Probability Notes

Pramana Saldin

Spring 2024

A rigorous introduction to probability theory at an advanced undergraduate level. Only a minimal amount of measure theory is used, in particular, the theory of Lebesgue integrals is not needed. It is aimed at math majors and Master's degree students, or students in other fields who will need probability in their future careers. Gives an introduction to the basics (Kolmogorov axioms, conditional probability and independence, random variables, expectation) and discusses some classical results with proofs (DeMoivre-Laplace limit theorems, the study of simple random walk on the one dimensional lattice, applications of generating functions).

Professor: Tatyana Shcherbyna.

Contents

1. Events and their probabilities

The following is called *Kolgomorov's axioms for probability*.

January 25, 2024 **Definition 1.1.** A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

1. Ω is a set of possible outcomes (**sample space**)
2. $\mathcal{F}$ is a collection of subsets of Ω called **events**, which we assume to be a σ-**algebra**, i.e., it satisfies the following properties:
   a) $\Omega \in \mathcal{F}$,
   b) if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$, where $A^C := \Omega \setminus A$ is the complement of A in Ω,
   c) if $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_k A_k \in \mathcal{F}$.
3. $\mathbb{P}\colon \mathcal{F} \to [0,1]$ is a **probability measure**:
   a) $\mathbb{P}(\varnothing) = 0$ and $\mathbb{P}(\Omega) = 1$,
   b) (σ-**additivity**) if $A_1, A_2, \ldots$ are countably many disjoint events, then

$$\mathbb{P}\left(\bigcup_k A_k\right) = \sum_k \mathbb{P}(A_k).$$

1.1. Properties of $\mathbb{P}$

January 30, 2024 **Proposition 1.1** (Complements of events)**.**

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

σ-additivity works when the $\{A_i\}_{i=1}^N$ are pairwise disjoint. If instead they are not disjoint, in the case $N = 2$ we have

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

Theorem 1.2 (Inclusion-exculsion formula)**.** Let $\{A_i\}_{i=1}^n$ be a set of events.

$$\begin{aligned}\mathbb{P}(A_1 \cup \cdots \cup A_n) &= \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i_1<i_2} \mathbb{P}(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &\quad + \cdots + (-1)^{n-1}\mathbb{P}(A_1 \cap \cdots \cap A_n) \\ &= \sum_{k=1}^n (-1)^{k-1} \sum_{i_1<\cdots<i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}).\end{aligned}$$

Proof. We induct on n. The base case is clear.

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{k=1}^{n+1} A_k\right) &= \mathbb{P}\left(\left(\bigcup_{k=1}^{n} A_k\right) \cup A_{n+1}\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{n} A_k\right) + \mathbb{P}(A_{n+1}) - \mathbb{P}\left(\left(\bigcup_{k=1}^{n} A_k\right) \cap A_{n+1}\right) \\
&= \sum_{k=1}^{n} (-1)^{k-1} \sum_{i_1<\cdots<i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) + \mathbb{P}(A_{n+1}) \\
&\quad - \sum_{k=1}^{n} (-1)^{k-1} \sum_{i_1<\cdots<i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k} \cap A_{n+1}). \\
&= \sum_{k=1}^{n+1} (-1)^{k-1} \sum_{i_1<\cdots<i_k} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}).
\end{aligned}$$

□

February 01, 2024 **Theorem 1.3** (Monotonicity of probability measure)**.**

1. If $A, B \in \mathcal{F}$ and $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
2. If $A_k \in \mathcal{F}$ for $k \in \mathbb{N}$, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

   If the A_k are disjoint, we have equality.

Proof. (1) Write $B = A \cup (B \setminus A)$. These sets are disjoint, so

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A).$$

(2) We construct disjoint sets $\{B_k\}$ such that $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$. Take $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus (A_1 \cup A_2)$. In other words,

$$B_k \coloneqq A_k \setminus (A_1 \cup \cdots A_{k-1}) = A_k \cap A_1^C \cap \cdots \cap A_{k-1}^C.$$

$\{B_k\}$ are disjoint by construction. Now we show that

$$\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} B_k.$$

By construction, $B_k \subseteq A_k$, so the right-to-left inclusion is done. For the other, suppose $\omega \in \bigcup_{k=1}^{\infty} A_k$. Let k_0 be the first value where $\omega \in A_{k_0}$ (i.e. $\omega \in A_{k_0}$ and $\omega \notin A_1, \ldots, A_{k_0-1}$). Then $\omega \in B_{k_0}$. With this,

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(B_k) && (\mathbb{P}(B_k) \leq \mathbb{P}(A_k)) \\
&\leq \sum_{k=1}^{\infty} \mathbb{P}(A_k).
\end{aligned}$$

□

Definition 1.2. Let A be an event and (A_n) a sequence of events. We say A_n **increases up to** A, denoted $A_n \nearrow A$, if $A_1 \subseteq A_2 \subseteq \cdots$ and $\bigcup_{k=1}^{\infty} A_k = A$. Similarly, we say A_n **decreases down to** A, denoted $A_n \searrow A$, if $A_1 \supseteq A_2 \supseteq \cdots$ and $\bigcap_{k=1}^{\infty} A_k = A$.

Proposition 1.4 (Continuity of probability)**.** If $A_n \nearrow A$ or $A_n \searrow A$, then

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Proof. ($A_n \nearrow A$) Create the sets $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus A_2$, continuing with $B_k \coloneqq A_k \setminus A_{k-1}$. We have

$$\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k = A,$$

so

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}(B_k) \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) \\
&= \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

($A_n \searrow A$)

$$A = \bigcap_{k=1}^{\infty} A_k \implies A^C = \bigcup_{k=1}^{\infty} A_k^C.$$

So we can follow the same proof as before to show

$$\mathbb{P}(A^C) = \lim_{n\to\infty} \mathbb{P}(A_n^C). \qquad \square$$

1.2. Conditional probability

February 06, 2024

Definition 1.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Given $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the **conditional probability of** A **given** B is

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

Theorem 1.5. Fix $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Consider $\mathbb{P}(A \mid B)$ for all $A \in \mathcal{F}$. Then $\mathbb{P}(\bullet \mid B)$ is a probability measure $(\Omega, \mathcal{F}, \mathbb{P}(\bullet \mid B))$ **concentrated on** B (i.e. $\mathbb{P}(B \mid B) = 1$).

Proof. We check the axioms of a probability measure (1.1):

(a) $\mathbb{P}(\varnothing \mid B) = \frac{\mathbb{P}(\varnothing \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\varnothing)}{\mathbb{P}(B)} = 0.$

$\mathbb{P}(\Omega \mid B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$

(b) Let $\{A_k\}$ be a countable collection of disjoint events. Then $\{A_k \cap B\}$ are disjoint as well. Therefore,

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

$\square$

Remark 1.6. Some events under this measure are indistinguishable. For instance,

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

It may make more sense to consider $\mathbb{P}(\bullet \mid B)$ as a probability measure on $\mathcal{F} \cap B \coloneqq \{A \cap B \mid A \in \mathcal{F}\}$.

By rearranging the terms in the definition of conditional probability, $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B)$. Let's further generalize this.

Proposition 1.7. Suppose $A_i \in \mathcal{F}$ for $i = 1, \ldots, n$ and $\mathbb{P}(A_1 \cap \cdots \cap A_n) > 0$. Then

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_1 \mid A_2)\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

We write $A_1 \cap \cdots \cap A_n = A_1 \cdots A_n$ as shorthand.

Proof. Expanding,

$$\mathbb{P}(A_1) \cdot \frac{\mathbb{P}(A_1A_2)}{\mathbb{P}(A_1)} \cdot \frac{\mathbb{P}(A_1A_2A_3)}{\mathbb{P}(A_1A_2)} \cdots \frac{\mathbb{P}(A_1 \cdots A_n)}{\mathbb{P}(A_1 \cdots A_{n-1})} = \mathbb{P}(A_1 \cdots A_n) \qquad \square$$

Recall a **partition** of Ω is a collection of pairwise disjoint sets whose union is Ω.

Theorem 1.8 (Law of total probability)**.** Let $B_i \in \mathcal{F}$ be a countable partition of Ω. For any $A \in \mathcal{F}$,

$$\mathbb{P}(A) = \sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i).$$

Proof. Recall $\mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i) = \mathbb{P}(A \cap B_i)$. So

$$\sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i) = \sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A \cap B_i) = \mathbb{P}(A). \qquad \square$$

Theorem 1.9 (Bayes' formula)**.** Let $B_i \in \mathcal{F}$ be a countable partition of Ω, and let $A \in \mathcal{F}$, $\mathbb{P}(A) > 0$. Then for all k such that $\mathbb{P}(B_k) > 0$, we have

$$\mathbb{P}(B_k \mid A) = \frac{\mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}{\sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Proof.

$$\begin{aligned}\mathbb{P}(B_k \mid A) &= \frac{\mathbb{P}(A \cap B_k)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A \mid B_k)\,\mathbb{P}(B_k)}{\sum_{i:\mathbb{P}(B_i)>0} \mathbb{P}(A \mid B_i) \cdot \mathbb{P}(B_i)}.\end{aligned}$$

Where the numerator comes from rearranging the definition of conditional probability and the denominator comes from Theorem 1.8. $\square$

February 11, 2024

Example 1.10 (Medical test)**.** This is the classic example for the application of how we should interpret Bayes' formula. Suppose a test detects a disease 96% of the time with a 2% chance of being a false positive (that is, a person without the disease receives a positive test). Suppose 0.5% of people carry the disease. If a random person tests positive, what is the probability that they actually carry it?

Solution. Let $D = \{\text{person carries the disease}\}$ and $A = \{\text{test is positive}\}$. The problem statement gives us that

$$\mathbb{P}(D) = 0.005, \qquad \mathbb{P}(A \mid D) = 0.96, \qquad \mathbb{P}(A \mid D^C) = 0.02.$$

Applying Bayes' formula,

$$\begin{aligned}\mathbb{P}(D \mid A) &= \frac{\mathbb{P}(A \mid D)\mathbb{P}(D)}{\mathbb{P}(A \mid D)\mathbb{P}(D) + \mathbb{P}(A \mid D^C)\mathbb{P}(D^C)} \\ &= \frac{0.96 \cdot 0.005}{0.96 \cdot 0.005 + 0.02 \cdot 0.995} \\ &\approx 0.194.\end{aligned}$$

$\square$

The takeaway here is randomly choosing people to test for a disease is not a good idea.

1.3. Independent events

Definition 1.4. $A, B \in \mathcal{F}$ are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

A more natural way (for me) to think about this is that A and B are independent if (when $\mathbb{P}(B) > 0$),

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

That is, the outcome of A does not depend on the outcome of B.

Example 1.11 (A, perhaps unintuitive, example of independence)**.** Roll a D20 (a 20-sided die) once. Let A and B be the events that the value of the roll is divisible by 4 and 5 respectively. Then since

$$\mathbb{P}(A)\mathbb{P}(B) = \frac{1}{4} \cdot \frac{1}{5} = \frac{1}{20} = \mathbb{P}(20) = \mathbb{P}(A \cap B),$$

A and B are independent.

If we replace B with the roll divisible by 6, then A and B are *not* independent:

$$\mathbb{P}(A)\,\mathbb{P}(B) = \frac{1}{4} \cdot \frac{3}{20} = \frac{3}{80} \neq \frac{1}{20} = \mathbb{P}(A \cap B).$$

Proposition 1.12. If A and B are independent, then A^* and B^* are independent, where $*$ represents either doing nothing or taking the complement.

Proof (A^C *and* B*).*

$$\mathbb{P}(A^C \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(1 - \mathbb{P}(A)). \qquad \square$$

Definition 1.5. We say that events $A_1, \ldots, A_n$ are **(mutually) independent** if for any collection $A_{i_1}, \ldots, A_{i_k}$ (where $2 \leq k \leq n$ and (i_j) is a strictly increasing multi-index of size k), we have

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Definition 1.6. The events $A_1, \ldots, A_n$ are **pairwise independent** if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $1 \leq i < j \leq n$.

Clearly (mutual) independence implies pairwise independence. The converse is false, as the next example illustrates.

Example 1.13. Flip 3 coins. Let A be the event that the first flip equals the second flip, let B be the event that the first flip equals the third flip, and let C be the event that the second flip equals the third flip.

The events themselves are probability $\frac{1}{2}$. Their pairwise intersections are probability $\frac{1}{4}$, so they are pairwise independent. However, the intersection of all 3 events has probability $\frac{1}{4}$, so they are not independent.

Proposition 1.14. If $A_1, \ldots, A_n$ are independent, then so are $A_1^*, \ldots, A_n^*$, where $*$ is either doing nothing or taking the complement.

Definition 1.7. We say the *infinite* sequence of events $\{A_i\}_{i=1}^{\infty}$ are **independent** if $A_1, \ldots, A_n$ are independent for any $n \in \mathbb{N}$.

Theorem 1.15. Let $\{A_i\}_{i=1}^{\infty}$ be independent. Let $1 \leq k_1 < k_2 < \cdots$. Suppose B_1 is created by set operations on $A_1, \ldots, A_{k_1}$, B_2 is created by set operations on $A_{k_1+1}, \ldots, A_{k_2}$, and so on for all B_n. Then the events $B_1, B_2, \ldots$ are independent.

