Probability

Jonathan Gai

4th April 2022

Contents

1 Formal Setup **2**
1.1 Examples of Probability Spaces 4
1.2 Combinatorial Analysis 5
1.3 Random Walks 6
1.4 Properties of Probability Measures 8
1.5 Bonferroni Inequalities 11
1.6 Counting with IEP 11
1.7 Independence 14
1.8 Conditional Probability 15
1.9 Law of Total Probability & Bayes' Formula 17

2 Discrete Random Variables **20**
2.1 Discrete Probability Distributions 22
2.2 More Than One RV 24
2.3 Expectation 25

Lecture 1: Probability Space

20 Jan. 11:00

Example. If we have a die with outcomes $1, 2, \ldots, 6$.

1. $\mathbb{P}(2) = \frac{1}{6}$
2. $\mathbb{P}(\text{multiple of } 3) = \frac{2}{6} = \frac{1}{3}$
3. $\mathbb{P}(\text{pair or a multiple of } 3) = \frac{4}{6} = \frac{2}{3}$

1 Formal Setup

We try to define a probability space rigorously in this section.

Definition 1.1: Probability Space

We have the following,

1. Sample space Ω, a set of outcomes.
2. $\mathcal{F}$, a collection of subsets of Ω (called events).
3. $\mathcal{F}$ is a σ-algebra if
   a) **F1**: $\Omega \in \mathcal{F}$
   b) **F2**: if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
   c) **F3**: For all countable collections $\{A_n\}$ in $\mathcal{F}$, $\cup_n A_n \in \mathcal{F}$.

Given σ-algebra $\mathcal{F}$ on Ω, function $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure if

1. **P1**: The probability function is nonnegative.
2. **P2**: $\mathbb{P}(\Omega) = 1$
3. **P3**: For all countable collection $\{A_n\}$ of disjoint events in $\mathcal{F}$, we have $\mathbb{P}(\cup_n A_n) = \sum\limits_{n=1}^{\infty} \mathbb{P}(A_n)$.

Then $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.

Problem. Why $\mathbb{P} : \mathcal{F} \to [0, 1]$, not $\mathbb{P} : \Omega \to [0, 1]$?

We will justify the definition in the following examples.

Example. When Ω is finite or countable,

1. In general: $\mathcal{F} = \mathcal{P}(\Omega)$.

2. $\mathbb{P}(2)$ is shorthand for $\mathbb{P}(\{2\})$.

3. $\mathbb{P}$ is determined by $\mathbb{P}(\{w\}), \forall w \in \Omega$.

Remark. When Ω is uncountable, a probability space behaves differently, as shown in the following example.

Example. If $\Omega = [0,1]$, and we want to choose a real number, all equally likely.

If $\mathbb{P}\{0\} = \alpha > 0$, then $\mathbb{P}(\{0, 1, \frac{1}{2}, \ldots, \frac{1}{n}\} = n\alpha)$. This cannot happen if n large, because we would have $\mathbb{P} > 1$. So $\mathbb{P}(\{0\}) = 0$ or undefined.

Example. When Ω is infinitely countable (e.g., $\Omega = \mathbb{N}$ or $\Omega = \mathbb{Q} \cap [0,1]$), however, it is not possible to choose uniformly. Suppose it is possible, there are two possibilities

- If $\mathbb{P}\left(\{\omega\}\right) = \alpha \quad \forall \omega \in \Omega$,

  then $\mathbb{P}\left(\Omega\right) = \sum\limits_{\omega \in \Omega} \mathbb{P}\left(\{\omega\}\right) = \infty$. $\lightning$

- If $\mathbb{P}\left(\{\omega\}\right) = 0 \quad \forall \omega \in \Omega$,

  then $\mathbb{P}\left(\Omega\right) = \sum\limits_{\omega \in \Omega} \mathbb{P}\left(\{\omega\}\right) = 0$. $\lightning$

So it is not possible to have one such uniform probability space. But that's fine as there exists many other interesting probability measures on a infinite countably set.

Property. From the axioms, we want to prove the following properties of a probability space.

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

   Proof. A, A^c disjoint. $A \cup A^c = \Omega$. So $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$ ∎

2. $\mathbb{P}(\varnothing) = 0$

3. If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

1.1 Examples of Probability Spaces

Example. Here we list some concrete examples of probability spaces.

1. Ω finite, $\Omega = \{w_1, \ldots, w_n\}$, $\mathcal{F}$ = all subsets under uniform choice.

   $\mathbb{P} : \mathcal{F} \to [0,1], \mathbb{P}(A) = \frac{|A|}{|\Omega|}$. In particular: $\mathbb{P}(\{w\}) = \frac{1}{|\Omega|} \forall w \in \Omega$.

2. If we are choosing without replacement n indistinguishable marbles that are labelled $\{1, \ldots, n\}$. Pick $k \leq n$ marbles uniformly at random.

   Here we have $\Omega = \{A \subseteq \{1, \ldots, n\}$, $|A| = k$, $|\Omega| = \binom{n}{k}$.

3. If we have a well-shuffled deck of cards, and we uniformly chose permutation of 52 cards.

   $\Omega = \{\text{all permutations of 52 cards}\}$. $|\Omega| = 52!$.

   Then we have

   $$\mathbb{P}(\text{first three cards have the same suit}) = \frac{52 \cdot 12 \cdot 11 \cdot 49!}{52!} = \frac{22}{425}.$$

Lecture 2: Finite Probability Space

22 Jan. 11:00

Example (Coincidental Birthday)**.** There we have n people, what is the probability that at least two share a birthday? To be precise, we first make the following assumptions,

- No leap years; (365 days in a year)
- All birthdays are equally likely.

We have the probability space

$$\begin{aligned}
\Omega &= \{1, \ldots, 365\}^n \\
\mathcal{F} &= \mathcal{P}(\Omega) \\
A &= \{\text{at least 2 people share birthday}\} \\
A^c &= \{\text{all } n \text{ birthdays are different}\}.
\end{aligned}$$

So we have the probability

$$\mathbb{P}\left(A^c\right) = \frac{365 \times 364 \times \ldots \times (365 - n - 1)}{365^n},$$
$$\mathbb{P}\left(A\right) = 1 - \frac{365 \times 364 \times \ldots \times (365 - n - 1)}{365^n}.$$

Remark.

- We note several special n values,

$$\begin{aligned} n = 22 \quad &: \quad \mathbb{P}\left(A\right) \approx 0.479 \\ n = 23 \quad &: \quad \mathbb{P}\left(A\right) \approx 0.507 \\ n \geq 366 \quad &: \quad \mathbb{P}\left(A\right) = 1 \end{aligned}$$

- The probability of birthday is not equal in real life though. It is more likely to be born about 9 months after christmas.

- Sometimes it would be easier to calculate the probability of the complement of an event.

1.2 Combinatorial Analysis

If Ω is a finite set such that $|\Omega| = n$,

Problem. How many ways to partition Ω into k disjoint subsets $\Omega_1, \ldots \Omega_k$ with $|\Omega_i| = n_i$ ($\sum\limits_{i=1}^{k} n_i = n$)?

The total number of ways M is

$$
\begin{aligned}
M &= \binom{n}{n_i}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-n_2\cdots-n_{k-1}}{n_k} \\
&= \binom{n}{n_i}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_k}{n_k} \\
&= \frac{n!}{n!(n-n_1)!}\times\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\times\cdots\times\frac{(n-n_1-n_2-\cdots-n_{k-1})!}{x_k!0!} \\
&= \frac{n!}{n_1!n_2!\cdots n_k!} \\
&= \binom{n}{n_1,n_2,\ldots,n_k}
\end{aligned}
$$

which is called the *multinomial coefficient*, and denoted by the last term in the equations.

Remark. The ordering of the subsets do matter in this setting.

