one writes

$$\mathbb{E}[XY] = \sum_{x\in S_X}\sum_{y\in S_Y} xy\rho_{X,Y} = \sum_{x\in S_X}\sum_{y\in S_Y} xy\rho_X\rho_Y = \sum_{x\in S_X} x\rho_X \sum_{y\in S_Y} y\rho_Y = \mathbb{E}X\mathbb{E}Y$$

One can also define $g_i : \mathbb{R} \to \mathbb{R}$, and, where all expectations of $g_i(X_i)$ are finite, and we'll have $\mathbb{E}[g_1(X_1) \cdot ... \cdot g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot ... \cdot \mathbb{E}[g_n(X_n)]$. The proof for this is a mostly trivial induction.

A similar statement holds for independent $\{X_i, i \in [n]\}$, where $\mathbb{E}[X_i^2] < \infty$, though a proof won't be provided:

**8.3 Variance of Independent Sums**
For $X_1, ..., X_n$ as described above, we have

$$\text{Var}(X_1 + ... + X_n) = \sum_{i=1}^{n} \text{Var}(X_i)$$

**Example:** Define $X \sim \text{NegBin}(k, p)$, the *negative binomial* distribution, which describes the time one waits until they see a sum $k$ of $\text{Bin}(p)$ events. $X$ is then the sum of times between each successful flip. If we denote these times $T_i$ for the $i^{\text{th}}$ success, then $X = T_1 + (T_2 - T_1) + (T_3 - T_2) + ... + (T_k - T_{k-1})$. Note that one can write $T_1 = T_1 - T_0$, where $T_0$ is the $0^{\text{th}}$ flip. Also note, since the flips are IID variables, they are exchangeable, so $T_i - T_{i-1}$ are all distributed as $T_1 - T_0 = T_1$, which is a $\text{Geom}(p)$ variable.

i.e. time until one sees $k$ heads, where probability on flips heads is $p$

Thus, $X = kG$, where $G \sim \text{Geom}(p)$. By linearity of expectation and our variance theorem above, we conclude

$$\mathbb{E}X = \mathbb{E}[kG] = k\mathbb{E}[G] = \frac{k}{p} \quad \text{and} \quad \text{Var}(X) = \text{Var}(kG) = k\text{Var}(G) = \frac{k(1-p)}{p^2}$$

*Moment Generating Function for Sums*

Recall from Part V the moment generating function of a variable $X$: $M_X(t) = \mathbb{E}[e^{tX}]$. We'll show the following using the facts we've established about expectation:

**8.4 MGF of Independent Sums**
If $X, Y$ are independent with MGFs $M_X(t)$ and $M_Y(t)$, respectively, then $M_{X+Y}(t) = M_X(t)M_Y(t)$, and, in fact

$$M_{X_1+...+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) \quad \text{with } X_i \text{ all independent}$$

PROOF.

We have that $M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}e^{tY}]$. With $g_1(X) = e^{tX}$ and $g_2(Y) = e^{tY}$, and using the expectation of independent products:

$$\mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t)$$

A generalization to $n$ elements can be shown in much the same spirit as previous inductions over expectation. □

SAMPLE MEAN AND VARIANCE

Given IID random variables $X_1, ..., X_n$, define the *sample mean*, denoted, $\overline{S_n}$, to be $\frac{1}{n}(X_1 + ... + X_n)$. For any given $X_i$, we have $\mathbb{E}[\overline{S_n}] = \mathbb{E}[X_i]$.

PROOF.

$$\mathbb{E}[\overline{S_n}] = \mathbb{E}\left[\tfrac{1}{n}(X_1 + ... + X_n)\right] = \tfrac{1}{n}\mathbb{E}[X_1 + ... + X_n] = \tfrac{1}{n}\mathbb{E}[nX_i] = \mathbb{E}[X_i]$$

Furthermore, $\text{Var}(S_n) = \frac{1}{n}\text{Var}(X_i)$ for any $X_i$. Recall that $\text{Var}(aX) = a^2\text{Var}(X)$.