1.3.1. Conditional independence

February 13, 2024 **Definition 1.8.** $A_1, \ldots, A_n$ are said to be **conditionally independent given** B if $\mathbb{P}(B) > 0$ and for any $1 \leq i_1 < \cdots < i_k \leq n$,

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k} \mid B) = \mathbb{P}(A_{i_1} \mid B) \cdots \mathbb{P}(A_{i_k} \mid B).$$

2. Random variables

As usual, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

We also allow $X\colon \Omega \to \mathbb{R} \cup \{\infty\}$.

Definition 2.1. A **random variable** is a map $X\colon \Omega \to \mathbb{R}$ such that $\{\omega \in \Omega \mid X(\omega) \leq C\} \in \mathcal{F}$ for all $C \in \mathbb{R}$.

Remark 2.1. Another name for a function satisfying this condition is an **($\mathcal{F}$-)measurable function**.

Remark 2.2. If Ω is discrete (i.e. $\mathcal{F} = 2^{\Omega}$), then any $X\colon \Omega \to \mathbb{R}$ is a random variable.

Definition 2.2. An **indicator variable** is a random variable defined as

$$I_B(\omega) = \begin{cases} 1 & \text{if } \omega \in B, \\ 0 & \text{if } \omega \notin B. \end{cases}$$

Since

$$\{\omega \mid I_B(\omega) \leq C\} = \begin{cases} \Omega & \text{if } c \geq 1, \\ B^C & \text{if } 0 \leq c < 1, \\ \varnothing & \text{if } c < 0, \end{cases}$$

I_B is indeed a random variable.

Example 2.3 (Non-measurable functions)**.** We have two examples for non-measurable functions in finite and infinite probability spaces.

1. Let $\Omega = \{1, 2, 3\}$, $\mathcal{F} = \{\varnothing, \{3\}, \{1, 2\}, \Omega\}$. Let $X(\omega) = \omega$. Then

   $$\{\omega \mid X(\omega) \leq 1\} = \{1\} \notin \mathcal{F},$$

   so X is non-measurable. This is a strange probability space, so this example may seem contrived. There turn out to be non-measurable functions in spaces we are used to as well.

Recall that $\mathcal{F} \neq 2^{[0,1]}$!

2. Let $\Omega = [0, 1]$ and $\mathcal{F}$ be the Borel sets on $[0, 1]$. If $B \notin \mathcal{F}$, then I_B is not a random variable.

As shorthand, we write $\{X \leq C\} \coloneqq \{\omega \mid X(\omega) \leq C\} \in \mathcal{F}$.

Example 2.4. Let $\Omega = D^2 \subseteq \mathbb{R}^2$ be the unit disc. Let $\mathcal{F}$ be the Borel sets on Ω, and let $\mathbb{P}$ be given by the area of the event. Define

$$R\colon \Omega \to [0, 1]\colon (x, y) \mapsto \sqrt{x^2 + y^2}.$$

To prove this is a random variable, note that $\{R \leq C\}$ is either $\varnothing$, Ω, or a circle of radius C. All of these are in $\mathcal{F}$.

2.1. Probability distribution of random variables

February 15, 2024

Definition 2.3. Suppose X is a random variable. The **(probability) distribution** of X on $\mathbb{R}$ is a *probability measure* μ_X *on* $\mathbb{R}$ given by

$$\begin{aligned} &\mu_X\colon \mathscr{B}\mathbb{R} \to \mathbb{R}, \\ &\qquad B \mapsto \mathbb{P}(X \in B) \coloneqq \mathbb{P}\left(X^{-1}(B)\right) = \mathbb{P}\left(\omega \in \Omega \mid X(\omega) \in B\right), \end{aligned}$$

for all Borel sets B on $\mathbb{R}$.

Example 2.5. Roll a die infinitely many times. Let N be the random variable representing the roll which we first get six. What is the distribution of N?

$\mathbb{P}(X = k)$ is the same as $\mathbb{P}(X^{-1}\{k\})$.

Solution. In the context of the question, it only really makes sense to look at when $X = k$ for some $k \in \mathbb{N} \cup \{\infty\}$. Computing this values, we get

$$\begin{aligned}\mu_N(k) &= \mathbb{P}(X \in \{k\}) = \mathbb{P}(X = k)\\ &= \left(\frac{5}{6}\right)^{k-1}\left(\frac{1}{6}\right).\\ \mu_N(\infty) &= \mathbb{P}(\text{no } 6)\\ &= 0.\end{aligned}$$

□

Definition 2.4. We say a random variable X is **discrete** if there exists a finite or countable set $B \subseteq \mathbb{R}$ such that $\mu_X(B) = 1$.

Values $y \in B$ such that $\mu_X(X = y) > 0$ are called **possible values**, so X being discrete is the same as it having finitely or countably many possible values.

If Ω is discrete, then X is discrete, but X being discrete *does not* mean that Ω is countable, since there could be uncountably many values where a random variable has probability zero.

Definition 2.5. The **probability mass function (PMF)** of a *discrete* random variable X is a function f_X such that

$$f_X(y) = \mathbb{P}(X = y)$$

for all possible values y of X.

Example 2.6. With the previous die example, $f_N(k) = \left(\frac{5}{6}\right)^{k-1}\left(\frac{1}{6}\right)$.

Note that

$$\mu_X(B) = \sum_{\substack{y \in B\\ y \text{ possible value}}} \mathbb{P}(X = y) = \sum_{\substack{y \in B\\ y \text{ possible value}}} f_X(y).$$

Example 2.7 (Examples of common distributions)**.**

1. Let $p \in [0, 1]$. We say that X has a **Bernoulli distribution** with parameter p if the PMF of X is $f_X(0) = p$ and $f_X(1) = 1 - p$. We write $X \sim \text{Ber}(p)$.
2. Let $p \in [0, 1]$. We say that X has a **geometric distribution** with parameter p if the PMF of X is $f_X(k) = (1-p)^{k-1}p$. We write $X \sim \text{Geom}(p)$.
3. Let n be a positive integer. X has a **binomial distribution** with parameters n, p if the PMF of X is
$$f_X(k) = \binom{n}{k}p^k(1-p)^{n-k}, \quad k = 0, \ldots, n.$$
We write $X \sim \text{Binom}(n, p)$.

Definition 2.6. The **cumulative density function (CDF)** of a random variable X (not necessarily discrete!) is given by

$$F_X(s) = \mathbb{P}(X \leq s).$$

Example 2.8. If $X \sim \text{Ber}(p)$, then

$$F_X(s) = \mathbb{P}(X \leq s) = \begin{cases} 0 & \text{if } s < 0,\\ p & \text{if } 0 \leq s < 1,\\ 1 & \text{ifs} \geq 1.\end{cases}$$

In general, if X is a discrete random variable,

$$F_X(s) = \mathbb{P}(X \le s) = \sum_{\substack{y \le C \\ y \text{ possible}}} f_X(y).$$

Proposition 2.9 (Properties of a CDF)**.** Let F_X be a cumulative density function.

0. $F_X(s) \in [0,1]$
1. $F_X(s)$ is a monotone increasing function; i.e., $s_1 \le s_2 \implies F_X(s_1) \le F_X(s_2)$.
2. $F_X(s)$ is a **right-continuous** function; i.e. $\lim_{s \downarrow t} F_X(s) = F_X(t)$.
3. $\lim_{t\to\infty} F_X(t) = 1$, $\lim_{t\to-\infty} F_X(t) = 0$.

Proof of (2). Let (s_n) be a sequence converging to t from above. Then let $A_n = \{X \le s_n\}$. So $A_1 \supseteq A_2 \supseteq \cdots$, $\bigcap_{n=1}^{\infty} A_n = A = \{X \le t\}$. Hence, $\mathbb{P}(A_n) \xrightarrow{n\to\infty} \mathbb{P}(A)$. □

Theorem 2.10 (Existence of CDF on random variable)**.** If $F\colon \mathbb{R} \to [0,1]$ satisfies (1)-(3), then there exists a probability space and random variable X on it such that $F_X = F$.

The important takeaway is that a CDF works in the place of a non-discrete random variable.

Proposition 2.11.

1. $\lim_{s \uparrow t} F_X(s) = \mathbb{P}(X < t)$.
2. Define the above value to be $F_X(t^-)$. Then

$$\mathbb{P}(X = t) = \mathbb{P}(X \le t) - \mathbb{P}(X < t) = F_X(t) - F_X(t^-).$$

February 20, 2024

Definition 2.7. A random variable X is **absolutely continuous** if there exists an integrable, non-negative function $f\colon \mathbb{R} \to \mathbb{R}_+$ such that

$$\underbrace{F_X}_{\text{CDF}}(t) = \int_{-\infty}^{t} f(s)\,ds.$$

f is called the **probability density function (PDF)** of X.

Remark 2.12. If f is continuous, then $f(t) = \frac{dF_X}{ds}(t)$

Additionally, we can evaluate the CDF at infinity:

$$F_X(\infty) = \mathbb{P}(X < \infty) = \int_{-\infty}^{\infty} f(s)\,ds = 1.$$

Example 2.13. Let X be the random variable on the probability space of uniformly at random picking a point in $[0,1]$, given by $X(\omega) = \omega$. So

$$\mathbb{P}(X \le t) = F_X(t) = \begin{cases} 0 & t \le 0 \\ 1 & t \ge 1 \\ t & 0 \le t \le 1 \end{cases}.$$

So

$$F_X'(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

Note that the PDF is *never* unique, since it can take any value at finitely many points without affecting the CDF.

Theorem 2.14. If a CDF of X is continuous and has derivative at all but countably many points, then X is absolutely continuous and the PDF of X is equal to $p(t) = F'_X(t)$ (when $F'_X(t)$ is not defined, it can take on any value).

Remark 2.15. If a CDF is *not* continuous, then X is not absolutely continuous.

We can calculate the probability distribution on any Borel set $B \in \mathscr{B}\mathbb{R}$ by using an indicator function as follows:

$$\mu_X(B) = \int_B f(t)dt = \int_{-\infty}^{\infty} f(t)I_B(t)\,dt.$$

Proposition 2.16. If $f\colon \mathbb{R} \to \mathbb{R}_+$ is integrable and $\int_{-\infty}^{\infty} f(t)\,dt = 1$, then f is a PDF for some random variable X.

Proof. Define the function $F_X\colon \mathbb{R} \to \mathbb{R}_+$ as

$$F_X(t) = \int_{-\infty}^{t} f(s)\,ds$$

is continuous. Note that $f \geq 0$, so

$$\int_{-\infty}^{t} f(s)\,ds \leq \int_{-\infty}^{r} f(s)\,ds$$

if $t \leq r$. Therefore, F_X is monotone increasing. Finally, $\lim_{t\to\infty} F_X(t) = 1$ and $\lim_{t\to-\infty} F_X(t) = 0$. Therefore, we can apply Theorem 2.10 to get a random variable X. □

Remark 2.17 (Probability density $\neq$ probability mass)**.** It's natural to assume that if $f(t)$ is a PDF, then $f(t)$ is $\mathbb{P}(X = t)$. However, this is not the case. If f is continuous at t, then

$$\mathbb{P}(t-\varepsilon < X < t+\varepsilon) = \int_{t-\varepsilon}^{t+\varepsilon} f(s)\,ds \approx f(t)2\varepsilon.$$

So $\mathbb{P}(X \in \text{small interval}) = f(t) \cdot \text{length(small interval)}$. Hence, a PDF is not necessarily a PMF!

Example 2.18 (A hybrid function)**.** Consider the CDF

$$F(t) = \begin{cases} 0 & t \leq 0, \\ \frac{t}{2} & 0 \leq t < 1, \\ 1 & t \geq 1. \end{cases}$$

This is neither discrete nor absolutely continuous.

Remark 2.19. F being continuous does not imply that F is absolutely continuous. A common counterexample is the Cantor function.

In summary:

If a random variable X has a...	Discrete	Absolutely Continuous
PMF (p_X / f_X)	Yes	No
PDF (f_X)	No	Yes
CDF (F_X)	Yes	Yes

2.2. Random vectors

Definition 2.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A d-dimensional **random vector** on this space is a measurable function

$$\mathbf{X}\colon \Omega \to \mathbb{R}^d.$$

This means that for $B \in \mathscr{B}\mathbb{R}^d$, $\{X \in B\} \in \mathcal{F}$.

Equivalently, a d-dimensional random vector $\mathbf{X}$ is a tuple/vector $(X_1, \ldots, X_d)$ of random variables $X_i\colon \Omega \to \mathbb{R}$.

Definition 2.9. A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is **discrete** if all X_i are discrete. The PMF of X is

$$p_X(y_1, \ldots, y_d) = \mathbb{P}(X_1 = y_1, \ldots, X_d = y_d).$$

Example 2.20. Consider a fair coin that lands 1 or 2. Let R_i be the result of the ith flip. Let

$$\mathbf{X} = \begin{bmatrix} R_1 + R_2 \\ R_1 R_2 \end{bmatrix}.$$

Then we have

$$\mathbb{P}\left(\mathbf{X} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) = \frac{1}{4}, \quad \mathbb{P}\left(\mathbf{X} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}\right) = \frac{1}{2}, \quad \mathbb{P}\left(\mathbf{X} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}\right) = \frac{1}{4}.$$

There is a way to recover the PMF of $X_1, \ldots, X_n$ from $\mathbf{X}$. We call this the **marginal PMF/distribution** of $\mathbf{X}$.