1.3 Random Walks

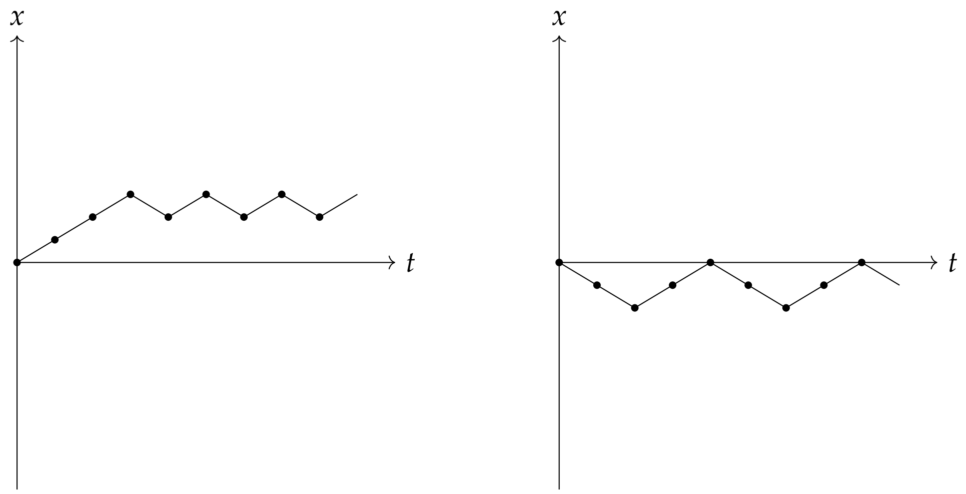

Figure 1: Random Walks

We have the following uniform probability space

$$
\begin{aligned}
&\Omega = \{(x_0,x_1,\ldots,x_n) \mid x_0 = 0, |x_k - x_{k-1}| = 1, k = 1,\ldots,n\}, \\
&|\Omega| = 2^n.
\end{aligned}
$$

Problem. What's $\mathbb{P}\left(x_n = 0\right)$ and $\mathbb{P}\left(x_n = n\right)$?

We have $\mathbb{P}(x_n = n) = \frac{1}{2^n}$.

When n is odd, $\mathbb{P}(x_n = 0) = 0$ because after every step the value changes parity. To find the probability when n is even, we need to choose $\frac{n}{2}$ ks for which $x_k = x_{k-1} + 1$, and the rest $x_k = x_{k-1} - 1$. So

$$\begin{aligned}\mathbb{P}(x_n = 0) &= 2^{-n}\binom{n}{n/2} \\ &= \frac{n!}{2^n[(\frac{n}{2})!]^2}.\end{aligned}$$

Problem. What happens when n is large?

We next present Stirling's Formula, and we adopt the following notation for the time being.

Notation. If (a_n), b_n are two sequences, we say $a_n \sim b_n$ as $n \to \infty$ if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$.

Theorem 1.1: Stirling's Formula

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \quad \text{as} \quad n \to \infty.$$

We also have the weaker version

$$\log(n!) \sim n \log n.$$

Lecture 3

25 Jan. 11:00

Proof. We have

$$\log(n!) = \log 2 + \log 3 + \ldots + \log n.$$

So

$$\begin{aligned}\int_1^n \log x dx &\le \log(n!) \le int_1^{n+1} \log x dx \\ \underbrace{n \log n - n + 1}_{n \log n} &\le \log(n!) \le \underbrace{(n+1)\log(n+1) - n}_{n \log n}.\end{aligned}$$

$\log(n!)$ is sandwiched between the lower and upper integrals, so $\log(n!)$ must be approximately $n \log n$ as well. In this calculation, these facts helped

1. $\log x$ is increasing, so it's easier to bounded by the integrals.

2. $\log x$ has a nice integral. So the integrals have closed forms.

■

(Ordered) Compositions

Definition 1.2

A *composition* of m with k parts is sequence $(m1, \ldots, m_k)$ of non-negative integers with $\sum_{i=1}^{k} m_i = m$.

We use stars and bars. There are m stars and $k-1$ bars, and

$$\#\text{Compositions} = \binom{m+k-1}{m}.$$

1.4 Properties of Probability Measures

Recall Definition 1.1. We prove the following properties.

Property.

1. Countable sub-additivity

   Let $(A_n)_{n\geq 1}$ sequence of events in $\mathcal{F}$. Then

   $$\mathbb{P}\left(\cup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}\left(A_n\right).$$

   Proof. We rewrite $\cup_{n\geq 1}$ as a disjoint union.

   Define $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup \ldots \cup A_{n-1})$.

   So

   - $\cup_{n\geq 1} B_n = \cup_{n\geq 1} A_n$,

- $(B_n)_{n\geq 1}$ disjoint (by construction),
- $B_n \subseteq A_n \implies \mathbb{P}(B_n) \leq \mathbb{P}(A_n)$.

And we have

$$\mathbb{P}(\cup_{n\geq 1} A_n) = \mathbb{P}(\cup_{n\geq 1} B_n) = \sum_{n\geq 1} \mathbb{P}(B_n) = \sum_{n\geq 1} \mathbb{P}(A_n).$$

■

2. Continuity $(A_n)_{n\geq 1}$ increasing sequence of events in $\mathcal{F}$ that is $A_n \subseteq A_{n+1}$ for all n.

   In fact, $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\cup_{n\geq 1} A_n)$.

   Proof. We reuse the B_ns, and we have

   - $\sqcup_{k=1}^{n} B_k = A_n$,
   - $\cup_{n\geq 1} B_n = \cup_{n\geq 1} A_n$.

   So we have

$$\mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(B_k) \to \sum_{k\geq 1} \mathbb{P}(B_k) = \mathbb{P}(\cup_{n\geq 1} B_n) = \mathbb{P}(\cup_{n\geq 1} A_n).$$

■

3. Inclusion-Exclusion Principle

   Background: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

   Similarly, for $A, B, C \in \mathcal{F}$,

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) =& \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(B \cap C) \\ &- \mathbb{P}(C \cap A) + \mathbb{P}(A \cap B \cap C).\end{aligned}$$

   The full Inclusion-Exclusion principle statement is the following. Let $A_1, \ldots, A_n \in$

$\mathcal{F}$, then

$$\begin{aligned}\mathbb{P}\left(\cup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \mathbb{P}\left(A_i\right) - \sum_{1\le i_1<i_2\le n} \mathbb{P}\left(A_{i_1}\cap A_{i_2}\right) + \dots \\ &\quad + (-1)^{n+1}\mathbb{P}\left(A_1\cap\ldots\cap A_n\right) \\ &= \sum_{\substack{I\subseteq\{1,\ldots,n\}\\ I\neq\varnothing}} (-1)^{|I|+1}\mathbb{P}\left(\cap_{i\in I}A_i\right).\end{aligned}$$

Lecture 3: Inclusion-Exclusion Principle

27 Jan. 2022

Proof. We used induction. The $n=2$ case is proved in the example sheet.

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \mathbb{P}\left((\bigcup_{i=1}^{n-1} A_i)\bigcup A_n\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{n-1} A_i\right) + \mathbb{P}\left(A_n\right) - \mathbb{P}\left((\bigcup_{i=1}^{n-1} A_i)\bigcap A_n\right).\end{aligned}$$

Note that for $J\subseteq\{1,\ldots,n-1\}$,

$$\bigcap_{i\in J}(A_i\cap A_n) = \bigcap_{i\in J\cup\{n\}} A_i.$$

The inductive statement tells us

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{\substack{J\subseteq\{1,\ldots,n-1\}\\ J\neq\varnothing}} (-1)^{|J|+1}\mathbb{P}(\bigcap_{i\in J} A_i) + \mathbb{P}\left(A_n\right) \\ &\quad - \sum_{\substack{J\subseteq\{1,\ldots,n-1\}\\ J\neq\varnothing}} (-1)^{|J|+1}\mathbb{P}\left(\bigcap_{i\in J\cup\{n\}} A_i\right) \\ &= \sum_{\substack{I\subseteq\{1,\ldots,n-1\}\\ I\neq\varnothing}} (-1)^{|I|+1}\mathbb{P}(\bigcap_{i\in I} A_i) + \mathbb{P}\left(A_n\right) \\ &\quad + \sum_{\substack{I\subseteq\{1,\ldots,n-1\}\\ n\in I, |I|\geq 2}} (-1)^{|I|+1}\mathbb{P}\left(\bigcap_{i\in I} A_i\right) \\ &= \sum_{\substack{I\subseteq\{1,\ldots,n\}\\ I\neq\varnothing}} (-1)^{|I|+1}\mathbb{P}\left(\bigcap_{i\in I} A_i\right).\end{aligned}$$

■

1.5 Bonferroni Inequalities

Problem. What if you truncate Inclusion-Exclusion Principle?

Recall countable subadditivity states that $\mathbb{P}\left(\cup A_i\right) \leq \sum \mathbb{P}\left(A_i\right)$, also known as union bound. We have the following inequalities.

- $\mathbb{P}\left(\cup_{i=1}^n A_i\right) \leq \sum_{k=1}^{r} (-1)^{k+1} \sum_{i_1<\ldots<i_k} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right)$ when r is odd;
- $\mathbb{P}\left(\cup_{i=1}^n A_i\right) \geq \sum_{k=1}^{r} (-1)^{k+1} \sum_{i_1<\ldots<i_k} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right)$ when r is even.

Problem. When is it good to truncate at, for example, $r = 2$?