PROOF.

$$\text{Var}(\overline{S_n}) = \text{Var}\left(\tfrac{1}{n}\sum_{i=1}^{n} X_i\right) = \tfrac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \tfrac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \tfrac{1}{n}\text{Var}(X_i) \quad \square$$

Now define the *sample variance* of IID $X_1, ..., X_n$ as $\overline{V_n} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{S_n})^2$. We have that $\mathbb{E}[\overline{V_n}] = \text{Var}(X_i)$. Denote $\mu = \mathbb{E}X_i$.

PROOF.

$$\begin{aligned}
\mathbb{E}\left[(X_i - \overline{S_n})^2\right] &= \mathbb{E}\left[(X_i - \mu + \mu - \overline{S_n})^2\right] \\
&= \underbrace{\mathbb{E}[(X_i - \mu)^2]}_{=\text{Var}(X_i)} + 2\mathbb{E}\left[(X_i - \mu)(\mu - \overline{S_n})\right] + \underbrace{\mathbb{E}[(\overline{S_n} - \mu)^2]}_{=\text{Var}(\overline{S_n}) = \frac{\text{Var}(X_i)}{n}} \\
&= \frac{n+1}{n}\text{Var}(X_i) - 2\mathbb{E}\left[(X_i - \mu)(\overline{S_n} - \mu)\right] \\
\implies \sum_{i=1}^{n}\mathbb{E}\left[(X_i - \overline{S_n})^2\right] &= (n+1)\text{Var}(X_i) - 2\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \mu)(\overline{S_n} - \mu)\right] \\
&= (n+1)\text{Var}(X_i) - 2\mathbb{E}\left[\sum_{i=1}^{n} n(\overline{S_n} - \mu)^2\right] \\
&= (n+1)\text{Var}(X_i) - 2n\text{Var}(\overline{S_n}) \\
&= (n+1)\text{Var}(X_i) - 2\text{Var}(X_i) = (n-1)\text{Var}(X_i) \\
\implies \overline{V_n} &= \text{Var}(X_i) \quad \square
\end{aligned}$$

Note that $\sum_{i=1}^{n} X_i - \mu = X_1 + ... + X_n - n\mu = n\overline{S_n} - n\mu = n(\overline{S_n} - \mu)$

## COVARIANCE AND CORRELATION

Let $X, Y$ be random variables defined on a common space. We define the *covariance* of $X$ and $Y$, $\text{Cov}(X, Y)$, to be $\mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right]$.

One can rearrange to yield an alternative formula, $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y$. We also have $\text{Cov}(X + a, Y + b) = \text{Cov}(X, Y)$ and $\text{Cov}(cX, dY) = cd\text{Cov}(X, Y)$, which thus means that covariance is not normalized. Both these results follow directly from the definition.

Now define the *correlation* of $X, Y$ to be

$$\text{Corr}(X, Y) = \text{Cov}\left(\frac{X - \mathbb{E}X}{\sqrt{\text{Var}X}}, \frac{Y - \mathbb{E}Y}{\sqrt{\text{Var}Y}}\right) = \text{Cov}(X, Y) \cdot \frac{1}{\sqrt{\text{Var}X\text{Var}Y}}$$

This quantity is simply a normalization of covariance, and takes values between -1 and 1 when expectations are finite and Var $\in (0, \infty)$ for both $X$ and $Y$. There is actually a good deal to prove in that statement, though.

One says that two variables are *positively*, *negatively*, or *un-* correlated if their covariance is positive, negative, or 0, respectively (this implies Corr is +/-/0 as well).

---

**Example:** We'll consider sampling without replacement from a bin of $n$ balls, $A$ of which are yellow. Define the variables $X = \mathbb{1}_{i^{\text{th}} \text{ sample yellow}}$ and $Y = \mathbb{1}_{j^{\text{th}} \text{ sample yellow}}$ By exchangeability, we can let $i = 1$ and $j = 2$ (this follows from the example on p. 35). Then $\mathbb{E}[XY] = \mathbb{P}(\text{first 2 samples yellow}) = \frac{A}{n}(\frac{A-1}{n-1})$. Similarly, $\mathbb{E}X = \frac{A}{n}$ and $\mathbb{E}Y = \frac{A}{n}$. We conclude $\text{Cov}(X, Y) = \frac{A(A-1)}{n(n-1)} - \frac{A^2}{n^2} < 0$, so $X$ and $Y$ are negatively correlated.