Proposition 2.21. If $\mathbf{X} = (X_1, \ldots, X_d)$ is a discrete random vector, then

$$f_{X_i}(\overline{y}_i) = \sum_{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d} p_{\mathbf{X}}(y_1, \ldots, \overline{y_i}, \ldots, y_d).$$

In other words, the value of the marginal PMF p_{X_i} at $\overline{y}_i$ is recovered by summing over all (possible) values in the PMF $p_{\mathbf{X}}$ with $\overline{y}_i$ in the ith entry.

February 22, 2024

Example 2.22. Let $\mathbf{X} = (X_1, X_2)$ be a 2-dimensional random vector given by

$$X_1 = \begin{cases} 1 & \text{probability } \frac{1}{2} \\ -1 & \text{probability } \frac{1}{2} \end{cases}, \qquad X_2 = \begin{cases} 1 & \text{probability } \frac{1}{2} \\ 2 & \text{probability } \frac{1}{2} \end{cases}.$$

Then

$$\mu_X([0,1] \times [0,1]) = \mathbb{P}(\mathbf{X} = (1,1)).$$

Definition 2.10. Let $\mathbf{X}$ be a d-dimensional random vector. The **joint CDF** of $X_1, \ldots, X_d$ (or the CDF of $\mathbf{X}$) is

$$F_X(t_1, \ldots, t_d) = \mathbb{P}(X_1 \le t_1, \ldots, X_d \le t_d).$$

Definition 2.11. Let $\mathbf{X}$ be a d-dimensional random vector. $X_1, \ldots, X_d$ are **jointly absolutely continuous** (or $\mathbf{X}$ is absolutely continuous) if there exists an integrable function $f_{\mathbf{X}}\colon \mathbb{R}^d \to \mathbb{R}_+$ such that

$$F_{\mathbf{X}}(t_1, \ldots, t_d) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(s_1, \ldots, s_d) ds_1 \cdots ds_d.$$

Remark 2.23. Again, if $f_{\mathbf{X}}$ is continuous, then

$$f_{\mathbf{X}}(s_1, \ldots, s_d) = \frac{\partial^d \left(F_{\mathbf{X}}(s_1, \ldots, s_d)\right)}{\partial s_1 \cdots \partial s_d}.$$

The distribution is given by

$$\mu_{\mathbf{X}}(B) = \mathbb{P}(X \in B) = \int_B f_{\mathbf{X}}(s_1, \ldots, s_d)\, ds_1 \cdots ds_d = \int_{\mathbb{R}^d} I_B(s_1, \ldots, s_d) f_{\mathbf{X}}(s_1, \ldots, s_d)\, ds_1 \cdots ds_d.$$

Moreover,

$$\int_{\mathbb{R}^d} f_{\mathbf{X}}(s_1, \ldots, s_d)\, ds_1 \cdots ds_d = 1.$$

Theorem 2.24. If a non-negative function f is integrable and $\int_{\mathbb{R}^d} f = 1$, then f is the PDF of some random vector $\mathbf{X}$.

Proposition 2.25. Suppose $\mathbf{X}$ is an absolutely continuous random vector with PDF $f_{\mathbf{X}}$. Then the marginal PDF is given by

$$f_{X_i}(t_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(s_1, \ldots, s_{i-1}, t_i, s_{i+1}, \ldots, s_d)\, ds_1 \cdots ds_{i-1} ds_{i+1} ds_d.$$

Proof for $d = 2$. Let (X, Y) be a random vector with PDF $f_{X,Y}$. Then

$$\begin{aligned}
\mathbb{P}(X \leq t) &= \mathbb{P}(X \leq t, Y \leq \infty) \\
&= \mathbb{P}((X, Y) \in (-\infty, t) \times \mathbb{R}) \\
&= \int_{-\infty}^{t} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy \\
&= \int_{-\infty}^{t} \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \right) dx
\end{aligned}$$

The inner integral $\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$ gives us the PDF of X by definition. □

Example 2.26. Let $\mathbf{X}$ be a random vector that represents uniformly choosing a point at random on the unit disc $D^2 \subseteq \mathbb{R}^2$. Then

$$f(x, y) = \frac{1}{\pi} I_{D^2}(x, y).$$

Suppose we want the PDF of X_1. We get that

$$\begin{aligned}
p_{X_1}(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy \\
&= \int_{-\infty}^{\infty} \frac{1}{\pi} I_{D^2}\, dy \\
&= \begin{cases} 0 & |x| \geq 1 \\ \frac{1}{\pi} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} 1\, dy & \text{otherwise.} \end{cases} \\
&= \begin{cases} 0 & |x| \geq 1 \\ \frac{2\sqrt{1-x^2}}{\pi} & |x| < 1. \end{cases}
\end{aligned}$$

We call this the **semicircle distribution**.

Example 2.27. An absolutely continuous distribution on $D \subseteq \mathbb{R}^d$ has PDF

$$p(\mathbf{x}) = \frac{1}{\operatorname{vol}(D)} I_D(\mathbf{x}).$$

If $\mathbf{X}$ is an absolutely continuous vector, then all $X_1, \ldots, X_d$ are absolutely continuous. The converse is not true.

Example 2.28. Let X be an absolutely continuous random variable. Suppose

$$Y(\omega) = X(\omega), \qquad \forall \omega \in \Omega.$$

So Y is absolutely continuous. But (X, Y) is *not* absolutely continuous since $\mathbb{P}(X = Y) = 1$, but the probability of anything in this set in $\mathbb{R}^2$ is 0.

2.3. Notions of equality

Definition 2.12. Two random variables (or vectors) X and Y over the same probability space are equal **almost everywhere (almost surely)**, denoted $X = Y$ **a.s.**, if they disagree on a measure zero set; that is,

$$\mathbb{P}(X \neq Y) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \neq Y(\omega)\}) = 0.$$

Equivalently,

$$\mathbb{P}(X = Y) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = Y(\omega)\}) = 1.$$

Definition 2.13. Suppose X and Y are random variables (or vectors) (not necessarily over the same probability space!). We say that X and Y are **equal in distribution** $X \overset{d}{=} Y$ if $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ for all Borel sets B.

Example 2.29. Let X represent flipping a fair coin (where tails is 0 and heads is 1) and Y represent either choosing 0 or 1 with probability $\frac{1}{2}$. Then $X \overset{d}{=} Y$.

February 27, 2024

Proposition 2.30. If X, Y are random variables on the same space and $X = Y$ a.s., then $X \overset{d}{=} Y$.

Proof. Notice that $\mathbb{P}(X \neq Y) = 0$, so any subset of $\{X \neq Y\}$ also has probability zero.

$$\begin{aligned}\mathbb{P}(X \in B) &= \mathbb{P}(X \in B, X = Y) + \mathbb{P}(X \in B, X \neq Y)\\ &= \mathbb{P}(Y \in B, X = Y) + 0\\ &= \mathbb{P}(Y \in B, X = Y) + \mathbb{P}(Y \in B, X \neq Y)\\ &= \mathbb{P}(Y \in B).\end{aligned}$$

□

Proposition 2.31. Let $\mathbf{X}$, $\mathbf{Y}$ be d-dimensional random vectors. Then

1. $X_i = Y_i$ a.s. for all $1 \leq i \leq d$ implies $\mathbf{X} = \mathbf{Y}$ a.s.
2. $X_i \overset{d}{=} Y_i$ does not necessarily imply $\mathbf{X} \overset{d}{=} \mathbf{Y}$.

Proof. (1)

$$\mathbb{P}(\mathbf{X} \neq \mathbf{Y}) = \mathbb{P}\left(\bigcup_{i=1}^{d} \{X_i \neq Y_i\}\right) \leq \sum_{i=1}^{d} \mathbb{P}(X_i \neq Y_i) = 0.$$

(2) Consider the probability space of flipping a pair of fair coins. Define

$$\begin{aligned}\mathbf{X} &= (X_1, X_2)\colon (\omega_1, \omega_2) \mapsto (\omega_1, \omega_2),\\ \mathbf{Y} &= (Y_1, Y_2)\colon (\omega_1, \omega_2) \mapsto (\omega_2, \omega_1).\end{aligned}$$

All X_i, Y_i $(i = 1, 2)$ are $\mathrm{Ber}\left(\frac{1}{2}\right)$ random variables, but

$$\mathbb{P}(\mathbf{X} \in \{(0,0),(1,1)\}) = \frac{1}{2}, \qquad \mathbb{P}(\mathbf{Y} \in \{(0,0),(1,1)\}) = 1.$$

□

Both of these propositions indicate that equality in distribution is a weaker condition that almost sure equality.

2.4. Functions of random variables

Let X be a random variable and $g\colon \mathbb{R} \to \mathbb{R}$ be measurable (that is, $\{x \mid g(x) \leq C\}$ is a Borel set). Let $Y = g(X)$. Then the distribution of Y is

$$\mathbb{P}(Y \in B) = \mathbb{P}(g(X) \in B) = \mathbb{P}\left(X \in g^{-1}(B)\right) = \mu_X(g^{-1}(B)),$$

g is not necessarily invertible, so $g^{-1}(B) = \{x \mid g(x) \in B\}$ is the *preimage* of B.

Example 2.32. Suppose $Y = (X-2)^2$. Express the CDF of Y in terms of the CDF of X.

Solution. If $s < 0$, then $F_Y(s) = 0$. If $s = 0$, then $F_Y(s) = \mathbb{P}(X = 2)$. What remains to check is when $s > 0$. We have

$$\begin{aligned} F_Y(s) &= \mathbb{P}\left((X-2)^2 \le s\right) \\ &= \mathbb{P}\left(-\sqrt{s} \le X - s \le \sqrt{s}\right) \\ &= \mathbb{P}\left(2-\sqrt{s} \le X \le 2+\sqrt{s}\right) \\ &= F_X(\sqrt{5}+2) - F_X((2-\sqrt{5})^-). \end{aligned}$$

i.e. the left limit

Since X is absolutely continuous, the left limit is equal to the limit, and $\mathbb{P}(X = 2) = 0$, so this equation becomes nicer:

$$F_X(s) = \begin{cases} 0 & s \le 0, \\ F_X(\sqrt{5}+2) - F_X(2-\sqrt{5}) & s > 0. \end{cases}$$

□

2.4.1. Functions of discrete variables

If X discrete, then $Y = g(X)$ is also discrete (its possible values are $g(x_i)$). So

$$\mathbb{P}(Y = y_i) = \sum_{x \in g^{-1}(y_i)} \mathbb{P}(X = x).$$

Problem 2.1. Uniformly choose a value in $\Omega = \{-2, -1, 0, 1, 2\}$. Let $X(\omega) = \omega$, and $Y = X^2$. Compute the PMF of X and Y.

2.4.2. Continuous variables with discontinuous functions

Example 2.33. Let X be a random variable with CDF

$$F_X(x) = \begin{cases} \frac{1}{x^2} & x \ge 1, \\ 0 & \text{otherwise.} \end{cases}$$

Define $Y = \lfloor X \rfloor$, which is not continuous. We compute the PMF for $k = 1, 2, \ldots$ as

$$\mathbb{P}(Y = k) = \mathbb{P}(\lfloor X \rfloor = k) = \mathbb{P}(k \le X \le k+1) = \int_k^{k+1} \frac{1}{x^2}\,dx = \frac{1}{k(k+1)}.$$

2.5. Change of variables

This section covers some technical facts not from class about functions of two random variables. Consider the jointly continuous random variables X and Y with joint probability density function $f_{X,Y}$. Let another pair of random variables (U, V) be defined as functions of (X, Y) as

$$U = g(X, Y), \quad V = h(X, Y).$$

We can find $f_{U,V}$ through a change of variables. Suppose K is a region such that $f_{X,Y}(x, y) = 0$ outside K. This implies $\mathbb{P}((X, Y) \in K) = 1$. Define

$$G(x, y) = (g(x, y), h(x, y)).$$

This is a bijective function from K onto its image (call it L). Denote $G^{-1}(u, v) = (q(u, v), r(u, v))$. Assume further that

(i) q and r has continuous partial derivatives $\frac{\partial q}{\partial u}$, $\frac{\partial q}{\partial v}$, $\frac{\partial r}{\partial u}$, and $\frac{\partial r}{\partial v}$ are continuous on L.

(ii) The Jacobian

$$\operatorname{Jac}(u,v) = \det \begin{bmatrix} \frac{\partial q}{\partial u}(u,v) & \frac{\partial q}{\partial v}(u,v) \\ \frac{\partial r}{\partial u}(u,v) & \frac{\partial r}{\partial v}(u,v) \end{bmatrix} \neq 0$$

on L.

Theorem 2.34. Under the above assumptions, the joint PDF of (U,V) is given by

$$f_{U,V}(u,v) = f_{X,Y}(q(u,v), r(u,v)) \, |\operatorname{Jac}(u,v)|$$

for $(u,v) \in L$ and is 0 for (u,v) outside L.

This can be generalized to more dimensions with the Jacobian of higher-dimensional functions. Notice that this is effectively just change of variables from multivariable calculus.