Proof. We induct on r and n. When r is odd

$$
\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) =& \mathbb{P}\left(\bigcup_{i=1} A_i\right) + \mathbb{P}\left(A_n\right) - \mathbb{P}\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right) \\
\leq & \sum_{\substack{J\subseteq\{1,\ldots,n-1\} \\ 1\leq|J|\leq r}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i\in J} A_i\right) + \mathbb{P}\left(A_n\right) \\
& - \sum_{\substack{J\subseteq\{1,\ldots,n-1\} \\ 1\leq|J|\leq r-1}} (-1)^{|J|+1} \mathbb{P}\left(\bigcap_{i\in J\cup\{n\}} A_i\right) \\
\leq & \sum_{\substack{I\subseteq\{1,\ldots,n\} \\ 1\leq|I|\leq r}} (-1)^{|I|+1} \mathbb{P}\left(\bigcap_{i\in I} A_i\right).
\end{aligned}
$$

And a similar argument follows when r is even. ∎

1.6 Counting with IEP

Inclusion Exclusion Principle gives up a route to solve questions that do not have a closed form answer.

When we have a uniform probability measure on Ω with $|\Omega| < \infty$,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \ \forall A \subseteq \Omega.$$

Then $\forall A_1, \ldots, A_n \subseteq \Omega$,

$$|A_1 \cup \ldots \cup A_n| = \sum_{k=1}^{n} (-1)^{n+1} \sum_{i_1 < \ldots < i_k} \left| A_{i_1} \cap \ldots \cap A_{i_k} \right|,$$

and similarly for Bonferroni inequalities.

Example. We count the number of surjections $f : \{1, \ldots, n\} \to \{1, \ldots, m\}$ with $n \geq m$.

We have the probability space and event

$$\begin{aligned} \Omega &= \{f : \{1, \ldots, n\} \to \{1, \ldots, m\}\}, \\ A &= \{f : \mathrm{im}(f) = \{1, \ldots, m\}\}. \end{aligned}$$

For all $i \in \{1, \ldots, m\}$, let $B_i = \{f \in \Omega \mid i \notin \mathrm{im}(f)\}$. We have the following key observations:

- $A = B_1^c \cap \ldots B_m^c = (B_1 \cup \ldots \cup B_m)^c$.
- $\left| B_{i_1} \cap \ldots \cap B_{i_k} \right|$ is nice to calculate, and we have
  $$\left| B_{i_1} \cap \ldots \cap B_{i_k} \right| = |\{f \in \Omega \mid i_1, \ldots, i_k \notin \mathrm{im}(f)\}| = (m-k)^n.$$
  So by IEP, we have
  $$\begin{aligned} |B_1 \cup \ldots \cup B_m| &= \sum_{k=1}^{m} (-1)^{k+1} \sum_{i_1 < \ldots < i_k} \left| B_{i_1} \cap \ldots \cap B_{i_k} \right| \\ &= \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n. \end{aligned}$$
  So $|A| = m^n - \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n = \sum_{k=0}^{m} (-1)^k \binom{m}{k} (m-k)^n$.

Lecture 5: Independence

29 Jan. 2022

Example (Derangements)**.** We try to find the number of permutations with no fixed points, for a Secret Santa for example. We have the sample space and event

$$\begin{aligned} \Omega &= \{\text{permutations of } \{1, \ldots, n\}\}, \\ D &= \{\sigma \in \Omega \mid \sigma(i) \neq i \ \forall i = 1, \ldots, n\}. \end{aligned}$$

For all $i \in 1, \ldots, n$, let $A_i = \{\sigma \in \Omega \mid \sigma(i) = i\}$.

Problem. Is $\mathbb{P}(D)$ large or small when $n \to \infty$.

Similar to the last example, $D = A_1^c \cap \ldots \cap A_n^c = (\cup_{i=1}^n A_i)^c$, and

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \frac{(n-k)!}{n!}.$$

So by IEP, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) &= \sum_{k=1}^n (-1)^{k+1} \sum_{i_1 < \ldots < i_k} \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) \\ &= \sum_{k=1}^n (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!}.\end{aligned}$$

So $\mathbb{P}(D) = 1 - \mathbb{P}\left(\cup_{i=1}^n A_i\right) = 1 - \sum\limits_{k=1}^n \frac{(-1)^{k+1}}{k!} = \sum\limits_{k=0}^n \frac{(-1)^k}{k!}$.

In fact, when $n \to \infty$, $\mathbb{P}(D) \to \sum\limits_{k=0}^\infty \frac{(-1)^k}{k!} = e^{-1} \approx 0.37$.

Note. What if instead $\Omega' = \{\text{all functions } f : \{1, \ldots, n\} \to \{1, \ldots, n\}\}$?

We have $D = \{f \in \Omega' \mid f(i) \neq i \ \forall i = 1, \ldots, n\}$, and

$$\mathbb{P}(D) = \frac{(n-1)^n}{n^n} = (1 - \frac{1}{n})^n \to e^{-1}.$$

Can we just say $\mathbb{P}(D) = (\frac{n-1}{n})^n$? We would need independence to say that.

Also note that $f(i)$ is a random quantity associated to Ω. We will study these later as a random variable.

We are allowed to toss a fair coin n times, but we can't toss an unfair coin n times so far.

1.7 Independence

Definition 1.3

Events $A, B \in \mathcal{F}$ are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)\,. \text{ (denoted as } A \perp\!\!\!\perp B\text{)}$$

A countable collection of events (A_n) is *independent* if for all distinct $i_1, \ldots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right).$$

Remark. *Pairwise independence* does not imply independence.

Example. If we have the uniform probability space

$$\Omega = \{(H,H), (H,T), (T,H), (T,T)\},$$

and $\mathbb{P}(\{\omega\}) = \frac{1}{4}$ for all $\omega \in \Omega$. And we define the following events

$$\begin{aligned} A &= \text{first coin } H = \{(H,H), (H,T)\} \\ B &= \text{second coin } H = \{(H,H), (T,H)\} \\ C &= \text{same outcome} = \{(H,H), (T,T)\} \end{aligned}$$

Note that probability of each of these happening is $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$, and $A \cap B = A \cap C = B \cap C = \{(H,H)\}$, so they are pairwise independent. But

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)\,.$$

The three events are not independent.

Example.

- If we have $\Omega' = \{\text{all functions } f : \{1, \ldots, n\} \to \{1, \ldots, n\}\}$, and let $A_i = \{f \in \Omega' \mid f(i) = i\}$. Then,

$$\begin{aligned} \mathbb{P}(A_i) &= \frac{n^(n-1)}{n^n} = \frac{1}{n} \\ \mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) &= \frac{n^{n-k}}{n^n} = \frac{1}{n^k} = \prod_{j=1}^{k} \mathbb{P}\left(A_{i_j}\right). \end{aligned}$$

Here, (A_i) are independent events.

- If we have $\Omega = \{\sigma \mid \text{permutation of } \{1, \dots, n\}\ \}$, and let $A_i = \{\sigma \in \Omega \mid \sigma(i) = i\}$. Then,

$$\mathbb{P}(A_i) = \frac{n^(n-1)}{n^n} = \frac{1}{n}$$
$$\mathbb{P}\left(A_i \cap A_j\right) = \frac{(n-1)!}{n!} = \frac{1}{n(n-1)} \neq \mathbb{P}(A_i)\,\mathbb{P}\left(A_j\right).$$

Here, (A_i) are not independent.

Property.

1. If A is independent of B then A is also independent of B^c.

   Proof.
   $$\begin{aligned}\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) \\ &= \mathbb{P}(A)\,(1 - \mathbb{P}(B)) \\ &= \mathbb{P}(A)\,\mathbb{P}(B^c).\end{aligned}$$
   ∎

2. A is independent of $B = \Omega$ and of $C = \varnothing$.

   Proof. $\mathbb{P}(A \cap \Omega) = \mathbb{P}(A) = \mathbb{P}(A)\,\mathbb{P}(\Omega)$, and $A \perp\!\!\!\perp \varnothing$ by part 1. ∎

3. $\mathbb{P}(B) = 0$ or 1 Then A is independent of B.

1.8 Conditional Probability

Definition 1.4: Conditional Probability

If we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ as before. Consider $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, and we have $\mathbb{P}(A)$, The *conditional probability of A given B* is

$$\mathbb{P}(A \mid B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We can interpret this informally as the probability of A if we know B happened.

Example. If A, B are independent events,

$$\mathbb{P}\left(A \mid B\right) = \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}\left(B\right)} = \frac{\mathbb{P}\left(A\right)\mathbb{P}\left(B\right)}{\mathbb{P}\left(B\right)} = \mathbb{P}\left(A\right).$$

Informally, we know that if A, B are independent, then knowing where B happened doesn't affect probability of A.

Lecture 6

1 Feb. 2022

Property.