There are a few ways of interpreting this result: one sees that $X$ negatively influences the outcome of $Y$, i.e. the probability of $Y$ decreases as $X$ occurs, since $X$ "takes away" a sample that may lead to the success of $Y$. If, on $X$ occurring, $X$ "puts back" its yellow ball, this effect is negated, one expects correlation to be 0, and this is verified when conceptualizing our new setup as sampling with replacement (if $X$ and $Y$ are independent, they would have no correlation–this stems directly from the definition of covariance.)

Generally speaking, if [$X$ lying above its mean $\implies$ $Y$ lies below its], then they are negatively correlated. If [$X$ lying above its mean $\implies$ $Y$ *also* lies above it], then they are positively correlated (helping eachother out, per se).

---

# IX More Limit Theorems and Approximations

*Some Inequalities*

For a random variable $X$ and $t \in \mathbb{R} > 0$, we have

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}X}{t}$$

this is called *Markov's Inequality*. It's actually quite straightforward to prove:

PROOF.

$$\mathbb{E}|X| \geq \mathbb{E}\left[|X|\mathbb{1}_{|X|\geq t}\right] \geq \mathbb{E}\left[t\mathbb{1}_{|X|\geq t}\right] = t\mathbb{E}\left[\mathbb{1}_{|X|\geq t}\right] = t\mathbb{P}(|X| \geq t) \qquad \square$$

We also have *Chebyshev's Inequality*, which states for a random variable $X$, $t \in \mathbb{R} > 0$, and/or an independent sum $S_n = X_1 + ... + X_n$:

$$\mathbb{P}(|X - \mathbb{E}X| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2} \quad \text{and} \quad \mathbb{P}(|S - \mathbb{E}S| \geq t) \leq \sum_{k=1}^{n} \mathrm{Var}(X_k)\frac{1}{t^2}$$

## LAWS OF LARGE NUMBERS

For a sequence of random variables $(S_n, 1 \leq n \leq \infty)$, we say that $S_n$ *converges in probability* to $S_\infty$, and write $S_n \xrightarrow{\mathbb{P}} S_\infty$, if

$$\forall\ \varepsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}(|S_n - S_\infty| > \varepsilon) = 0$$

As an example, if we consider $S_n \sim \mathrm{Ber}(1/n)$, then $S_n \xrightarrow{\mathbb{P}} 0$. One sees that $\mathbb{P}(|S_n - 0| > \varepsilon) = \mathbb{P}(S_n = 1) \to 0$ as $n \to \infty$.

**9.1 Weak Law of Large Numbers**
Let $(X_n, n \geq 1)$ be a sequence of independent random variables, with $\mathbb{E}(X_n^2) < \infty$ always. Let $S_n := X_1 + ... + X_n$. Then

$$\frac{S_n - \mathbb{E}X_n}{n} \xrightarrow{\mathbb{P}} 0$$

PROOF.

By Chebyshev, we know that

$$\mathbb{P}(|S_n - \mathbb{E}X_n| \geq t) \leq \frac{\mathrm{Var}(S_n)}{t^2} = \sum_{n\geq 1} \mathrm{Var}(X_i)\frac{1}{t^2} \leq \frac{cn}{t^2}$$

where the last step follows from the fact that $\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}X)^2 \leq c - (\mathbb{E}X)^2 \leq c$. Since $\mathbb{E}[X_i^2]$ is finite, we choose $c \in \mathbb{R}$ arbitrarily.

We need to show $\mathbb{P}(|S_n - \mathbb{E}X_n| > \varepsilon n) \to 0$, but this is easy now, since $\mathbb{P}(|S_n - \mathbb{E}X_n| > \varepsilon n) \leq \frac{cn}{(\varepsilon n)^2} = \frac{c}{\varepsilon^2 n}$, which clearly goes to 0 for any fixed $\varepsilon, c$.

$$\mathbb{P}(|S_n - \mathbb{E}X_n| > \varepsilon n) \to 0 \implies \lim_{n\to\infty} \mathbb{P}\left(\frac{|S_n - \mathbb{E}X_n|}{n} - 0 > \varepsilon\right) = 0 \quad \square$$

Once again for random variables $(S_n, n \geq 1)$, we say that $S_n$ *converges almost surly* to $S_\infty$, and write $S_n \overset{\text{a.s.}}{\to} S_\infty$, if

$$\forall \varepsilon > 0 \; \exists N : \; \mathbb{P}\left(\forall n \geq N \; |S_n - S_\infty| < \varepsilon\right) = 1$$

or, equivalently, if $\mathbb{P}(\lim_{n\to\infty} X_n = X_\infty) = 1$.