Example 2.35. Let (X,Y) be a uniformly chosen point on the unit disc D^2. Let (R,Θ) be the polar coordinates of the chosen point. What is the joint PDF of (R,Θ)?

Solution. We have $x = r\cos\theta$, $y = r\sin\theta$. The Jacobian is

$$\operatorname{Jac}(r,\theta) = \det \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = r\cos^2\theta + r\sin^2\theta = r.$$

The joint PDF of (X,Y) is given by $\frac{1}{\pi}$ in D^2 and 0 outside. Then

$$f_{R,\Theta}(r,\theta) = \begin{cases} \frac{1}{\pi} r & (r,\theta) \in D^2, \\ 0 & \text{otherwise.} \end{cases}$$

□

3. Independence of random variables

Recall that the events $A_1, \dots, A_n$ are independent if for all $1 \le i_1 < \cdots < i_k \le n$,

$$\mathbb{P}\left(A_{i_1} \cdots A_{i_k}\right) = \mathbb{P}\left(A_{i_1}\right) \cdots \mathbb{P}\left(A_{i_k}\right).$$

Definition 3.1. Let $X_1, \dots, X_n$ be random variables on some probability space. Then $X_1, \dots, X_n$ are **independent** if

$$\mathbb{P}\left(X_1 \in B_1, \dots, X_n \in B_n\right) = \mathbb{P}\left(X_1 \in B_1\right) \cdots \mathbb{P}\left(X_n \in B_n\right)$$

for all Borel sets $B_1, \dots, B_n \in \mathscr{B}\mathbb{R}$.

We avoid checking every $1 \le i_1 < \cdots < i_k \le n$ by setting some $B_i = \mathbb{R}$.

If $\{X_i\}_{i=1}^{\infty}$ is a countable collection of random variables, we say that $\{X_i\}$ are **independent** if $X_1, \dots, X_n$ are independent for all $n \in \mathbb{N}$.

Remark 3.1. It is hard to check *all* Borel sets. Luckily, it suffices to check for all $B_i = (-\infty, c_i]$ for $c_i \in \mathbb{R}$, $i = 1, \dots, n$. In other words,

$$\mathbb{P}\left(X_1 \le c_1, \dots, X_n \le c_n\right) = \mathbb{P}\left(X \le c_1\right) \cdots \mathbb{P}\left(X \le c_n\right)$$

or

$$F_{\mathbf{X}}(c_1, \dots, c_n) = F_{X_1}(c_1) \cdots F_{X_n}(c_n).$$

Example 3.2. If $X_1, \dots, X_n$ are discrete random variables, $X_1, \dots, X_n$ are independent $\iff$

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \mathbb{P}\left(X_1 = x_1\right) \cdots \mathbb{P}\left(X_n = x_n\right)$$

for all possible values $x_1, \dots, x_n$.

Proof. ($\impliedby$) is obvious. ($\implies$, $n = 2$) check the definition:

$$\mathbb{P}\left(X_1 \in B_1, X_2 \in B_2\right) = \sum_{\substack{x_i \in B_i \\ x_i \text{ possible}}} \mathbb{P}\left((X_1, X_2) = (x_1, x_2)\right) = \sum_{\substack{x_i \in B_i \\ x_i \text{ possible}}} \mathbb{P}\left(X_1 = x_1\right) \mathbb{P}\left(X_2 = x_2\right).$$

□

Example 3.3. Flip a fair coin and let $X_i = \begin{cases} 1 & i\text{th coin is H,} \\ 0 & \text{otherwise.} \end{cases}$ (so $X_i \sim \text{Ber}\left(\frac{1}{2}\right)$). Then

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \frac{1}{2^n} = \mathbb{P}\left(X_1 = x_1\right) \cdots \mathbb{P}\left(X_n = x_n\right).$$

February 29, 2024

Remark 3.4. If we can write $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = g_1(x_1) \cdots g_n(x_n)$ for some functions $g_1, \dots, g_n$, then $X_1, \dots, X_n$ are independent, and there are constants $c_1, \dots, c_n$ such that $\prod_{j=1}^{n} c_j = 1$, and $g_j(x) = c_j \mathbb{P}\left(X_j = x\right)$ for all x. This method is somewhat "by inspection," by looking at the PMF and seeing if we can factorize it into an x and y term.

Example 3.5. Let $X, Y \colon \mathbb{N} \to [0, 1]$ be random variables such that

$$\mathbb{P}\left(X = k, Y = \ell\right) = \frac{1}{3^{k-1} 2^{2\ell - 1}}.$$

Therefore,

$$\mathbb{P}\left(X = k\right) = \sum_{\ell=1}^{\infty} \mathbb{P}\left(X = k, Y = \ell\right) = \sum_{\ell=1}^{\infty} \frac{1}{3^{k-1} 2^{2\ell-1}} = \frac{2}{3^k}.$$

$$\mathbb{P}\left(Y = \ell\right) = \sum_{k=1}^{\infty} \mathbb{P}\left(X = k, Y = \ell\right) = \sum_{k=1}^{\infty} \frac{1}{3^{k-1} 2^{2\ell-1}} = \frac{3}{2^{2\ell}}.$$

Taking the product of these two shows us that X and Y are independent. We could have also immediately noticed by inspection that $\frac{1}{3^{k-1} 2^{2\ell-1}} = \frac{1}{3^{k-1}} \cdot \frac{1}{2^{2\ell-1}}$.

3.1. Absolutely continuous independence

Theorem 3.6. Let $X_1, \ldots, X_n$ be jointly absolutely continuous. Then $X_1, \ldots, X_n$ are independent $\iff p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$.

Definition 3.2. Let $X_1, \ldots, X_n$ be random variables on the same space. We say $X_1, \ldots, X_n$ are **independent identically distributed (i.i.d.)** if they are (1) independent and (2) all X_i have the same distribution.

3.2. Functions of independent random variables

Theorem 3.7. Let $X_1, \ldots, X_n$ be independent random variables. Let g_i be measurable functions. Then $g_1(X_1), \ldots, g_n(X_n)$ are also independent.

Proof.

$$\begin{aligned}\mathbb{P}(g_1(X_1) \in B_1, \ldots, g_n(X_n) \in B_n) &= \mathbb{P}\left(g_1^{-1}(B_1), \ldots, g_n^{-1}(B_n)\right) \\ &= \mathbb{P}\left(g_1^{-1}(B_1)\right) \cdots \mathbb{P}\left(g_n^{-1}(B_n)\right) \quad (g_i^{-1}(B_i) \text{ are Borel}) \\ &= \mathbb{P}(g_1(X_1) \in B_1) \cdots \mathbb{P}(g_n(X_n) \in B_n).\end{aligned}$$

□

Theorem 3.8. Let $\{X_i\}_{i=1}^{\infty}$ be independent. Then if Y_1 is a measurable function of $X_1, \ldots, X_{k_1}$, Y_2 is a measurable function of $X_{k_1+1}, \ldots, X_{k_2}$, etc., then $Y_1, Y_2, \ldots$ are independent as well.

3.3. Random trials and some named distributions

March 05, 2024 Many of our named distributions aries from considering series of random trials. That is, picking $X_1, X_2, \ldots$ i.i.d. and considering random variables formed out of them.

For example, if $X_1, X_2, \ldots, \sim \text{Ber}(p)$ are independent, then the distribution of the first success $N = \min_n \{X_n = 1\}$ is $\text{Geom}(p)$. Another example is that the number of successes in the first n trials is $\text{Binom}(n, p)$.

Problem 3.1. What is the probability distribution of the *second* success (call this N_2 for now)? What about the rth success (N_r)?

Solution. For $r = 2$, the possible values are $k = 2, 3, \ldots$. One of the first $k-1$ trials will be a success. There are $\binom{k-1}{1}$ ways to choose it. The probability for each event is $(1-p)^{k-2}p^2$. Hence, $\mathbb{P}(N_2 = k) = \binom{k-1}{1}(1-p)^{k-2}p^2$.

It's clear that $\mathbb{P}(N_r = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$. □

We call this distribution the **negative binomial distribution**: $\text{Negbin}(r, p)$. As a sanity check, $\text{Negbin}(1, p) = \text{Binom}(p)$.

Consider counting rare events. If we have $X_n \sim \text{Binom}(n, \frac{\lambda}{n})$, then as n grows, we have more trials, but a smaller chance of success. We compute (for possible k: $0, 1, 2, \ldots, n$),

$$\mathbb{P}(X_n = k) = \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Problem 3.2. What is the limit as $n \to \infty$? Verify this is a distribution.

Solution.

$$\lim_{n\to\infty} \frac{n\cdots(n-k+1)}{(k\cdots 2\cdot 1)n^k}\left(1-\frac{\lambda}{n}\right)^{n-k}\lambda^k = \frac{e^{-\lambda}\lambda^k}{k!}.$$

To check this is indeed a distribution, we need to check the probabilities sum to 1. Indeed,

$$\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1. \qquad \square$$

We call this the **Poisson distribution**: Poisson(λ).

Now suppose we are waiting for a rare event to happen. Let $T_n \sim \text{Geom}(\frac{\lambda}{n})$ and consider the random variable $\frac{T_n}{n}$. The possible values of T_n/n will get dense in $\mathbb{R}$ as $n \to \infty$, so we would assume this becomes a continuous random variable in the limit. Let's look at the CDF:

$$\mathbb{P}(T_n/n \le x) = 1 - \mathbb{P}(T_n/n > x) = 1 - \mathbb{P}(T_n > nx).$$

Problem 3.3. Fix λ. What is the limit of $\mathbb{P}(T_n/n \le x)$ as $n \to \infty$?

Solution.

$$\lim_{n\to\infty}\mathbb{P}(T_n > nx) = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{\lfloor nx\rfloor} = e^{-\lambda x}.$$

So

$$\lim_{n\to\infty}\mathbb{P}(T_n/n > x) = 1 - e^{-\lambda x} \qquad \square$$

$T = T_n/n$ is absolutely continuous so we can compute its PDF:

$$f_T(x) = \begin{cases}\lambda e^{-\lambda x} & \text{if } x > 0\\ 0 & \text{otherwise.}\end{cases}$$

We call this the **exponential distribution**: Exp(λ).

Proposition 3.9. Exp(λ) has a **memoryless property**. That is, given $t, s > 0$,

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

Imagine T measures the life of something. This says that past t time units have no influence on its survival for the next s time units.

Proof. The RHS is $e^{-\lambda s}$. The LHS is

$$\frac{\mathbb{P}(T > t+s, T > t)}{\mathbb{P}(T > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s}. \qquad \square$$

In fact, this is the *only* continuous distribution with this property. Indeed, suppose $G(t) = \mathbb{P}(T > t)$ is memoryless. Then

$$\frac{G(t+s)}{G(t)} = G(s) \implies G(t+s) = G(t)G(s).$$

Taking the log,

$$\log(G(t+s)) = \log(G(t)) + \log(G(s)).$$

This is a Cauchy function equation. It has a *unique continuous solution* and some evil non-continuous solutions.

3.3.1. Trials with multiple outcomes

Suppose we get $i = 1, \dots, r$ with probability p_i (such that $\sum_{i=1}^{r} p_i = 1$). Let X_d be the number of trials with outcome d. Then

$$X_d \sim \text{Binom}(n, p_d).$$

We can look at the joint distribution $X_1, \dots, X_r$. We call this a **multinomial distribution**: $\text{Multinom}(n, r, p_1, \dots, p_r)$. The probability of seeing event j, a_j times ($\sum_{j=1}^{r} a_j = n$) is

$$\begin{aligned}\mathbb{P}(X_1 = a_1, \dots, X_r = a_r) &= \binom{n}{a_1, \dots, a_r} p_1^{a_1} \cdots p_r^{a_r} \\ &= \frac{n!}{a_1! \cdots a_r!} p_1^{a_1} \cdots p_r^{a_r}\end{aligned}$$

3.4. Distribution of sum of independent random variables

March 07, 2024

Let X, Y be random variables. A simple operation to take on these variables is addition. What is the distribution of $X + Y$?

If X, Y are discrete, then $X + Y$ is discrete. If we further suppose that X, Y are independent, we can compute

$$\begin{aligned}\mathbb{P}(X + Y = n) &= \sum_a \mathbb{P}(X = a, Y = n - a) \\ &= \sum_a \mathbb{P}(X = a)\,\mathbb{P}(Y = n - a) \\ &= \sum_a p_X(a) p_Y(n - a).\end{aligned}$$

This sum is a **discrete convolution** of p_X and p_Y. Another way to write this is

$$p_X * p_Y(u) = \sum_a p_X(a) p_Y(u - a).$$

Notice that this sum only needs to be taken over a where a is a possible value of X and $u - a$ is a possible value of Y.

If X and Y absolutely continuous and independent then $X + Y$ is absolutely continuous (on the other hand, if we drop the independent assumption, we could consider X and $-X$, whose only possible value is 0).