1. $\mathbb{P}\left(A \mid B\right) \geq 0$.

2. $\mathbb{P}\left(B \mid B\right) = \mathbb{P}\left(\Omega \mid B\right) = 1$.

3. (A_n) disjoint events in $\mathcal{F}$, we claim

$$\mathbb{P}\left(\cup_{n\geq 1} A_n \mid B\right) = \sum_{n\geq 1} \mathbb{P}\left(A_n \mid B\right).$$

Proof.
$$\begin{aligned}\mathbb{P}(\cup_{n\geq 1} A_n \mid B) &= \frac{\mathbb{P}\left((\cup_n A_n) \cap B\right)}{\mathbb{P}\left(B\right)} \\ &= \frac{\mathbb{P}\left(\cup_n (A_n \cap B)\right)}{\mathbb{P}\left(B\right)} \quad \text{numerator is a disjoint union} \\ &= \frac{\sum_n \mathbb{P}\left(A_n \cap B\right)}{\mathbb{P}\left(B\right)} = \sum_{n\geq 1} \mathbb{P}\left(A_n \mid B\right).\end{aligned}$$

To prove it, we used the definition, and applied **P1**, **P2**, **P3** to numerator. ■

4. $\mathbb{P}\left(\cdot \mid B\right)$ is a function from $\mathcal{F} \to [0, 1]$ that satisfies the rules to be a probability measure in Ω. It is often useful to restrict the function to

$$\begin{aligned}\Omega' &= B \\ \mathcal{F}' &= \mathcal{P}(B),\end{aligned}$$

especially in finite/ countable setting. Then $(\Omega', \mathcal{F}', \mathbb{P}\left(\cdot \mid B\right))$ also satisfies rules to be a probability measure on Ω'.

We have

$$\begin{aligned}\mathbb{P}\left(A \cap B\right) =& \mathbb{P}\left(A\right)\mathbb{P}\left(B \mid A\right) \\ \mathbb{P}\left(A_1 \cap A_2 \cap \cdots \cap A_n\right) =& \mathbb{P}\left(A_1\right)\mathbb{P}\left(A_2 \mid A_1\right)\mathbb{P}\left(A_3 \mid A_1 \cap A_2\right) \\ & \cdots \mathbb{P}\left(A_n \mid A_1 \cap \cdots \cap A_{n-1}\right)\end{aligned}$$

Example. Uniform permutation $(\sigma(1), \sigma(2), \dots, \sigma(n)) \in \Sigma_n$. We claim that

$$\begin{aligned} &\mathbb{P}\left(\sigma(k) = i_k \mid \sigma(1) = i, \dots, \sigma(k-1) = i_{k-1}\right) \\ =&\begin{cases} 0, & \text{if } i_k \in \{i, \dots, i_{k-1}\} \\ \dfrac{1}{n-k+q}, & \text{if otherwise} \end{cases} \end{aligned}$$

Proof. We have

$$\begin{aligned} &\mathbb{P}\left(\sigma(k) = i_k \mid \sigma(1) = i_1, \dots, \sigma(k-1) = i_{k-1}\right) \\ =&\frac{\mathbb{P}\left(\sigma(1) = i_1, \dots, \sigma(k) = i_k\right)}{\mathbb{P}\left(\sigma(1) = i_1, \dots, \sigma(k-1) = i_{k-1}\right)} \\ =&\frac{\frac{(n-k)!}{n!}}{\frac{(n-k+1)!}{n!}} = \frac{1}{n-k+1}. \end{aligned}$$

■

1.9 Law of Total Probability & Bayes' Formula

Definition 1.5

$(B_1, B_2, \dots) \subseteq \Omega$ is a *partition* of Ω if $\Omega = \cup_n B_n$ and (B_n) are disjoint.

Theorem 1.2

(B_n) a finite or countable partition of Ω with $B_n \in \mathcal{F}$ for all n such that $\mathbb{P}(B_n) > 0$. Then for all $A \in \mathcal{F}$:

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \mid B_n)\, \mathbb{P}(B_n).$$

This is also called "Partition Theorem".

Proof. Note that $\cup_n (A \cap B_n) = A$. So we have

$$\mathbb{P}(A) = \sum_{n \geq 1} \mathbb{P}(A \cap B_n) = \sum_n \mathbb{P}(A \mid B_N)\, \mathbb{P}(B_n).$$

■

Theorem 1.3: Bayes' Formula

With the same setup as above, we have

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(A \cap B_N)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_n)\,\mathbb{P}(B_n)}{\sum\limits_m \mathbb{P}(A \mid B_m)\,\mathbb{P}(B_m)}.$$

Rephrasing for $n = 2$, we have $\mathbb{P}(B \mid A)\underbrace{\mathbb{P}(A)}_{given} = \underbrace{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}_{given} = \mathbb{P}(A \cap B)$.

Example. Lecture course has $\frac{2}{3}$ of the lectures on weekdays and $\frac{1}{3}$ on weekends. We have

$$\mathbb{P}(\text{forget notes} \mid \text{weekday}) = \frac{1}{8}$$
$$\mathbb{P}(\text{forget notes} \mid \text{weekend}) = \frac{1}{2}$$

What is $\mathbb{P}(\text{weekend} \mid \text{forget notes})$?

We have $B_1 = \{\text{weekday}\}$ and $B_2 = \{\text{weekend}\}$ and $A = \{\text{forget notes}\}$. So we have

$$\mathbb{P}(A) = \frac{2}{3} \cdot \frac{1}{8} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{12} + \frac{1}{6} = \frac{1}{4}.$$

And by Bayes' Formula, we have

$$\mathbb{P}(B_2 \mid A) = \frac{\frac{1}{3} \cdot \frac{1}{2}}{\frac{1}{4}} = \frac{2}{3}.$$

Example (Disease testing). If p are infected and $1 - p$ are not, and we have

$$\mathbb{P}(\text{positive} \mid \text{infected}) = 1 - \alpha$$
$$\mathbb{P}(\text{positive} \mid \text{not infected}) = \beta.$$

Ideally, you want both α, β to be small. Of course, we want p to be small as well. We want to find $\mathbb{P}(\text{infected} \mid \text{positive})$. By LTP, we have

$$\mathbb{P}(\text{positive}) = p(1 - \alpha) + (1 - p)\beta.$$

Using Bayes', we have

$$\mathbb{P}(\text{infected} \mid \text{positive}) = \frac{p(1 - \alpha)}{p(1 - \alpha) + (1 - p)\beta}.$$

Suppose $p \ll \beta$, we have $p(1 - \alpha) \ll (1 - p)\beta$. The probability is approximately $\frac{p(1-\alpha)}{(1-p)\beta} \sim \frac{p}{\beta}$ which is small.

Example (Simpson's Paradox)**.** If the scientists want to know if jelly beans make your tongue change color? Studies give results:

Oxford	Change	No change	% change
Blue	15	22	41 %
Green	5	8	38 %

Cambridge	Change	No change	% change
Blue	10	3	77 %
Green	23	14	62 %,

but if you add them up, you get

Total	Change	No change	% change
Blue	25	25	50 %
Green	28	22	56 %.

Lecture 7

3 Feb. 2022

We continue from the Simpson's Paradox example. Let $A = \{\text{change color}\}$, $B = \{\text{blue}\}$, $B^c = \{\text{green}\}$, $C = \{\text{Cambridge}\}$ and $C^c = \{\text{Oxford}\}$. We have

$$\begin{aligned} \mathbb{P}\left(A \mid B \cap C\right) &> \mathbb{P}\left(A \mid B^c \cap C\right) \\ \mathbb{P}\left(A \mid B \cap C^c\right) &> \mathbb{P}\left(A \mid B^c \cap C^c\right). \end{aligned}$$

But it is not true that $\mathbb{P}\left(A \mid B\right) > \mathbb{P}\left(A \mid B^c\right)$. LTP for conditional probabilities is the following. Suppose $C_1, C_2, \dots$ is a partition of B, and we have

$$\begin{aligned} \mathbb{P}\left(A \mid B\right) &= \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}\left(B\right)} = \frac{\mathbb{P}\left(A \cap (\cup_n C_n)\right)}{\mathbb{P}\left(B\right)} \\ &= \frac{\mathbb{P}\left(\cup_n (A \cap C_n)\right)}{\mathbb{P}\left(B\right)} = \frac{\sum_n \mathbb{P}\left(A \cap C_n\right)}{\mathbb{P}\left(B\right)} \\ &= \frac{\sum_n \mathbb{P}\left(A \mid C_n\right)\mathbb{P}\left(C_n\right)}{\mathbb{P}\left(B\right)} = \sum_n \mathbb{P}\left(A \mid C_n\right)\frac{\mathbb{P}\left(B \cap C_n\right)}{\mathbb{P}\left(B\right)} \end{aligned}$$

So in conclusion, we have

$$\mathbb{P}\left(A \mid B\right) = \sum_n \mathbb{P}\left(A \mid C_n\right)\mathbb{P}\left(C_n \mid B\right).$$

Special Case:

- If all $\mathbb{P}(C_n)$ are equal, then $\mathbb{P}(C_n \mid B)$ are all equal.
- If $\mathbb{P}(A \mid C_n)$ are all equal. Note that $\sum_n \mathbb{P}(C_n \mid B) = 1$. Then we have

$$\mathbb{P}(A \mid B) = \mathbb{P}(A \mid C_n).$$

Example. Uniform permutation $(\sigma(1), \sigma(2), \ldots, \sigma(52)) \in \Sigma_{52}$ ("well-shuffled cards"). We call $\{1, 2, 3, 4\}$ the aces. We consider $A = \{\sigma(1), \sigma(2) \text{ aces}\}$, and $B = \{\sigma(1) \text{ ace}\} = \{\sigma(1) \leq 4\}$, $C_i = \{\sigma(1) = i\}$.