---

**Example:** These symbols provide little intuition, so here is some food for thought: suppose we roll a die many, many times, and record a variable $S_i$, being the average of rolls thus far. After 1000 rolls, we may see a value $S_{1000} = 3.573$. We let $S_\infty = 3.5$ exactly, since this is the expectation of a roll. $S_n$ will get close to $S_\infty$ over time, so $S_n$ converges in probability to $S_\infty$.

Now suppose one rolls this same die many, many times, but stops after rolling 100 6's in a row. Here, our random variables are simply the number rolled. It's not so useful to think of these variables *approaching* any number, because they don't. However, we *do* know that, at some point (say, the $N^{\text{th}}$ roll), 100 6's will be rolled, and thus all $X_i$ for $i \geq N$ will be 0. In this sense, $X_i$ are *almost surely* going to 0. Note that, in any finite setting, there is always positive probability that 100 6's haven't been rolled yet.

It is a fact that a.s. convergence implies convergence in probability, but not the other way around, so it is the stronger condition.

Borel-Cantelli Lemma: if $X_n$ are independent, $\sum_{n\geq 1} \mathbb{P}(X_n) < \infty \implies \mathbb{P}(X_n \text{ occurs } \infty \text{ often}) = 0$. If $\sum_{n\geq 1} \mathbb{P}(X_n) = \infty$, then this probability is 1. This is unexaminable, but is an easier way of showing that $\text{Ber}(1/n) \overset{\text{a.s.}}{\not\to} 0$, and further that $\text{Ber}(1/n^2)$ *does* converge almost surly to 0

As a concrete example, consider the independent sequence $S_n \sim \text{Ber}(\frac{1}{n})$. We've shown above that this converges in probability to 0. Fix any $N$. We want to consider $\mathbb{P}(\forall n \geq N \; |S_n| < \varepsilon)$. This probability is less than $\mathbb{P}(\forall n \in [N, \tilde{N}] \; |S_n| < \varepsilon)$. By independence, this is

$$\prod_{n=N}^{\tilde{N}} \mathbb{P}(S_n < \varepsilon) = \prod_{n=N}^{\tilde{N}} \mathbb{P}(S_n = 0) = \prod_{n=N}^{\tilde{N}} (1 - \frac{1}{n}) = \prod_{n=N}^{\tilde{N}} (\frac{n-1}{n}) = \frac{N-1}{\tilde{N}}$$

This is clearly less than 1, and we are done.

---

### 9.2 Strong Law of Large Numbers

Let $(X_n, n \geq 1)$ be IID with $\mathbb{E}X_i = 0$. Let $S_n := X_1 + ... + X_n$. Then $\frac{S_n}{n} \overset{\text{a.s.}}{\to} 0$

The proof for this is significantly more involved than that of the weak law, and in doable fashion requires the assumption that $\mathbb{E}[X_i^4]$ is finite.

**9.3 Central Limit Theorem**

Let $-\infty < a < b < \infty$ and $X_1, ..., X_n$ be IID random variables with finite mean $\mu$ and variance $\sigma^2$. Let $S_n = X_1 + ... + X_n$. Then we have

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \in [a, b]\right) \to \int_a^b \varphi(x) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = \Phi(b) - \Phi(a)$$

# X Conditional Distributions

## DISCRETE SETTING

Given a discrete random variable $X$ and an event $E$ with $\mathbb{P}(E) > 0$, we define the *conditional probability mass function*

$$\rho_{X|E}(x) = \mathbb{P}(X = x|E) = \frac{\mathbb{P}(X = x, E)}{\mathbb{P}(E)}$$

This is indeed a PMF, since

$$\sum_{x \in S_X} \rho_{X|E} = \sum_{x \in S_x} \frac{\mathbb{P}(X = x, E)}{\mathbb{P}(E)} = \frac{1}{\mathbb{P}(E)} \mathbb{P}(x \in S_X, E) = \frac{\mathbb{P}(E)}{\mathbb{P}(E)} = 1$$

We similarly define the *conditional expectation* of $X$, given $E$, as $\mathbb{E}[X|E] = \sum_{x \in S_X} x \rho_{X|E}(x) = \sum_{x \in S_X} x \mathbb{P}(X = x \mid E)$.