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

To derive this, consider the CDF:

$$\begin{aligned}\mathbb{P}(X + Y \le z) &= \iint_{\substack{(x,y) \\ x+y \le z}} f(x, y)\,dx\,dy = \iint_{\substack{(x,y) \\ x+y \le z}} f_X(x) f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{z} f_X(x) f_Y(u - x)\,du.\end{aligned}$$

We define the **(continuous) convolution** of f_X and f_Y as

$$f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\,dx.$$

Example 3.10. Let $X, Y \sim \text{Exp}(\lambda)$. Then

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(t) f_Y(z-t)\,dt \\ &= \int_0^z \lambda e^{-\lambda t} \lambda e^{-\lambda(z-t)}\,dt \\ &= \lambda^2 \int_0^z e^{-\lambda z}\,dt \\ &= \lambda^2 z e^{-\lambda z}. \end{aligned}$$

In general, if $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ are i.i.d., then

$$f_{X_1+\cdots+X_n}(z) = \begin{cases} \frac{\lambda^n z^{n-1}}{(n-1)!} e^{-\lambda z} & z > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We call this distribution the **gamma distribution**, and denote it $\text{Gamma}(n, \lambda)$. You can extend the definition to $n \in \mathbb{R}$ with the actual Γ function.

3.5. Exchangeability

If we draw from a deck without replacement, asking for the probability of the 10th card being an ace and the 37th draw being a king should be the same as asking if the first draw is an ace and the second draw is a king.

Definition 3.3. The random variables $X_1, \ldots, X_n$ with joint CDF F are **exchangeable** if for any permutation $\sigma\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$,

$$F(x_1, \ldots, x_n) = F(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

If $X_1, \ldots, X_n$ are discrete, then we just need

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}\left(X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)}\right).$$

If $X_1, \ldots, X_n$ are jointly absolutely continuous with PDF f, then we need

$$f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

Example 3.11. I.i.d. random variables are exchangeable.

Example 3.12 (Sampling without replacement)**.** Let $(X_1, \ldots, X_n)$ sample without replacement from $\{1, \ldots, n\}$. Then

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \begin{cases} 0 & \text{if } (x_1, \ldots, x_n) \text{ not a permutation of } \{1, \ldots, n\} \\ \frac{1}{n!} & \text{if } (x_1, \ldots, x_n) \text{ is a permutation of } \{1, \ldots, n\}, \end{cases}$$

which implies $X_1, \ldots, X_n$ are exchangeable.

Proposition 3.13. If $X_1, \ldots, X_n$ are exchangeable, then for any distinct $k_1, \ldots, k_m$, the joint distributions of $X_1, \ldots, X_m$ and $X_{k_1}, \ldots, X_{k_m}$ are the same.

Proposition 3.14. Let $X_1, \ldots, X_n$ be exchangeable. Then for a measurable function g, $g(X_1), \ldots, g(X_n)$ are exchangeable.

Example 3.15. Let $X_1, \ldots, X_8$ be i.i.d. random variables from $\text{Unif}[1, 2]$. What is the probability that the largest variable is X_4?

Solution. At least one of the variables is the largest (equality happens with probability zero), so

$$\sum_{i=1}^{8} \mathbb{P}\left(X_i \text{ is the largest}\right) = 1.$$

Since X_i are i.i.d., and hence exchangeable,

$$\mathbb{P}\left(X_i \text{ is the largest}\right) = \mathbb{P}\left(X_i > X_j \mid j \neq i\right) = \mathbb{P}\left(X_{\sigma(i)} > X_{\sigma(j)} \mid j \neq i\right) = \mathbb{P}\left(X_{\sigma(i)} \text{ is the largest}\right).$$

So

$$1 = \sum_{i=1}^{8} \mathbb{P}\left(X_i \text{ is the largest}\right) = 8\mathbb{P}\left(X_4 \text{ is the largest}\right) \implies \mathbb{P}\left(X_4 \text{ is the largest}\right) = \frac{1}{8}. \qquad \square$$

4. Expectation of random variables

March 12, 2024

If X is a discrete random variable, then its **expectation** is defined as

$$\mathbb{E}[X] = \sum_k k\mathbb{P}(X = k).$$

If X is absolutely continuous with PDF f, then its expectation is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\,dx.$$

It can be thought of as the mean/average of a random variable. In fact, when we say the *mean* of a random variable, we mean its expectation.

4.1. Construction

Our class went twice through defining how to get the expectation of random variables. Once with discrete and absolutely continuous variables, then with a formal measure theoretic construction of the expectation. This section will cover the formal construction.

Let $X\colon \Omega \to \mathbb{R}$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. The goal of this section is to define the expectation as

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,d\mathbb{P}(\omega),$$

where this integral is the **Lebesgue integral** of X over $(\Omega, \mathbb{P})$. We construct this Lebesgue integral from a series of approximations.

Let $X = I_A$ be an indicator for an event $A \in \mathcal{F}$. Then we expect

$$\int_\Omega I_A(\omega)\,d\mathbb{P}(\omega) = \mathbb{P}(A).$$

A variable X is **simple** if it takes only *finitely many* real values. Let those values be $\alpha_1, \ldots, \alpha_m$ and define $A_i \coloneqq \{\omega \mid X(\omega) = \alpha_i\}$. Notice that

$$X(\omega) = \sum_m \alpha_i I_{A_i}(\omega),$$

where A_i are disjoint sets. Next we define $\int_\Omega d\mathbb{P}$ for a linear sum:

$$\int_\Omega X(\omega)\,d\mathbb{P} \coloneqq \sum_{i=1}^{m} \alpha_i \mathbb{P}(A_i) = \sum_{i=1}^{m} \alpha_i \mathbb{P}(X = \alpha_i).$$

To go beyond simple random variables, we consider several cases.

First: X **is a non-negative random variable**. In theory, if we could find simple random variables $X_1, X_2, \ldots$ such that $0 \le X_n(\omega) \nearrow X(\omega)$, then we could define

$$\mathbb{E}[X] \coloneqq \lim_{n\to\infty} \mathbb{E}[X_n].$$

The main problem with this is that we don't know that this is well-defined. That is, if we had another collection of simple random variables $X'_n \nearrow X$, then would $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X'_n]$?

To encompass all possible approximations, we define

$$\mathbb{E}[X] = \sup\{\mathbb{E}[W] \mid W \text{ is a simple random variable such that } 0 \le W(\omega) \le X(\omega) \text{ for all } \omega \in \Omega\}.$$

That is, $\mathbb{E}[X]$ is the upper limit of the expectation of all approximations by simple variables (which we can compute!) of X from below. Now the following proposition is reassuring for us, because it says that any approximation converges to our definition of expectation.

Proposition 4.1. Any approximation $\{X_n\}$ of simple random variables such that $X_n \nearrow X$ has

$$\mathbb{E}[X] = \lim_{n\to\infty} \mathbb{E}[X_n].$$

So it suffices to find *some* approximation of X with simple random variables. The actual construction is defining

$$X_n(\omega) = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} < X(\omega) \leq \frac{k+1}{2^n}, \qquad k = 0, 1, \ldots, n2^n - 1, \\ n & \text{if } X(\omega) > n. \end{cases}$$

By how we defined X_n, if $X(\omega) \in [0, n]$, then

$$0 \leq X(\omega) - X_n(\omega) \leq \frac{1}{2^n}.$$

Hence, we have convergence.

Next: X **is a** $[-\infty, \infty]$**-valued random variable**. We define the **positive and negative parts** of X as

$$X^+(\omega) \coloneqq X(\omega) \vee 0 = \max\{X(\omega), 0\},$$

and

$$X^-(\omega) \coloneqq -X(\omega) \vee 0 = \max\{-X(\omega), 0\}$$

respectively. Then X^+ and X^- are non-negatively-valued random variables, so $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$ are well-defined. Therefore, we let

$$\mathbb{E}[X] \coloneqq \mathbb{E}\left[X^+\right] - \mathbb{E}\left[X^-\right].$$

If $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$ both diverge to ∞, then we will say $\mathbb{E}[X]$ does not exist. In all other cases where $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$ converge or diverge to ∞, $\mathbb{E}[X]$ is a value in $[-\infty, \infty]$.

4.1.1. For practical purposes...

There are two cases where we can always compute $\mathbb{E}[X]$ and have it "make sense."

1. If X has a single sign almost surely: $\mathbb{P}(X \geq 0) = 1$ or $\mathbb{P}(X \leq 0) = 1$, or
2. if X is **absolutely integrable**. That is, $\mathbb{E}[|X|]$ is finite (hence $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}\left[X^-\right]$ are both finite).
   a) If X is discrete, this means $\sum_i p_i|x_i| < \infty$.
   b) If X is continuous, this means $\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty$.

Example 4.2 (Five discrete and five continuous examples of expecatation)**.** Let's compute some expectations.

1. Let I_A be an indicator variable for $A \in \mathcal{F}$. Then
$$\mathbb{E}[I_A] = 0 \cdot \mathbb{P}(I_A = 0) + 1 \cdot \mathbb{P}(I_A = 1) = \mathbb{P}(I_A = 1) = \mathbb{P}(A).$$
2. Let X be the value of a fair die roll. Then
$$\mathbb{E}[X] = \sum_{i=1}^{6} i \cdot \frac{1}{6} = \boxed{\frac{7}{2}}.$$
3. Let $X \sim \text{Poisson}(\lambda)$ (recall this is $\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, 2, \ldots$). Then
$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) = \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!}e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \boxed{\lambda}.$$

4. Let X be a random variable such that $\mathbb{P}\left(X = 2^k\right) = \frac{1}{2^k}$ for $k = 1, 2, \ldots$. Notice that $\sum_{k=1}^{\infty} \frac{1}{2^k} = 1$, so this is indeed a discrete random variable. Notice that

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} 2^k \frac{1}{2^k},$$

which diverges to ∞, so we say X has expectation ∞.

5. Now consider X such that $\mathbb{P}\left(X = (-1)^k 2^k\right) = \frac{1}{2^k}$ for $k = 1, 2, \ldots$. Then

$$\sum_{k=1}^{\infty} (-1)^k 2^k \frac{1}{2^k} = \sum_{k=1}^{\infty} (-1)^k,$$

which is not convergent. To show that it does not have an expectation, we look at X^+ and X^-:

$$X^+ = \begin{cases} 2^{2k} & \text{with probability } \frac{1}{2^{2k}}, \\ 0 & \text{with probability } \frac{2}{3}, \end{cases} \qquad X^- = \begin{cases} 2^{2k+1} & \text{with probability } \frac{1}{2^{2k+1}}, \\ 0 & \text{with probability } \frac{1}{3}. \end{cases}$$

It follows that

$$\mathbb{E}\left[X^+\right] = \infty, \qquad \mathbb{E}\left[X^-\right] = \infty,$$

so $\mathbb{E}[X]$ does not exist.

6. Let $X \sim \text{Unif}[a, b]$. Recall that the PDF f of X is

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

So

$$\mathbb{E}[X] = \int_a^b x \frac{1}{b-a}\, dx = \frac{b^2 - a^2}{(b-a)2} = \boxed{\frac{a+b}{2}}.$$

7. Let $X \sim \text{Exp}(\lambda)$, which has PDF

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We compute

$$\mathbb{E}[X] = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx \overset{\text{(I.B.P.)}}{=} 0 + \int_0^{\infty} e^{-\lambda x}\, dx = \boxed{\frac{1}{\lambda}}.$$

8. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, the normal distribution with parameters μ and σ^2, which has PDF

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

So

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx \\ &= \int_{-\infty}^{\infty} (y + \mu) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}\, dy && (y = x - \mu) \\ &= \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}\, dy + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}\, dy \\ &= \boxed{\mu}. \end{aligned}$$

9. Let X be a random variable with PDF given by

$$f(x) = \begin{cases} \frac{1}{2}x^{-\frac{3}{2}} & x \geq 1, \\ 0 & x < 1. \end{cases}$$

Then its integral on $[1, \infty)$ is 1, but

$$\mathbb{E}[X] = \int_1^\infty x\frac{1}{2}x^{-\frac{3}{2}}\,dx = \frac{1}{2}\int_1^\infty x^{-\frac{1}{2}}\,dx,$$

which diverges to ∞.

10. We demonstrate an absolutely continuous variable with no expectation. Let X be a random variable with PDF given by

$$f(x) = \begin{cases} \frac{1}{4}|x|^{-\frac{3}{2}} & |x| \geq 1, \\ 0 & |x| < 1. \end{cases}$$

Then

$$\mathbb{E}\left[X^+\right] = \int_0^\infty x\frac{1}{4}x^{-\frac{3}{2}}I_{x\geq 1}\,dx = \frac{1}{4}\int_1^\infty x^{-\frac{1}{2}}\,dx = \infty,$$

and

$$\mathbb{E}\left[X^-\right] = \int_{-\infty}^0 (-x)f(x)\,dx = \infty.$$

So $\mathbb{E}[X]$ does not exist.