Note $\mathbb{P}(A \mid C_i) = \mathbb{P}(\sigma(2) \in \{1, 2, 3, 4\} \mid \sigma(1) = i) = \frac{3}{51}$ for $i \leq 4$ by previous example. And we have $\mathbb{P}(C_i) = \frac{1}{52}$. So we have $\mathbb{P}(A \mid B) = \frac{3}{51}$. In total, we have

$$\mathbb{P}(A) = \mathbb{P}(B) \times \mathbb{P}(A \mid B) = \frac{4}{52} \times \frac{3}{51}.$$

2 Discrete Random Variables

Motivation: Roll two dices. $\Omega = \{1, \ldots, 6\}^2 = \{(i, j) \mid 1 \leq i, j \leq 6\}$. If we restrict attention to first dice $\{(i, j) \mid i = 3\}$; sum of dices $\{(i, j) \mid i + j = 8\}$; max of dice $\{(i, j) \mid i, j \leq 4, i \text{ or } j = 4\}$.

Goal: "Random real-valued measurements".

Definition 2.1

A *discrete random variable* X (often denoted by RV) on a probability space $(\Omega, \mathcal{F}, \mathbb{P}())$ is a function $X : \Omega \to \mathbb{R}$ such that

1. $\{\omega \in \Omega \mid X(\omega) = x\} \in \mathcal{F}$.
2. $\text{im}(X)$ is finite or countable (subset of $\mathbb{R}$).

We can write $\{\omega \in \Omega \mid X(\omega) = x\}$ as $\{X = x\}$. So $\mathbb{P}(X = x)$ is valid. And the image is often $\mathbb{Z}$ or $\{0, 1\}$ for example, instead of $\{\text{Heads}, \text{Tails}\}$.

If Ω is finite or countable, and $\mathcal{F} = \mathcal{P}(\Omega)$, both requirements hold automatically.

Example (Part II Applied Probability)**.** If we consider the arrival problem, we have

$\Omega = \{\text{countable subsets } (a_1, a_2, \dots) \text{ of } (0, \infty)\}$. Then,

$$\begin{aligned} N_t &= \text{number of arrivals by time t} \\ &= |\{a_i \mid a_i \leq t\}| \in \{0, 1, 2, \dots\} \end{aligned}$$

is a discrete RV for each time t.

Definition 2.2

The *probability mass function* (p.m.f.) of discrete RV X is the function $p_X : \mathbb{R} \to [0, 1]$ given by

$$p_X(x) = \mathbb{P}(X = x) \quad \forall x \in \mathbb{R}.$$

Note.

- If $x \notin \operatorname{im}(X)$ (that is, $X(\omega)$ never takes value x), then

$$p_X(x) = \mathbb{P}(\omega \in \Omega \mid X(\omega) = x) = \mathbb{P}(\varnothing) = 0.$$

- $$\begin{aligned} \sum_{x \in (X)} p_X(x) &= \sum_{x \in \operatorname{im}(x)} \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}) \qquad . \\ &= \mathbb{P}\left(\cup_{x \in \operatorname{im}(X)} \{\omega \in \Omega \mid X(\omega) = x\}\right) = \mathbb{P}(\Omega) = 1 \end{aligned}$$

Example (Indicator Function)**.** Event $A \in \mathcal{F}$, define $\mathbf{1}_A : \omega \to \mathbb{R}$ by

$$\mathbf{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A \\ 0, & \text{if } \omega \notin A \end{cases}$$

called the *indicated function* of A. $\mathbf{1}_A$ is a discrete RV with $\operatorname{im}(\mathbf{1}) = \{0, 1\}$. The probability mass function is

$$\begin{aligned} p_{\mathbf{1}_A}(1) &= \mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A) \\ p_{\mathbf{1}_A}(0) &= \mathbb{P}(\mathbf{1}_A = 0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A) \\ p_{\mathbf{1}_A}(x) &= 0 \quad \forall x \notin \{0, 1\}. \end{aligned}$$

It encodes "did A happen" as a real number.

Remark. Given a probability mass function, we can always construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a RV defined on it with this pmf.

- $\Omega = \mathrm{im}(X)$. That is, $\{x \in \mathbb{R} \mid p_X(x) > 0\}$;
- $\mathcal{F} = \mathcal{P}(\Omega)$;
- $\mathbb{P}(\{x\}) = p_X(x)$ and extend it to all $A \in \mathcal{F}$.

Lecture 8

5 Feb. 2022

2.1 Discrete Probability Distributions

We first start with distributions with Ω finite.

2.1.1 Bernoulli Distribution ("biased coin toss")

We have $X \sim \mathrm{Bern}(p)$ with $p \in [0,1]$, and

$$\begin{aligned} \mathrm{im}(X) &= \{0,1\} \\ p_X(1) &= \mathbb{P}(X=1) = p \\ p_X(0) &= \mathbb{P}(X=0) = 1-p. \end{aligned}$$

Example. $\mathbf{1}_A \sim \mathrm{Bern}(p)$ with $p = \mathbb{P}(A)$.

2.1.2 Binomial Distribution

We have $X \sim \mathrm{Bin}(n,p)$ with $n \in \mathbb{Z}^+, p \in [0,1]$. ("Toss coin n times, count number of heads") We have

$$\begin{aligned} \mathrm{im}(X) &= \{0,1,\dots,n\} \\ p_X(k) &= \mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

Do check that $\sum_{k=0}^{n} p_X(k) = 1$ by binomial expansion. Next, we consider $\Omega = \mathbb{N}$. ("Ways of choosing a random integer")

Next, we consider the case when Ω is countable. This is slightly deviating from the order which they were taught in the lectures.

2.1.3 Geometric Distribution ("Waiting for success")

We have $X \sim \mathrm{Geom}(p)$ with $p \in (0,1]$. ("Toss a coin with $\mathbb{P}$ (head) $= p$ until a head appears. Count how many trials were needed") So

$$\begin{aligned} \mathrm{im}(X) &= \{1, 2 \dots\} \\ p_X(k) &= \mathbb{P}\left((n-1) \text{ failures, then success on last}\right) = (1-p)^{k-1}p. \end{aligned}$$

Indeed, we have

$$\sum_{k \geq 1} (1-p)^{k-1} p = p \sum_{\ell \geq 0} (1-p)^{\ell} = \frac{p}{1-(1-p)} = 1.$$

Alternatively, we can count how many failures before a success. So

$$\begin{aligned} \mathrm{im}(Y) &= \{0, 1, 2, \dots\} \\ p_Y(k) &= \mathbb{P}\left(k \text{ failures, then success on next}\right) = (1-p)^k p. \end{aligned}$$

Similarly, we have

$$\sum_{k \geq 0} (1-p)^k p = 1.$$

2.1.4 Poisson Distribution

We have $X \sim \mathrm{Po}(\lambda)$ (or $\mathrm{Poi}(\lambda)$ with parameter λ), and

$$\begin{aligned} \mathrm{im}(X) &= \{0, 1, 2, \dots\} \\ p_X(k) &= e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

Note that $\sum_{k \geq 0} \mathbb{P}\left(X = k\right) = e^{-\lambda} \sum_{k \geq 0} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$.