Generally speaking, if $E_1, ..., E_n$ partition $\Omega$, then we have the following: $\rho_X(x) = \sum_{i=1}^{n} \rho_{X|E_i}(x) \mathbb{P}(E_i)$, and thus $\mathbb{E}X = \sum_{x \in S_X} x \rho_X = \sum_{x \in S_X} x \sum_{i=1}^{n} \rho_{X|E_i}(x) \mathbb{P}(E_i)$.

Similarly, $\mathbb{E}X = \mathbb{E}[X|E]\mathbb{P}(E) + \mathbb{E}[X|E^c]\mathbb{P}(E^c)$

---

**Examples:**

1. Suppose we model the number of customers who are in a store at a given moment, and call this variable $X$. If it's raining outside, the event $R$ occurs, and if not, $R^c$ does. The distribution of $X$ is $\text{Poi}(\lambda)$ when it's raining, and $\text{Poi}(\mu)$ when not.

   $\mathbb{E}X = \mathbb{E}[X|R]\mathbb{P}(R) + \mathbb{E}[X|R^c]\mathbb{P}(R^c) = \lambda \mathbb{P}(R) + \mu(1 - \mathbb{P}(R))$.

2. Let $(X_i, i \geq 1)$ be IID $\text{Ber}(p)$ variables, and $N = \min(i : X_i = 1)$. Then $N \sim \text{Geom}(p)$. It's expectation we know to be $\frac{1}{p}$, but we'll derive it with conditional expectations:

   $\mathbb{E}N = \mathbb{E}[N|X_1 = 0](1 - p) + \mathbb{E}[N|X_1 = 1]p = \mathbb{E}[N + 1](1 - p) + p = (1 - p)(\mathbb{E}N + 1) + p \implies \mathbb{E}N = 1/p$

If $X, Y$ are discrete, then for $x \in S_X$ and $y \in S_Y$, we write $\rho_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y)$. If $X, Y$ are independent, then $\rho_{X|Y}(x|y) = \rho_X(x)$.

3. With that in mind, consider $X \sim \text{Poi}(\lambda)$, $Y \sim \text{Poi}(\mu)$, independent poissons, and $Z = X + Y$. One can show that $Z \sim \text{Poi}(\lambda + \mu)$ (this is a good exercise for material from much earlier). We are interested in $\rho_{X|Z}(k|l) = \mathbb{P}(X = k|Z = l)$:

$$\mathbb{P}(X = k|Z = l) = \frac{\mathbb{P}(X = k, Z = l)}{\mathbb{P}(Z = l)} = \frac{\mathbb{P}(X = k, X + Y = l)}{\mathbb{P}(Z = l)}$$

$$= \frac{\mathbb{P}(X = k, Y = l - k)}{\mathbb{P}(Z = l)} = \frac{\lambda^k e^{-\lambda}}{k!} \frac{\mu^{l-k} e^{-\mu}}{(l-k)!} \left( \frac{(\lambda + \mu)^l e^{-\lambda-\mu}}{l!} \right)^{-1}$$

$$= \frac{\lambda^k \mu^{l-k} l!}{k!(l-k)!(\lambda+\mu)^l} = \binom{l}{k} \left(\frac{\lambda}{\mu}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^l = \binom{l}{k} \left(\frac{\lambda}{\mu+\lambda}\right)^k \left(\frac{\mu}{\lambda+\mu}\right)^{l-k}$$

$$\implies \rho_{X|Z}(k|l) \sim \text{Bin}\left(l, \frac{\lambda}{\lambda+\mu}\right)$$

---

*Poisson Marking*

Let $P \sim \text{Poi}(\lambda)$ represent the number of customers who enter a store in some fixed time interval. Suppose each person receives, independently, a coupon of type 1, 2, or 3 with probability $p_1$, $p_2$, $p_3$, respectively. What is the joint density of $X_1, X_2, X_3$, the number of people who receive coupons of type $i$ in the time interval?