4.2. Properties of expectation

Proposition 4.3 (Functions of random variables)**.** Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector and let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Then, assuming $g \geq 0$ or $g(\mathbf{X})$ is absolutely integrable,

1. if $X_1, \ldots, X_d$ are discrete, then

$$\mathbb{E}[g(\mathbf{X})] = \sum_{\mathbf{k}\text{ possible}} g(\mathbf{k})\mathbb{P}(\mathbf{X} = \mathbf{k}),$$

2. if $X_1, \ldots, X_d$ are jointly absolutely continuous with PDF f, then

$$\mathbb{E}[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x})f(\mathbf{x})\,d\mathbf{x}.$$

Proof. For the discrete case, first notice that for some fixed $\mathbf{t} \in \mathbb{R}^d$,

$$\sum_{\mathbf{y}\text{ possible}} \mathbf{y}I(g(\mathbf{t}) = \mathbf{y}) = g(\mathbf{t}),$$

because there is only one value of $\mathbf{y}$ where $g(\mathbf{t}) = \mathbf{y}$: $g(\mathbf{t})$. From this, we prove the result with

rearrangement.

$$\begin{aligned}
\mathbb{E}\left[g(\mathbf{X})\right] &= \sum_{\mathbf{y} \text{ possible}} \mathbf{y}\mathbb{P}\left(\mathbf{X}=\mathbf{y}\right) \\
&= \sum_{\mathbf{y} \text{ possible}} \mathbf{y} \sum_{\mathbf{t}:g(\mathbf{t})=\mathbf{y}} \mathbb{P}\left(\mathbf{X}=\mathbf{t}\right) \\
&= \sum_{\mathbf{y} \text{ possible}} \mathbf{y} \sum_{\mathbf{t} \text{ possible}} I(g(\mathbf{t})=\mathbf{y})\mathbb{P}\left(\mathbf{X}=\mathbf{t}\right) \\
&= \sum_{\mathbf{t} \text{ possible}} \mathbb{P}\left(\mathbf{X}=\mathbf{t}\right) \sum_{\mathbf{y} \text{ possible}} \mathbf{y} I(g(\mathbf{t})=\mathbf{y}) \\
&= \sum_{\mathbf{t} \text{ possible}} \mathbb{P}\left(\mathbf{X}=\mathbf{t}\right) g(\mathbf{t}).
\end{aligned}$$

For the continuous case, we need the following claim:

Claim 4.1. If $X \geq 0$, then $\mathbb{E}\left[X\right] = \int_0^\infty \mathbb{P}\left(X > s\right) ds$. We call $\mathbb{P}\left(X > s\right)$ a **tail probability**.

Proof.

$$\begin{aligned}
\int_0^\infty \mathbb{P}\left(X > s\right) ds &= \int_0^\infty \left(\int_s^\infty f(x)\, dx\right) ds \\
&= \int_0^\infty \int_0^x ds f(x)\, dx \\
&= \int_0^\infty x f(x)\, dx \\
&= \int_{-\infty}^\infty x f(x)\, dx \\
&= \mathbb{E}\left[X\right].
\end{aligned}$$

■

For the absolutely continuous case now,

$$\begin{aligned}
\mathbb{E}\left[g(\mathbf{X})\right] &= \int_0^\infty \mathbb{P}\left(g(\mathbf{X}) > s\right) ds \\
&= \int_0^\infty \left(\int_{g(\mathbf{X})>s} f(x_1,\ldots,x_d)\, dx_1 \cdots dx_d\right) ds \\
&= \int_0^\infty \left(\int_{\mathbb{R}^d} I_{g(\mathbf{X})>s} f(x_1,\ldots,x_d)\, dx_1 \cdots dx_d\right) ds \\
&= \int_{\mathbb{R}^d} \left(\int_0^\infty I_{g(\mathbf{X})>s} ds\right) f(x_1,\ldots,x_d)\, dx_1 \cdots dx_d \\
&= \int_{\mathbb{R}^d} g(x_1,\ldots,x_d) f(x_1,\ldots,x_d)\, dx_1 \cdots dx_d.
\end{aligned}$$

In the case where $g \not\geq 0$, we can take $g^+ = g \vee 0$ and $g^- = (-g) \vee 0$ and get

$$\mathbb{E}\left[g(\mathbf{X})\right] = \mathbb{E}\left[g^+(\mathbf{X})\right] - \mathbb{E}\left[g^-(\mathbf{X})\right].$$

□

Example 4.4 (Computing expectation of singular function)**.** Let X have CDF

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{3} & x \in [0,1) \\ 1 & x \geq 1. \end{cases}$$

What is the expectation of X? [notice that X is not discrete nor absolutely continuous]

Solution. Let $U \sim \text{Unif}[0,3]$, and

$$g(u) = \begin{cases} u & u \in [0,1) \\ 1 & u \in [1,3]. \end{cases}$$

Now $\mathbb{P}(g(U) \leq s) = F(s)$, so applying the above theorem,

$$\mathbb{E}[X] = \mathbb{E}[g(U)] = \int_{-\infty}^{\infty} g(x) f_U(x)\,dx = \int_0^3 \frac{1}{3} g(u)\,du = \frac{1}{3}\left(\int_0^1 u\,du + \int_1^3 1\,du\right) = \boxed{\frac{5}{6}}. \quad \square$$

Theorem 4.5 (Properties of expectation)**.** Suppose X and Y are non-negative random variables, or $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are finite. Then

(a) (*linearity of expectation*) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and $\mathbb{E}[aX] = a\mathbb{E}[X]$ for $a \in \mathbb{R}$,

(b) (*monotonicity*) if $X \geq Y$, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$,

(c) if $\mathbb{P}(X = Y) = 1$, then $\mathbb{E}[X] = \mathbb{E}[Y]$,

(d) if $X \overset{d}{=} Y$, then $\mathbb{E}[X] = \mathbb{E}[Y]$,

(e) $|\mathbb{E}[X]| \leq \mathbb{E}[|X|]$,

(f) suppose $X \geq 0$ a.s. and $\mathbb{E}[X] = 0$. Then $X = 0$ a.s.

4.2.1. Linearity of expectation is OP

The linearity of expectation property is very useful because we don't need variables to be independent for it to apply.

Example 4.6. The expectation of $X \sim \text{Binom}(n,p)$ is easily computed by noticing $X = X_1 + \cdots + X_n$ for $X_i \sim \text{Ber}(p)$. Hence, by linearity of expectation,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \boxed{np}.$$

There are more examples in the rest of the notes and online of where linearity is used.

4.3. Variance and covariance

Definition 4.1. Let X be a random variable. Then $\mathbb{E}[X^s]$ (if it exists) is called the s-**moment** of the distribution. $\mathbb{E}[|X|^s]$ is the **absolute** s-**moment** of X.

Definition 4.2. $\text{Var}(X) \coloneqq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ is called the **variance** of X.

By linearity of expectation, the variance is just the second moment minus the expectation squared: $\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Example 4.7. Let $X \sim \text{Binom}(n,p)$. Then $X = X_1 + \cdots + X_n$, where $X_i \sim \text{Ber}(p)$ are i.i.d. So

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) = \boxed{np(1-p)}.$$

Example 4.8 (Variance of normal distribution)**.** The book uses integration by parts. We could also do the differentiation under the sign trick:

$$\int_{-\infty}^{\infty} y^2 e^{-ay^2}\,dy = -\int_{-\infty}^{\infty} \frac{\partial}{\partial a}(e^{-ay^2})\,dy = -\frac{\partial}{\partial a}\left(\int_{-\infty}^{\infty} e^{-ay^2}\,dy\right) = \sqrt{\pi}\frac{1}{2}a^{-\frac{3}{2}} = \sqrt{2\pi}.$$

Example 4.9. Suppose $X = I_1 + \cdots + I_n$ is given by a sum of indicators. Then $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[I_i]$. So the variance of X is given by

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[(I_1 + \cdots + I_n - \mathbb{E}[I_1] - \cdots \mathbb{E}[I_n])^2\right] \\ &= \mathbb{E}\left[\sum_{j=1}^{n} (I_j - \mathbb{E}\left[I_j\right])^2 + 2\sum_{j<k} (I_j - \mathbb{E}\left[I_j\right])(I_k - \mathbb{E}[I_k])\right].\end{aligned}$$

If the indicators are independent, then we have

$$\mathbb{E}\left[(I_j - \mathbb{E}\left[I_j\right])(I_k - \mathbb{E}[I_k])\right] = \mathbb{E}\left[(I_j - \mathbb{E}\left[I_j\right])\right]\mathbb{E}\left[(I_k - \mathbb{E}[I_k])\right] = 0.$$

Hence,

$$\operatorname{Var}(X) = \operatorname{Var}(I_1) + \cdots + \operatorname{Var}(I_n).$$

Replacing indicators with any random variables, we have the following:

Proposition 4.10. If $X_1, \ldots, X_n$ are independent, then

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

On the other hand, if any X_j and X_k were dependent, the value of

$$\mathbb{E}\left[(X_j - \mathbb{E}\left[X_j\right])(X_k - \mathbb{E}[X_k])\right]$$

would be nonzero. We define this to be the **covariance** of X_j and X_k, denoted $\operatorname{Cov}(X_j, X_k)$. So the variance of $X_1 + \cdots + X_n$ has the nice form

$$\sum_{j=1}^{n} \operatorname{Var}(X_j) + 2\sum_{j<k} \operatorname{Cov}(X_j, X_k).$$

By linearity of expectation, $\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$. Notice that $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$. If X and Y are independent, then $\operatorname{Cov}(X, Y) = 0$. The converse is not true.

Example 4.11. Let $X \sim \text{Unif}(\{-1, 0, 1\})$ and $Y = X^2$. Then

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}\left[X^3\right] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0.$$

But X and Y are certainly not independent.

Definition 4.3. If $\operatorname{Cov}(X, Y) = 0$, then we say X and Y are **uncorrelated**.

This means that Proposition 4.10 also holds if X_j and X_k are uncorrelated for all $j \neq k$. Variables being uncorrelated is not the same as them being independent.

Proposition 4.12 (Properties of covariance)**.**

(a) $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$.

(b) $\operatorname{Cov}(aX + b, Y) = a \operatorname{Cov}(X, Y)$.

(c) More generally, covariance is *bilinear*:

$$\operatorname{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \operatorname{Cov}(X_i, Y_j).$$

Since $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$, we derive $\operatorname{Var}(cX) = \operatorname{Cov}(cX,cX) = c^2 \operatorname{Cov}(X,X) = c^2 \operatorname{Var}(X)$.

Definition 4.4. The **correlation coefficient** of X and Y is given by

$$\operatorname{Corr}(X,Y) \coloneqq \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}.$$

Example 4.13. Let $X = I_A$ and $Y = I_B$ for A, B events in some probability space. Then

$$\operatorname{Cov}(X,Y) = \mathbb{E}[I_A I_B] - \mathbb{E}[I_A]\,\mathbb{E}[I_B] = \mathbb{P}(A \cap B) - \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(B)\left(\mathbb{P}(A \mid B) - \mathbb{P}(A)\right).$$

Since $\operatorname{Corr}(X,Y)$ has the same sign as $\operatorname{Cov}(X,Y)$,

- $\operatorname{Corr}(X,Y) = 0 \implies A$ and B are independent,
- $\operatorname{Corr}(X,Y) > 0 \implies \mathbb{P}(A \mid B) > \mathbb{P}(A)$, which means B increases the probability of A, and
- $\operatorname{Corr}(X,Y) < 0 \implies \mathbb{P}(A \mid B) < \mathbb{P}(A)$, which means B decreases the probability of A.

We may wonder why we would bother to look at the sign of $\operatorname{Corr}(X,Y)$ instead of the sign of $\operatorname{Cov}(X,Y)$. The key part is that $\operatorname{Corr}(X,Y)$ has some normalization.

Theorem 4.14. Let X, Y be random variables. All the following hold:

- $-1 \le \operatorname{Corr}(X,Y) \le 1$,
- $\operatorname{Corr}(X,Y) = 1 \implies X = aY + b$ a.s. for $a > 0$,
- $\operatorname{Corr}(X,Y) = -1 \implies X = aY + b$ a.s. for $a < 0$.

Proof. This proof is very similar to the Cauchy-Schwarz inequality proof from analysis. Consider

$$\operatorname{Var}(X - \alpha Y) = \operatorname{Var}(X) + \alpha^2 \operatorname{Var}(Y) - 2\alpha \operatorname{Cov}(X,Y).$$

This is a quadratic in α, and $\operatorname{Var}(X - \alpha Y) \ge 0$, so its discriminant is non-positive. Hence,

$$\operatorname{Cov}(X,Y)^2 - \operatorname{Var}(X)\operatorname{Var}(Y) \le 0. \qquad \square$$

March 21, 2024

Proposition 4.15 (Cauchy-Schwarz inequality)**.** Let X and Y be random variables such that $\mathbb{E}[X^2]$, $\mathbb{E}[Y^2]$, and $\mathbb{E}[XY]$ exist. Then

$$|\mathbb{E}[XY]| \le \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.$$

Proof. Consider $\mathbb{E}[(X - \alpha Y)^2] \ge 0$. $\square$

4.4. Convergence properties

We now return to the construction in subsection 4.1 to formally state/prove some results about expectation. A converging sequence does not necessarily consist of simple random variables.