Motivation: Consider $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$, we split time interval $[0, \lambda]$ into n small intervals. If the probability of arrival in each interval is p, and independent across intervals. The total number of arrivals is X_n, and note by fixing k and taking $n \to \infty$,

$$\begin{aligned} \mathbb{P}\left(X_n = k\right) &= \binom{n}{k} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\ &= \frac{n!}{n^k (n-k)!} \times \frac{\lambda^k}{k!} \times (1 - \frac{\lambda}{n})^n \times (1 - \frac{\lambda}{n})^{-k} \\ &\to 1 \times \frac{\lambda^k}{k!} \times e^{-\lambda} \times 1 = e^{-\lambda} \frac{\lambda^k}{k!}. \end{aligned}$$

2.2 More Than One RV

Motivation: Roll a die, and the outcome is $X \in \{1,2,3,4,5,6\}$. If we consider the events

$$A = \{1 \text{ or } 2\},\ B = \{1 \text{ or } 2 \text{ or } 3\},\ C = \{1 \text{ or } 3 \text{ or } 5\}.$$

We have

$$\mathbf{1}_A \sim \mathrm{Bern}(\frac{1}{3}),\ \mathbf{1}_B \sim \mathrm{Bern}(\frac{1}{2}),\ \mathbf{1}_C \sim \mathrm{Bern}(\frac{1}{2}).$$

Note $\mathbf{1}_A \leq \mathbf{1}_B$ for all outcomes, but $\mathbf{1}_A \leq \mathbf{1}_C$ is not true for all outcomes.

Definition 2.3

$X_1, \dots, X_n$ discrete RVs, then we say $X_1, \dots, X_n$ are *independent* if

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \mathbb{P}\left(X_1 = x_1\right) \cdots \mathbb{P}\left(X_n = x_n\right) \quad \forall x_1, \dots, x_n \in \mathbb{R}.$$

Remark. It suffices to check that $\forall x_i \in \mathrm{im}(X_i)$.

Example. $X_1, \dots, X_n$ independent RVs each with the $\mathrm{Bern}(p)$ distribution. We study $S_n = X_1 + \cdots + X_n$. Then

$$\begin{aligned}
\mathbb{P}\left(S_n = k\right) &= \sum_{\substack{x_1 + \cdots + x_n = k \\ x_i \in \{0,1\}}} \mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) \\
&= \sum_{\substack{x_1 + \cdots + x_n = k \\ x_i \in \{0,1\}}} \mathbb{P}\left(X_1 = x_1\right) \cdots \mathbb{P}\left(X_n = x_n\right) \text{ by independence} \\
&= \sum_{\substack{x_1 + \cdots + x_n = k \\ x_i \in \{0,1\}}} p^{|\{i | x_i = 1\}|} (1-p)^{|\{i | x_i = 0\}|} \\
&= \sum_{\substack{x_1 + \cdots + x_n = k \\ x_i \in \{0,1\}}} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k}.
\end{aligned}$$

So $S_n \sim \mathrm{Bin}(n, k)$.

Example. Consider the uniform permutation $(\sigma(1), \dots, \sigma(n))$ of the integers $1, 2, \dots, n$. We claim that $\sigma(1)$ and $\sigma(2)$ are not independent.

It suffices to find i_1, i_2 such that

$$\mathbb{P}\left(\sigma(1) = i_1, \sigma(2) = i_2\right) \neq \mathbb{P}\left(\sigma(1) = i_1\right)\mathbb{P}\left(\sigma(2) = i_2\right).$$

For example,

$$\mathbb{P}\left(\sigma(1) = 1, \sigma(2) = 1\right) = 0 \neq \mathbb{P}\left(\sigma(1) = 1\right)\mathbb{P}\left(\sigma(2) = 1\right) = \frac{1}{n} \cdot \frac{1}{n}.$$

We also have that if $X_1, \ldots, X_n$ are independent, $\forall A_1, \ldots, A_n \in \mathbb{R}$ countable,

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \mathbb{P}\left(X_1 \in A_1\right) \cdots \mathbb{P}\left(X_n \in A_n\right).$$

Lecture 9

8 Feb. 2022

2.3 Expectation

If we have $(\Omega, \mathcal{F}, \mathbb{P})$ and X a discrete RV. For now, X only takes non-negative values. "$X \geq 0$"

Definition 2.4

The expectation of X (or *expected value* or *mean*).

$$\mathbb{E}[X] = \sum_{x \in \mathrm{im}(X)} x\mathbb{P}(X = x) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}).$$

The latter definition is only used in a later proof once.

Remark. Informally, this is the "average of values taken by X, weighted by p_X".

Example. If we have X uniform on $\{1, 2, \ldots, 6\}$ (e.g., a die), we have

$$\mathbb{E}[X] = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \cdots + \frac{1}{6} \times 6 = 3.5.$$

Note that $\mathbb{E}[X] \notin \mathrm{im}(X)$.

Example. If $X \sim \text{Bin}(n, p)$. We have

$$\mathbb{E}[X] = \sum_{k=0}^{n} k\mathbb{P}(X = k) = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k}.$$

Note that

$$k\binom{n}{k} = \frac{k \times n!}{k! \times (n-k)!} = \frac{n!}{(k-1)!(n-k)!} = \frac{n \times (n-1)!}{(k-1)! \times (n-k)!} = n\binom{n-1}{k-1}.$$

So we have

$$\begin{aligned}\mathbb{E}[X] &= n\sum_{k=1}^{n} \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= np\sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\ &= np\sum_{\ell=0}^{n-1} \binom{n-1}{\ell} p^{\ell} (1-p)^{(n-1)-\ell} \\ &= np(p + (1-p))^{n-1} \\ &= np.\end{aligned}$$

Note. We would like to say that

$$\mathbb{E}[\text{Bin}(n, p)] = \mathbb{E}[\text{Bern}(p)] + \cdots + \mathbb{E}[\text{Bern}(p)].$$

We will show this later in the lecture.

Example. If $X \sim \text{Poisson}(\lambda)$,

$$\begin{aligned}\mathbb{E}[X] = \sum_{k \geq 0} k\mathbb{P}(X = k) &= \sum_{k \geq 0} k \cdot e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k \geq 1} e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\ &= \lambda \sum_{k \geq 1} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda \sum_{\ell \geq 0} e^{-\lambda} \frac{\lambda^{\ell}}{\ell!} \\ &= \lambda.\end{aligned}$$

Note. We would like to say that

$$\mathbb{E}[\text{Poisson}(\lambda)] \approx \mathbb{E}[\text{Bin}(n, \frac{\lambda}{n})] = \lambda.$$

But it is not true in general that $\mathbb{P}(X_n = k) \approx \mathbb{P}(X = k) \implies \mathbb{E}[X_n] \approx \mathbb{E}[X]$.

For a general X (not necessarily $X \geq 0$),

$$\mathbb{E}[X] = \sum_{x \in \text{im}(X)} x\mathbb{P}(X = x)$$

unless $\sum_{\substack{x>0 \\ x\in\text{im}(X)}} x\mathbb{P}(X = x) = +\infty$ and $\sum_{\substack{x<0 \\ x\in\text{im}(X)}} x\mathbb{P}(X = x) = -\infty$, then we say $\mathbb{E}[X]$ is not defined. (because we don't want to do arithmetic with infinity)

If only one of them holds, we say that $\mathbb{E}[X]$ is $+\infty$ and $-\infty$ respectively. (some people say that it is undefined, but the lecturer disagrees with it) If neither of them hold, we say X is *integrable*.

Example. Most examples in the course are integrable except the following. Let

$$\mathbb{P}(x = n) = \frac{6}{\pi^2} \times \frac{1}{n^2}. \qquad x \geq 1$$

Note that $\sum \mathbb{P}(X = n) = 1$. Then

$$\mathbb{E}[X] = \sum \frac{6}{\pi^2} \times \frac{1}{n} = +\infty.$$

If instead, let

$$\mathbb{P}(X = n) = \frac{3}{\pi^2} \times \frac{1}{n^2}. \qquad n \in \mathbb{Z} \setminus \{0\}$$

Then $\mathbb{E}[X]$ is not defined.

Example. $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$.

Property.

1. If $X \geq 0$, then $\mathbb{E}[X] \geq 0$ with equality if and only if $\mathbb{P}(X = 0) = 1$.

   Proof. $\mathbb{E}[X] = \sum_{\substack{x\in\text{im}(X) \\ x\neq 0}} x\mathbb{P}(X = x)$. ■

2. If $\lambda, c \in \mathbb{R}$, then

   a) $\mathbb{E}[X + c] = \mathbb{E}[x] + c$;

   b) $\mathbb{E}[\lambda X] = \lambda \mathbb{E}[X]$.