If we fix $P = k$, this question becomes much easier, and we conclude that $\mathbb{P}(X_1 = k_1, X_2 = k_2, X_3 = k_3 \mid P = k) = \binom{k}{k_1,k_2,k_3} p_1^{k_1} p_2^{k_2} p_3^{k_3}$

$$\rho(k_1, k_2, k_3) = \mathbb{P}(X_1 = k_1, X_2 = k_2, X_3 = k_3 \mid P = k)\mathbb{P}(P = k)$$

$$= \binom{k}{k_1, k_2, k_3} p_1^{k_1} p_2^{k_2} p_3^{k_3} \frac{\lambda^k e^{-\lambda}}{k!} = \frac{(\lambda p_1)^{k_1} e^{-\lambda p_1}}{k_1!} \frac{(\lambda p_2)^{k_2} e^{-\lambda p_2}}{k_2!} \frac{(\lambda p_3)^{k_3} e^{-\lambda p_3}}{k_3!}$$

When rearranging for the last step, one notes that $k_1 + k_2 + k_3 = k$ and $p_1 + p_2 + p_3 = 1$. This form is significant, since it splits $\rho$ into 3 separate PMFs, $\text{Poi}(\lambda p_1)$, $\text{Poi}(\lambda p_2)$, and $\text{Poi}(\lambda p_3)$, so we can conclude that the number of people holding type 1, 2, and 3 coupons, respectively, are given independent of one another.

We can generalize this process of "marking" elements of a Poisson process with types (in the example above there are 3). In an unmarked setting, $|P \cap I| \sim \text{Poi}(\lambda|I|)$ as we have seen before. Suppose now we mark the elements of the process with types $1, 2, ..., n$, doing so with probabilities $p_1, p_2, ..., p_n$ such that $p_1 + ... + p_n = 1$, and denote $X_i \subseteq P$ to be the set of type $i$ elements. We have the following for disjoint intervals $I_1, ..., I_k$:

$$|X_1 \cap I_1| \sim \text{Poi}(\lambda p_1|I_1|), \quad |X_2 \cap I_1| \sim \text{Poi}(\lambda p_2|I_1|), ..., \ |X_n \cap I_1| \sim \text{Poi}(\lambda p_n|I_1|)$$

$$|X_1 \cap I_2| \sim \text{Poi}(\lambda p_1|I_2|), \quad |X_2 \cap I_1| \sim \text{Poi}(\lambda p_2|I_2|), ..., \ |X_n \cap I_2| \sim \text{Poi}(\lambda p_n|I_2|)$$

$$\vdots \qquad\qquad \vdots$$

$$|X_1 \cap I_k| \sim \text{Poi}(\lambda p_1|I_k|), \quad |X_2 \cap I_k| \sim \text{Poi}(\lambda p_2|I_k|), ..., \ |X_n \cap I_k| \sim \text{Poi}(\lambda p_n|I_k|)$$

### CONTINUOUS SETTING

Let $X, Y$ have joint density $f_{X,Y}$. Just as we did for discrete variables, define

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{provided that } f_Y(y) > 0$$

Even though the probability that $Y$ is some fixed value is 0 in continuous settings, we can still ask $\mathbb{P}(X \in A|Y = y)$, and this is given by

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx = \frac{1}{f_Y(y)} \int_A f_{X,Y}(x,y)$$

One can verify that, as expected, $\int_{\mathbb{R}} f_{X|Y}(x|y) = 1$, since this is just

$$\frac{1}{f_Y(y)} \int_{\mathbb{R}} f_{X,Y}(x,y)dx = \frac{f_Y(y)}{f_Y(y)} = 1$$

Recall that we can recover the density of $f_X$ by integrating over the joint density: $f_X = \int_{\mathbb{R}} f_{X,Y}(x,y)dy$. From above, this is $\int_{\mathbb{R}} f_{X|Y}(x|y)f_Y(y)dy$.

That's it! Cheers