Theorem 4.16 (Monotone convergence theorem)**.** Let X and $\{X_n\}_{n=1}^{\infty}$ be non-negative random-variables such that $X_n \nearrow X$ (that is, $0 \le X_1(\omega) \le X_2(\omega) \le \cdots \le X(\omega)$ and $X_n(\omega) \to X(\omega)$). Then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Theorem 4.17 (Dominated convergence theorem)**.** Let $X_n \to X$ a.s. Suppose there exists a random variable Y with finite expectation such that $|X_n(\omega)| \le Y(\omega)$ for all $\omega \in \Omega$. Then

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

Proposition 4.18 (Independent expectation is multiplicative)**.** If X, Y are independent with finite expectation, then

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

Moreover, if $X_1, \ldots, X_n$ are independent and g_i are measurable functions,

$$\mathbb{E}[g_1(X_1)\cdots g_n(X_n)] = \mathbb{E}[g_1(X_1)]\cdots\mathbb{E}[g_n(X_n)]$$

We will only prove the first statement. We can prove the second one by noting that $g_1(X_1)$ and $g_2(X_2)\cdots g_n(X_n)$ are independent and using induction.

Proof. For the **discrete** case,

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{k,\ell} k\ell\,\mathbb{P}(X=k, Y=\ell) = \sum_{k} k\mathbb{P}(X=k)\cdot\sum_{\ell}\ell\mathbb{P}(Y=\ell)\\ &= \mathbb{E}[X]\,\mathbb{E}[Y].\end{aligned}$$

For **non-negative, bounded** X, Y, suppose $0 \le X, Y \le M$. Let

$$X_n = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} \le X(\omega) \le \frac{k+1}{2^n}, \quad k = 0, \ldots, M2^{n-1},\\ 0 & \text{otherwise,}\end{cases}$$

and similarly,

$$Y_n = \begin{cases} \frac{k}{2^n} & \text{if } \frac{k}{2^n} \le Y(\omega) \le \frac{k+1}{2^n}, \quad k = 0, \ldots, M2^{n-1},\\ 0 & \text{otherwise.}\end{cases}$$

Notice that $X_n \nearrow X$ and $Y_n \nearrow Y$ and

$$0 \le X_n \le X \le X_n + \frac{1}{2^n}, \quad 0 \le Y_n \le Y \le Y_n + \frac{1}{2^n},$$

which implies

$$X_nY_n \le XY \le \left(X_n + \frac{1}{2^n}\right)\left(Y_n + \frac{1}{2^n}\right).$$

By monotonicity of expectation,

$$\mathbb{E}[X_nY_n] \le \mathbb{E}[XY] \le \mathbb{E}\left[\left(X_n + \frac{1}{2^n}\right)\left(Y_n + \frac{1}{2^n}\right)\right].$$

The variables on the LHS and RHS are independent and discrete, so

$$\mathbb{E}[X_n]\,\mathbb{E}[Y_n] \le \mathbb{E}[XY] \le \mathbb{E}\left[\left(X_n + \frac{1}{2^n}\right)\right]\mathbb{E}\left[\left(Y_n + \frac{1}{2^n}\right)\right] = \mathbb{E}[X_n]\,\mathbb{E}[Y_n] + \frac{1}{2^n}\left(\mathbb{E}[X_n] + \mathbb{E}[Y_n]\right) + \frac{1}{2^{2n}}.$$

Taking the limit as $n \to \infty$, $\mathbb{E}[XY]$ is bounded on both sides by $\mathbb{E}[X]\,\mathbb{E}[Y]$, hence giving us the result.

Now suppose $X, Y \ge 0$ but are **possibly unbounded**. For any integer $M > 0$, define

$$X_M = \min\{X, M\}, \qquad Y_M = \min\{Y, M\}.$$

Then $X_M \to X$, $\quad Y_M \to Y$ and $0 \le X_M \le X$ and $0 \le Y_M \le Y$. Hence,

$$\mathbb{E}[X_n] \to \mathbb{E}[X], \qquad \mathbb{E}[Y_n] \to \mathbb{E}[Y].$$

Now by limit properties,

$$\mathbb{E}[XY] = \lim_{n\to\infty}\mathbb{E}[X_nY_n] = \lim_{n\to\infty}\mathbb{E}[X_n]\,\mathbb{E}[Y_n] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

The proof extends to arbitrary random variables by using X^+, X^-, Y^+, Y^-. □

5. Law of large numbers

April 02, 2024

Let $X_i \sim \text{Ber}(p)$ i.i.d. and

$$S_n := \#\text{ of successes} = \sum_{i=1}^{n} X_i.$$

Then $\frac{S_n}{n}$ is the share of successes in n trials (frequency). We expect $\frac{S_n}{n} \approx p$. Of course, we haven't defined what "$\approx$" means here. There are two different definitions we can use.

Definition 5.1. Let $\{X_n\}_{n=1}^{\infty}$ be defined on the same probability space. We say X_n converges to X **almost surely** ($\lim_{n\to\infty} X_n = X$ a.s.) if $\mathbb{P}(\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)) = 1$.

X_n converges to X **in probability** (denoted $X_n \xrightarrow{p} X$) if for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$.

The statement that $\lim_{n\to\infty} \frac{S_n}{n} = p$ almost surely is the *strong law of large numbers*. $\frac{S_n}{n} \xrightarrow{p} p$ is the *weak law of large numbers*.

5.1. Weak law of large numbers

Theorem 5.1 (Weak law of large numbers/WLNN)**.** Let X_i be i.i.d. with $\mathbb{E}[X_i] = \mu$, and $\text{Var}(X_i) < \infty$. If $S_n = X_1 + \cdots + X_n$, then $\frac{S_n}{n}$ converges to μ in probability.

For this, we need two inequalities.

Theorem 5.2 (Markov's inequality)**.** Let $X \geq 0$. Then for all $a > 0$,

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

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X(I_{X<a} + I_{X\geq a})] \\
&= \mathbb{E}[X(I_{X<a})] + \mathbb{E}[X(I_{X\geq a})] \\
&\geq \mathbb{E}[X(I_{X\geq a})] \\
&\geq \mathbb{E}[aI_{X\geq a}] \\
&= a\mathbb{P}(X \geq a).
\end{aligned}$$

□

Theorem 5.3 (Chebyshev's inequality)**.** Let X be a random variable and $\text{Var}(X) < \infty$, $\mathbb{E}[X] = \mu$. Then for any $a > 0$,

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\text{Var}(X)}{a^2}.$$

In particular, $\mathbb{P}(X \geq \mu + a) \leq \frac{\text{Var}(X)}{a^2}$ and $\mathbb{P}(X \leq \mu - a) \leq \frac{\text{Var}(X)}{a^2}$.

Proof. The random variable $(X-\mu)^2$ is non-negative, so by Markov's inequality,

$$\mathbb{P}(|X - \mu| \geq a) = \mathbb{P}\left((X-\mu)^2 \geq a^2\right) \overset{(5.2)}{\leq} \frac{\mathbb{E}\left[(X-\mu)^2\right]}{a^2} = \frac{\text{Var}(X)}{a^2}.$$

□

Proof of Theorem 5.1. The bound $0 \leq \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right)$ holds trivially. Notice that

$$\mathbb{E}\left[\frac{S_n}{n}\right] = \frac{1}{n}\mathbb{E}[S_n] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]) = \frac{1}{n}(\mu + \cdots + \mu) = \mu.$$

So by Theorem 5.3,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right) &\leq \frac{\operatorname{Var}\left(\frac{S_n}{n}\right)}{\varepsilon^2}\\ &= \frac{\operatorname{Var}(S_n)}{n^2\varepsilon^2}\\ &= \frac{\operatorname{Var}(X_1)+\cdots+\operatorname{Var}(X_n)}{n^2\varepsilon^2}\\ &= \frac{n\operatorname{Var}(X_1)}{n^2\varepsilon^2}\\ &= \frac{\operatorname{Var}(X_1)}{n\varepsilon^2}.\end{aligned}$$

Since ε and $\operatorname{Var}(X_1)$ are fixed, as $n\to\infty$, this value tends to 0. □

Remark 5.4. Some conditions can be relaxed while still having the statement true.

1. $X_1,\ldots,X_n$ do not need to be i.i.d. If we have $\mathbb{E}[X_i]=\mu_i$, then $\frac{S_n}{n}-\frac{\mu_1+\cdots+\mu_n}{n}\xrightarrow{p}0$.
2. We don't need variance to be the same. We just need independence and $\sup_i \operatorname{Var}(X_i)$ being bounded.
3. If $\sum_{i<j}\operatorname{Cov}(X_i,X_j)$ is bounded, then the proof also still works.
4. The variance need not be finite as long as the mean μ is finite.

Example 5.5. Consider flipping a fair coin arbitrarily many times. We will show

$$\mathbb{P}\,(>51\%\text{ of the first }n\text{ flips are H})\to 0$$

as $n\to\infty$. This is equivalent to showing that

$$\mathbb{P}\left(\frac{S_n}{n}>0.51\right)\xrightarrow{n\to\infty}0.$$

Notice that $\mathbb{E}[X_i]=\frac{1}{2}$. We can rewrite this as

$$\mathbb{P}\left(\frac{S_n}{n}-\frac{1}{2}>0.01\right)\xrightarrow{n\to\infty,\ (5.1)}0.$$

Example 5.6. Flip a biased coin with an unknown probability of heads p. Let S_n be the number of head in n trials. Then

$$\widehat{p}=\frac{S_n}{n}$$

is a good approximation of p.

5.2. Infinitely often happening events

We need to develop more tools to prove the strong law of large numbers. Consider a sequence of random variables $\{X_n\}_n$. Notice that $\lim_{n\to\infty}X_n(\omega)$ fails to converge to $X(\omega)$ if there exists an integer $k>0$ such that

$$|X_n(\omega)-X(\omega)|\geq\frac{1}{k}$$

for infinitely many n. We can write this as the event

$$\left\{\omega\mid\lim_{n\to\infty}X_n(\omega)\neq X(\omega)\right\}=\bigcup_{k=1}^{\infty}\left\{\omega\mid|X_n(\omega)-X(\omega)|\geq\frac{1}{k}\text{ for infinitely many }n\right\}.$$

Let $\{A_n\}$ be a sequence of events. Then

$$\begin{aligned}\{\omega \mid \omega \in A_n \text{ for infinitely many } n\} &= \{\omega \mid \text{for all } m > 1 \text{ there is some } n \geq m \text{ s.t. } \omega \in A_n\} \\ &= \bigcap_{m=1}^{\infty} \{\omega \mid \text{there is some } n \geq m \text{ such that } \omega \in A_n\} \\ &= \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n.\end{aligned}$$

We define this as the **limit superior (limsup)** of A_n, denoted $\limsup_{n\to\infty} A_n$.

The complementary event is

$$\begin{aligned}\{\omega \mid \omega \in A_n \text{ for finitely many } n\} &= \{\omega \mid \text{there is some } m \geq 1 \text{ s.t. for all } n \geq m, \omega \notin A_n\} \\ &= \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n^C.\end{aligned}$$

Notice by de Morgan's laws show this is indeed the complement of the limit superior. Replacign A_n^C with the events $\{B_n\}_n$, we call this the **limit inferior (liminf)** of B_n, denoted $\liminf_{n\to\infty} B_n := \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} B_n$, which is the set of ω that lie in all but finitely many B_n.

Hence, we can rewrite

$$\left\{\omega \mid \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right\} = \bigcup_{k=1}^{\infty} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \left\{\omega \mid |X_n(\omega) - X(\omega)| \geq \frac{1}{k}\right\}.$$

Theorem 5.7. If $X_n \to X$ a.s., then $X_n \xrightarrow{p} X$.

Proof. Let $\varepsilon > 0$.

$$\begin{aligned}\lim_{m\to\infty} \mathbb{P}(|X_m - X| \geq \varepsilon) &\leq \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} \{|X_n - X| \geq \varepsilon\}\right) \\ &= \mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{|X_n - X| \geq \varepsilon\}\right) \\ &= \mathbb{P}(|X_n - X| \geq \varepsilon \text{ happens for infinitely many } n) \\ &\leq \mathbb{P}\left(\lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right) = 0.\end{aligned}$$

□

Theorem 5.8 (Borel-Cantelli lemma)**.** Let $\{A_n\}$ be a sequence of events, in the same probability space. Suppose $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. Then

$$\mathbb{P}(\{\omega \mid \omega \in A_n \text{ for infinitely many } n\}) = 0.$$

Proof. Using the fact that if a series $\sum_{n=1}^{\infty} x_n$ converges, then $\lim_{m\to\infty} \sum_{n=m}^{\infty} x_n = 0$, we have

$$\begin{aligned}\mathbb{P}(\{\omega \mid \omega \in A_n \text{ for infinitely many } n\}) &= \mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n\right) \\ &= \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} A_n\right) \\ &\leq \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(A_n) = 0\end{aligned}$$

□

Proof with expectation. Let $N(\omega) = \sum_{n=1}^{\infty} I_{A_n}(\omega)$ (that is, N is a random variable representing how many events A_i contain $\omega \in \Omega$). N is a non-negative random variable with values in $\mathbb{N}_0 \cup \{\infty\}$, so we may take its expectation:

$$\begin{aligned}
\mathbb{E}[N] = \mathbb{E}\left[\sum_{n=1}^{\infty} I_{A_n}\right] &= \mathbb{E}\left[\lim_{m\to\infty} \sum_{n=1}^{m} I_{A_n}\right] \\
&= \lim_{m\to\infty} \mathbb{E}\left[\sum_{n=1}^{m} I_{A_n}\right] \\
&= \lim_{m\to\infty} \sum_{n=1}^{m} \mathbb{E}[I_{A_n}] \\
&= \lim_{m\to\infty} \sum_{n=1}^{m} \mathbb{P}(A_n) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty.
\end{aligned}$$

We got the third inequality by applying Theorem 4.16 to the random variables

$$X_m = \sum_{n=1}^{m} I_{A_n} \nearrow \sum_{n=1}^{\infty} I_{A_n}.$$

Hence, N is finite with probability one. So $\{\omega \mid \omega \in A_n \text{ happens for infinitely many } n\}$ has probability zero. □

April 9, 2024

The use of the Borel-Cantelli lemma for us is the following corollary.