3. a) For X, Y random variables (both integrable) on same probability space,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

   b) In fact, for $\lambda, \mu \in \mathbb{R}$,

$$\mathbb{E}[\lambda X + \mu Y] = \lambda \mathbb{E}[X] + \mu \mathbb{E}[Y].$$

Proof. For Ω countable, we have

$$\begin{aligned}
\mathbb{E}[\lambda X + \mu Y] &= \sum_{\omega \in \Omega} (\lambda X(\omega) + \mu Y(\omega)) \mathbb{P}(\{\omega\}) \\
&= \lambda \sum_{\omega \in \Omega} X(\omega) \mathbb{P}(\{\omega\}) + \mu \sum_{\omega \in \Omega} Y(\omega) \mathbb{P}(\{\omega\}) \\
&= \lambda \mathbb{E}[X] + \mu \mathbb{E}[Y].
\end{aligned}$$

■

Note that property (2) is a special case of property (3). Similarly, it extends to n RVs. It is called *linearity of expectation*.

Remark. 1. Independence is not a condition.

Lecture 10

10 Feb. 2022

Corollary 2.1

$X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[X]$.

Proof. Note $X = (X - Y) + Y$. By linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[X - Y] + \mathbb{E}[X].$$

Because $X - Y \geq 0$ and property 1, $\mathbb{E}[X - Y] \geq 0$. ■

Key applications of expectation are counting problems.

Example. Let $(\sigma(1), \ldots, \sigma(n))$ be uniform on Σ_n, and $Z = |\{i : \sigma(i) = i\}|$ be the number of fixed points. Let $A_i = \{\sigma(i) = i\}$. (recall that A_i are not independent) Note

$$Z = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}.$$

And we have

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[\mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}] \\ &= \mathbb{E}[\mathbf{1}_{A_1}] + \cdots + \mathbb{E}[\mathbf{1}_{A_n}] \\ &= \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) \\ &= \frac{1}{x} \times n = 1. \end{aligned}$$

Note that this is the same answer as $\mathrm{Bin}(n, \frac{1}{n})$, but they are not the same distribution.

Example. If X takes values in $\mathbb{Z}_{\geq 0}$.

$$\mathbb{E}[X] = \sum_{k \geq 1} \mathbb{P}(X \geq k).$$

Proof. Carefully re-arrange the summands. ∎

Proof. Write $X = \sum_{k \geq 1} \mathbf{1}_{X \geq k}$, and take expectation of both sides

$$\mathbb{E}[X] = \mathbb{E}[\sum \mathbf{1}_{X \geq k}] = \sum \mathbb{E}[\mathbf{1}_{X \geq k}] = \sum \mathbb{P}(X \geq k).$$

∎

Theorem 2.1: Markov's Inequality

Let $X \geq 0$ be a random variable. Then $\forall a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Remark. If we take $a = \frac{\mathbb{E}[X]}{2}$, it is not useful since it just tells us that the probability is less than 2. It gets more useful when a is large.

Proof. Observe that $X \geq a\mathbf{1}_{X \geq a}$. Taking expectations,

$$\mathbb{E}[X] \geq a\mathbb{E}[\mathbf{1}_{X \geq a}] = a\mathbb{P}(X \geq a),$$

and rearrange. ■

Remark. This is also true for continuous RVs.

Proposition 2.1

Let $f : \mathbb{R} \to \mathbb{R}$ be a function, then $f(X)$ is also a random variable. And

$$\mathbb{E}[f(X)] = \sum_{x \in \mathrm{im}(X)} f(x)\mathbb{P}(X = x)$$

when the expectation exists.

Proof. Let $A = \mathrm{im}(f(X)) = \{ f(x) \mid x \in \mathrm{im}(X) \}$. Starting with the right-hand side,

$$\begin{aligned}
\sum_{x \in \mathrm{im}(X)} f(x)\mathbb{P}(X = x) &= \sum_{y \in A} \sum_{\substack{x \in \mathrm{im}(X) \\ f(x) = y}} f(x)\mathbb{P}(X = x) \\
&= \sum_{y \in A} y \sum_{\substack{x \in \mathrm{im}(X) \\ f(x) = y}} \mathbb{P}(X = x) \\
&= \sum_{y \in A} y\mathbb{P}(f(X) = y) \\
&= \mathbb{E}[f(X)]
\end{aligned}$$

■

Consider the random variables.

$$\begin{aligned}
U_n &\sim \mathrm{Uniform}(\{-n, -n+1, \ldots, n\}) \\
V_n &\sim \mathrm{Uniform}(\{-n, n\}) \\
Z_n &= 0 \\
S_n &\sim n - 2\mathrm{Bin}(n, 1/2) \qquad \text{(random walk for } n \text{ step)}
\end{aligned}$$

All of these have expectation 0. *Variance* "measure how concentrated a RV is around its mean".

Definition 2.5

The *variance* of X is
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Property. 1. $\mathrm{Var}(X) \geq 0$ with equality if and only if $\mathbb{P}(X = \mathbb{E}[X]) = 1$.

2. Alternatively,
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

   Proof. Write $\mu = \mathbb{E}[X]$, then
$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[X^2 - 2\mu X + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2.\end{aligned}$$
   ■

3. If $\lambda, c \in \mathbb{R}$,

   - $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$;
   - $\mathrm{Var}(X + c) = \mathrm{Var}(X)$;

     Proof. $\mathbb{E}[X + c] = \mu + c$, and
$$\mathrm{Var}(X+c) = \mathbb{E}[(X + c - (\mu + c))^2 = \mathbb{E}[(X-\mu)^2] = \mathrm{Var}(X).$$
     ■

Lecture 11

12 Feb. 2022

Example. $X \sim \mathrm{Poisson}(\lambda)$, then $\mathbb{E}[X] = \lambda$, and we have
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \lambda^2.$$

"Falling factorial trick": sometimes $\mathbb{E}[X(X-1)]$ is easier than $\mathbb{E}[X^2]$.

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k\geq 2} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \lambda^2 e^{-\lambda}\sum_{k\geq 2}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.\end{aligned}$$

And by linearity of expectation,

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \lambda^2 + \lambda.$$

So the variance is $\operatorname{Var}(X) = x$.

Example. Take $Y \sim \operatorname{Geo}(p) \in \{1, 2, 3, \dots\}$, and $\mathbb{E}[Y] = \frac{1}{p}$, $\operatorname{Var}(Y) = \frac{1-p}{p^2}$.

Note. When λ is large, $\operatorname{Var}(X) = \mathbb{E}[X]$. When p is small, $\operatorname{Var}(Y) \approx \frac{1}{p^2} = (\mathbb{E}[X])^2$. So Poisson distribution is more concentrated.

Example. When $X \sim \operatorname{Bern}(p)$, we have $\mathbb{E}[X] = 1 \times p = p$, and $\mathbb{E}[X^2] = 1^2 \times p = p$, so

$$\operatorname{Var}(X) = p - p^2 = p(1-p).$$

Before we study the variance of binomial distribution, we develop some theory.

Lemma 2.1

If X, Y are independent RVs, and f, g functions $\mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[f(X)g(X)] = \mathbb{E}[f(X)]\,\mathbb{E}[f(Y)].$$

Proof. We have

$$\begin{aligned}\mathbb{E}[f(X)g(X)] &= \sum_{\substack{x\in\operatorname{im}(X)\\ y\in\operatorname{im}(Y)}} f(x)g(y)\mathbb{P}(X=x, Y=y) \\ &= \sum_{\substack{x\in\operatorname{im}(X)\\ y\in\operatorname{im}(Y)}} f(x)g(y)\,\mathbb{P}(X=x)\,\mathbb{P}(Y=y) \\ &= \sum_{x\in\operatorname{im}(X)} f(X)\,\mathbb{P}(X=x) \sum_{y\in\operatorname{im}(Y)} g(y)\,\mathbb{P}(Y=y) \\ &= \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].\end{aligned}$$

■

Example. If we have $f(x) = g(x) = x$, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

Example. When $f(x) = g(x) = z^x$ or $f(x) = g(x) = e^{tx}$, the lemma is useful.

Lemma 2.2

If $X_1, \ldots, X_n$ are independent,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

Proof. It suffices to prove the case when $n = 2$. Let $\mathbb{E}[X] = \mu$ and $\mathbb{E}[Y] = \nu$, and $\mathbb{E}[X + Y] = \mu + \nu$.

$$\begin{aligned}
\mathrm{Var}(X + Y) &= \mathbb{E}[(X + Y - \mu - \nu)^2] \\
&= \mathbb{E}[(X - \mu)^2] + \mathbb{E}[(Y - \nu)^2] + 2\,\mathbb{E}[(X - \mu)(Y - \nu)] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathbb{E}[X - \mu]\,\mathbb{E}[Y - \nu] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}$$

■

Definition 2.6

If X, Y are RVs. Their covariance is $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$.

Remark. It measures how dependent X, Y are and in which direction. $\mathrm{Cov}(X, Y) > 0$ means X large and Y large, and $\mathrm{Cov}(X, Y) < 0$ means X large and Y small.

Property.

1. $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
2. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
3. Alternatively,
$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$
It is often more useful, and it's nice if $\mathbb{E}[X] = 0$.

Proof.

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}[(X-\mu)(Y-\nu)] \\ &= \mathbb{E}[XY] - \mu\,\mathbb{E}[Y] - \nu\,\mathbb{E}[X] + \mu\nu \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

∎

4. For $\lambda, c \in \mathbb{R}$,
    - $\operatorname{Cov}(c, X) = 0$
    - $\operatorname{Cov}(X + c, Y) = \operatorname{Cov}(X, Y)$.
    - $\operatorname{Cov}(\lambda X, Y) = \lambda \operatorname{Cov}(X, Y)$.
5. $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y)$.
6. Covariance is linear in each argument. That is,
$$\operatorname{Cov}(\sum \lambda_i X_i, Y) = \sum \lambda_i \operatorname{Cov}(X_i, Y)$$
and
$$\operatorname{Cov}(\sum \lambda_i X_i, \sum \mu_j Y_j) = \sum\sum \lambda_i \mu_j \operatorname{Cov}(X_i, Y_j).$$
The special case is
$$\begin{aligned}\operatorname{Var}(\sum_{i=1}^n X_i) &= \operatorname{Cov}(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i) \\ &= \sum_{i=1}^n \operatorname{Var}(X_i) + \sum_{i\neq j} \operatorname{Cov}(X_i, Y_j).\end{aligned}$$

Remark. We know that X, Y independent implies $\operatorname{Cov}(X, Y) = 0$, but the converse is false.

Lecture 12

15 Feb. 2022

Example. $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$ if X and Y are independent.

Take $Y = -X$, $\operatorname{Var}(Y) = \operatorname{Var}(-X) = \operatorname{Var}(X)$. But
$$0 = \operatorname{Var}(0) = \operatorname{Var}(X + Y) \neq \operatorname{Var}(X) + \operatorname{Var}(Y) = 2\operatorname{Var}(X)$$
unless X and Y are deterministic.

Example. Again let $(\sigma(1), \dots, \sigma(n))$ uniformly on Σ_n, and let $A_i = \{\, \sigma \mid \sigma(i) = i \,\}$, and $N = \mathbf{1}_{A_1} + \cdots + \mathbf{1}_{A_n}$ be the number of fixed points. We've already seen

$$\mathbb{E}[N] = n \times \frac{1}{n} = 1.$$

Note that the A_is are not independent, and

$$\begin{aligned}
\mathrm{Var}(\mathbf{1}_{A_i}) &= \frac{1}{n}(1 - \frac{1}{n}) \\
\mathrm{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) &= \mathbb{E}[\mathbf{1}_{A_i}\mathbf{1}_{A_j}] - \mathbb{E}[\mathbf{1}_{A_i}]\,\mathbb{E}[\mathbf{1}_{A_j}] \\
&= \mathbb{E}[\mathbf{1}_{A_i \cap A_j}] - \mathbb{E}[\mathbf{1}_{A_i}]\,\mathbb{E}[\mathbf{1}_{A_j}] \\
&= \mathbb{P}(A_i \cap A_j) - \mathbb{P}(A_i)\,\mathbb{P}(A_j) \\
&= \frac{1}{n(n-1)} - \frac{1}{n} \times \frac{1}{n} \\
&= \frac{1}{n^2(n-1)} > 0.
\end{aligned}$$

So

$$\begin{aligned}
\mathrm{Var}(N) &= \sum_{i=1}^{n} \mathrm{Var}(\mathbf{1}_{A_i}) + \sum_{i \neq j} \mathrm{Cov}(\mathbf{1}_{A_i}, \mathbf{1}_{A_j}) \\
&= n \times \frac{1}{n}(1 - \frac{1}{n}) + n(n-1) \times \frac{1}{n^2(n-1)j} \\
&= 1 - \frac{1}{n} + \frac{1}{n} = 1.
\end{aligned}$$

Compare this with $\mathrm{Bin}(n, \frac{1}{n})$. The binomial distribution has expectation 1 and variance $1 - \frac{1}{n}$. So the binomial distribution is not too disimilar to the number of fixed points.

Theorem 2.2: Chebyshev's Inequality

Let X be a RV, $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$, then

$$\mathbb{P}(|X - \mu| \geq \lambda) = \frac{\mathrm{Var}(X)}{\lambda^2}.$$

Remark. It's easier to remember the proof, not the statement.

Proof. Apply Markov's inequality to $(X - \mu)^2$,

$$\mathbb{P}((X - \mu)^2 \geq \lambda^2) \leq \frac{\mathbb{E}[(X - \mu)^2}{\lambda^2} = \frac{\mathrm{Var}(X)}{\lambda^2}.$$

And we are done. ■

Remark.

1. If instead we apply Markov's inequality to $|X-\mu|$, $\mathbb{E}[|X-\mu|]$ is less nice than $\mathrm{Var}(X)$.
2. Chebyshev's inequality gives better bounds than Markov's inequality.
3. Note that it can apply to all RVs, not just ≥ 0.
4. $\mathrm{Var}(X) < \infty$ is a stronger condition than $\mathbb{E}[X] < \infty$.

Definition 2.7

Quantity $\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of X.

Remark. It has the same unit as X, but it does not have as many nice properties as variance.

If we write $\lambda = k\sqrt{\sigma^2}$ ("k standard deviations") in Chebyshev's inequality, then

$$\mathbb{P}(|X-\mu| \geq k\sqrt{\sigma^2}) \leq \frac{1}{k^2}.$$

This is a nice uniform statement.

Definition 2.8: Conditional Expectation

If we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we defined

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

The *conditional expectation* with the condition $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$ and X a RV is

$$\mathbb{E}[X \mid B] = \frac{\mathbb{E}[X\mathbf{1}_B]}{\mathbb{P}(B)}.$$

Example. If X is a die uniform on $\{1, \ldots, 6\}$,

$$\mathbb{E}[X \mid X \text{ prime}] = \frac{\frac{1}{6}(0+2+3+0+5+0)}{1/2} = \frac{1}{3}(2+3+5) = \frac{10}{3}.$$

Remark. An alternative characterization is

$$\mathbb{E}[X \mid B] = \sum_{x \in \mathrm{im}(X)} x\, \mathbb{P}(X = x \mid B).$$

Proof.

$$\begin{aligned}\sum_{x \in \mathrm{im}(X)} x\, \mathbb{P}(X = x \mid B) &= \sum \frac{x\, \mathbb{P}(\{X = x\} \cap B}{\mathbb{P}(B)} \\ &= \sum \frac{x\, \mathbb{P}(X\mathbf{1}_B = x)}{\mathbb{P}(B)}\end{aligned}$$

and note that $\mathbb{E}[X\mathbf{1}_B] = \sum x\, \mathbb{P}(X\mathbf{1}_B = x)$. ■

Theorem 2.3: Law of Total Expectation

If $(B_1, B_2, \dots)$ is a finite or countably-infinite partition of Ω with $B_n \in \mathcal{F}$ such that $\mathbb{P}(B_n) > 0$ and X a RV, then

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X \mid B_n]\, \mathbb{P}(B_n).$$

Proof.

$$\begin{aligned}\sum_n \mathbb{E}[X \mid B_n]\, \mathbb{P}(B_n) &= \sum_n \mathbb{E}[X\mathbf{1}_{B_n}] \\ &= \mathbb{E}[X \cdot (\mathbf{1}_{B_1} + \cdots + \mathbf{1}_{b_n})] \\ &= \mathbb{E}[X \cdot \mathbf{1}] = \mathbb{E}[X].\end{aligned}$$

■

Remark.

1. We recover Law of Total probability by taking $X = \mathbf{1}_A$.
2. Two-stage randomness where (B_n) describes what happens in stage 1.

Example (Random Sums)**.** If $(X_n)_{n \geq 1}$ are IID (independent and identically distributed) with $\mathbb{E}[X_n] = \mu$, and $N \in \{0, 1, 2, \dots\}$ is a random index independent of (X_n). The sum $S_n = X_1 + \cdots + X_n$ has $\mathbb{E}[S_n] = n\mu$. The random sum

$$\begin{aligned}
\mathbb{E}[S_N] &= \sum_{n\geq 0} \mathbb{E}[S_N \mid N = n]\, \mathbb{P}(N = n) \\
&= \sum \mathbb{E}[S_n]\, \mathbb{P}(N = n) \\
&= \sum n\mu\, \mathbb{P}(N = n) = \mu\, \mathbb{E}[N]
\end{aligned}$$

Lecture 16

24 Feb. 2022

Lecture 17

26 Feb. 2022