Corollary 5.9 (A.s. convergence from Borel-Cantelli)**.** Let $\{X_n\}_{n=1}^{\infty}$ and X be random variables on the same space. Suppose that for all ε,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) < \infty.$$

Then $X_n \to X$ a.s.

Proof. Define the event

$$B_k := \left\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| \geq \frac{1}{k} \text{ happens for only finitely many } n\right\}$$

for $k = 0, 1, \ldots$. By Theorem 5.8,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n - X| \geq \frac{1}{k}\right) < \infty \implies \mathbb{P}\left(|X_n - X| \geq \frac{1}{k} \text{ for infinitely many } n\right) = 0.$$

Hence, $\mathbb{P}(B_k) = 1$. Define $B := \bigcap_{k=1}^{\infty} B_k$. Then

$$\mathbb{P}\left(B^C\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} B_k^C\right) \leq \sum_{k=1}^{\infty} \mathbb{P}\left(B_k^C\right) = 0 \implies \mathbb{P}(B) = 1.$$

Let $\omega \in B$. So $\omega \in B_k$ for all k. Hence, there exists an N_k such that for $n > N_k$,

$$|X_n(\omega) - X(\omega)| < \frac{1}{k}.$$

Since this holds for all k, $X_n(\omega) \to X(\omega)$. □

5.3. Strong law of large numbers

Theorem 5.10 (Strong law of large numbers/SLNN)**.** Let X_i be i.i.d. with $\mathbb{E}[X_i] = \mu$, and $\mathrm{Var}(X_i) < \infty$. If $S_n = X_1 + \cdots + X_n$, then $\frac{S_n}{n}$ converges to μ a.s., i.e.

$$\mathbb{P}\left(\left\{\omega \mid \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\right\}\right) = 1.$$

For this proof, we assume that $\mathbb{E}[X_i^4]$ is finite (hence the first through fourth moments of X_i are finite).

Proof. Let $\overline{X_n} \coloneqq X_n - \mu$. So $\mathbb{E}[\overline{X_n}] = 0$. Define $\overline{S_n} \coloneqq \overline{X_1} + \cdots + \overline{X_n}$. Now we want to show

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{\overline{S_n}}{n}\right| > \varepsilon\right) < \infty^{1}$$

and then apply Corollary 5.9. By rearranging and taking the expression to the fourth power, we can apply Markov's inequality.

Taking the second power would give us the proof again of the WLNN.

$$\mathbb{P}\left(\left|\frac{\overline{S_n}}{n}\right| > \varepsilon\right) = \mathbb{P}\left(|\overline{S_n}| > \varepsilon n\right) = \mathbb{P}\left((\overline{S_n})^4 > \varepsilon^4 n^4\right) \overset{(5.2)}{\leq} \frac{\mathbb{E}[(\overline{S_n})^4]}{\varepsilon^2 n^4}.$$

Now we bound the expectation in the numerator. First expand $(\overline{S_n})^4 = (\overline{X_1} + \cdots + \overline{X_n})^4$. Notice that since X_i is independent, we can write, e.g.

$$\mathbb{E}\left[(\overline{X_1})^2 \overline{X_2}\,\overline{X_3}\right] = \mathbb{E}\left[(\overline{X_1})^2\right] \underbrace{\mathbb{E}[\overline{X_2}]}_{=0} \underbrace{\mathbb{E}[\overline{X_3}]}_{=0} = 0.$$

So any term in the expansion which has any $\overline{X_i}$ to the first power will vanish. Hence, we have

$$\begin{aligned}
\mathbb{E}\left[(\overline{S_n})^4\right] &= \mathbb{E}\left[(\overline{X_1} + \cdots + \overline{X_n})^4\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[(\overline{X_i})^4\right] + 3\sum_{i<j} \mathbb{E}\left[(\overline{X_i})^2\right] \mathbb{E}\left[(\overline{X_j})^2\right] \\
&= n\mathbb{E}\left[(\overline{X_1})^4\right] + 3n(n-1)\mathbb{E}\left[(\overline{X_1})^2\right] \mathbb{E}\left[(\overline{X_2})^2\right] \\
\implies \frac{\mathbb{E}[(\overline{S_n})^4]}{\varepsilon^2 n^4} &\leq \frac{Cn^2}{\varepsilon^2 n^4} \qquad \text{(for some fixed } C > 0\text{)} \\
&= \frac{C}{\varepsilon^2 n^2} = O(n^{-2}).
\end{aligned}$$

Notice that $\sum_{n=1}^{\infty} \frac{C}{\varepsilon^2 n^2} < \infty$, so Corollary 5.9 finishes. □

Proof with expecatation. Using the bound on $\mathbb{E}[(\overline{S_n})^4]$ from the first proof,

$$\mathbb{E}\left[\sum_{n=1}^{\infty} \left(\frac{\overline{S_n}}{n}\right)^4\right] = \sum_{n=1}^{\infty} \frac{\mathbb{E}[(\overline{S_n})^4]}{n^4} \leq \sum_{n=1}^{\infty} \frac{C}{n^2} < \infty.$$

[1] Notice that this looks *a lot* like the statement in the WLNN. However, if we do the same thing as in the WLNN proof, we will find that each probability is $O(n^{-1})$. But then the infinite sum will behave like the harmonic series $\frac{1}{1} + \frac{1}{2} + \cdots$, which diverges, and hence we cannot apply Borel-Cantelli.

We now consider the infinite series $\sum_{n=1}^{\infty}\left(\frac{\overline{S_n}}{n}\right)^4$ as a non-negative random variable. Since its expectation is finite, it is finite with probability one. So

$$\mathbb{P}\left(\left\{\omega \mid \sum_{n=1}^{\infty}\left(\frac{\overline{S_n}(\omega)}{n}\right)^4 \text{ converges}\right\}\right) = 1.$$

For each ω such that this holds, we have

$$\lim_{n\to\infty}\left(\frac{\overline{S_n}(\omega)}{n}\right)^4 = 0,$$

which is equivalent to

$$\lim_{n\to\infty}\frac{\overline{S_n}(\omega)}{n} = 0,$$

which gives

$$\lim_{n\to\infty}\frac{S_n(\omega)}{n} = \mu. \qquad \square$$

5.3.1. Application: renewal theory

Let X_i be i.i.d. non-negative random variables with mean $\mathbb{E}[X_i] = \mu > 0$. Think of each X_i as the time it takes for a lightbulb to go out. Let

$$T_n \coloneqq X_1 + \cdots + X_n,$$

(e.g. the time for the ith lightbulb to burn out). We want to examine the random variable

$$N_t \coloneqq \#\text{cycles up to time } t$$

(e.g. the number of lightbulbs burnt out) for $t \in \mathbb{N}_0$. The collection $\{N_t\}_{t=0}^{\infty}$ is an example of a **stochastic process**. $\frac{N_t}{t}$ measures how many cycles there are per unit time (e.g. how many lightbulbs you use per unit time).

Theorem 5.11. $\frac{N_t}{t} \to \frac{1}{\mu}$ a.s.

Proof. First notice that as $t \to \infty$, $N_t \to \infty$. Theorem 5.10 tells us that $\frac{T_n}{n} \to \mu$ a.s. We have

$$T_{N_t} \le t < T_{N_{t+1}} \implies \frac{T_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{T_{N_{t+1}}}{N_{t+1}} \cdot \frac{N_{t+1}}{N_t}.$$

Since $\left(\frac{T_{N_t}}{N_t}\right)_t$ is a subsequence of $\left(\frac{T_n}{n}\right)_n$,

$$\lim_{t\to\infty}\frac{T_{N_t}}{N_t} = \mu.$$

Hence,

$$\mu \le \lim_{t\to\infty}\frac{t}{N_t} \le \mu \lim_{t\to\infty}\frac{N_{t+1}}{N_t} = \mu. \qquad \square$$

5.4. Fluctuations

Example 5.12 (Coupon colletor's problem)**.** Stochastic processes show up in many situations. For example, suppose each time period, we sample a random number (with replacement) from $\{1,\ldots,n\}$ and store it in a bag. Let T_n be the time to get all the numbers from 1 to n. Define τ_k as the time it takes to collect k different coupons for $0 \le k \le n$. Define $X_n =$

$\tau_k - \tau_{k-1}$ as the time it takes to collect one more coupon after collecting $k-1$ of them. Notice we can write

$$T_n = \tau_n = (\tau_n - \tau_{n-1}) + \cdots + (\tau_1 - \tau_0) + \underbrace{\tau_0}_{=0} = X_n + \cdots + X_1.$$

Hence,

$$\mathbb{E}[T_n] = \sum_{k=1}^{n} \mathbb{E}[X_k].$$

The probability of choosing a number different from the $k-1$ you already have is $\frac{n-(k-1)}{n}$. As a result,

$$X_k \sim \text{Geom}\left(\frac{n-(k-1)}{n}\right) \implies \mathbb{E}[X_k] = \frac{n}{n-(k-1)}.$$

So

$$\mathbb{E}[T_n] = \sum_{k=1}^{n} \frac{n}{n-(k-1)} = n \cdot \left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right).$$

Define the **harmonic number** as

$$H_n \coloneqq 1 + \frac{1}{2} + \cdots + \frac{1}{n}.$$

From analysis, we have the asymptotics of the harmonic numbers:

$$H_n = \log n + \gamma + \frac{1}{2n} + O\left(\frac{1}{n^2}\right),$$

where γ is a constant (the Euler-Mascheroni constant). From this, we know that $\mathbb{E}[T_n]$ behaves like $n \log n$. That is,

$$\frac{\mathbb{E}[T_n]}{n \log n} \xrightarrow{n\to\infty} 1.$$

Suppose we want to know the **fluctuation** of T_n around $\mathbb{E}[T_n]$. Precisely, this is

$$\mathbb{P}\left(\left|\frac{T_n - \mathbb{E}[T_n]}{n \log n}\right| > \varepsilon\right).$$

We can rearrange and use Chebyshev's inequality to bound this value:

$$\mathbb{P}(|T_n - \mathbb{E}[T_n]| > \varepsilon n \log n) \overset{(5.3)}{\leq} \frac{\text{Var}(T_n)}{\varepsilon^2 n^2 (\log n)^2}.$$

By independence of $X_1, \ldots, X_k$, we have that

$$\begin{aligned}
\text{Var}(T_n) = \sum_{k=1}^{n} \text{Var}(X_k) &= \sum_{k=1}^{n} \frac{\frac{k-1}{n}}{\left(1 - \frac{k-1}{n}\right)^2} \\
&= \sum_{k=1}^{n} \frac{(k-1)n}{(n-k+1)^2} \\
&= n \sum_{\ell=1}^{n} \frac{n-\ell}{\ell^2} && (\ell = n - i + 1) \\
&\leq n^2 \sum_{\ell=1}^{n} \frac{1}{\ell^2} = O(n^2).
\end{aligned}$$

So

$$\frac{\operatorname{Var}(T_n)}{\varepsilon^2 n^2 (\log n)^2} = \frac{O(n^2)}{\varepsilon^2 n^2 (\log n)^2} = O\left(\frac{1}{(\log n)^2}\right) \xrightarrow{n\to\infty} 0.$$

So we actually know

$$\frac{T_n}{n \log n} \xrightarrow{n\to\infty} 1.$$

Remark 5.13. Recall that Theorem 5.1 says for i.i.d. X_i with mean μ and second moment σ^2, $\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\operatorname{Var}(S_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$. This is not the best possible bound; we saw in the proof of Theorem 5.10 that taking a higher power gave us an $O(n^{-2})$ bound. We can go even further than taking powers: for some cases, we have

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \leq e^{-c_\varepsilon n} = O(\exp(-n)).$$

April 11, 2024 **Example 5.14.** Let X_i be i.i.d. $\operatorname{Ber}(p)$ variables. The idea is that $f(x) = \exp(\alpha x)$ is an increasing function for any $\alpha > 0$ (i.e. $a < b \implies \exp(\alpha a) < \exp(\alpha b)$ for any $\alpha > 0$). Let $\varepsilon > 0$. Then we seek an upper tail bound:

$$\begin{aligned}
\mathbb{P}\left(\frac{S_n}{n} > p + \varepsilon\right) &= \mathbb{P}\left(S_n > n(p+\varepsilon)\right) \\
&= \mathbb{P}\left(e^{\alpha S_n} > e^{\alpha n (p+\varepsilon)}\right) && \text{(for } \textit{any}\ \alpha > 0\text{)} \\
&\overset{(5.2)}{\leq} \left(\frac{\mathbb{E}\left[e^{\alpha X_1}\right]}{e^{\alpha(p+\varepsilon)}}\right)^n \\
&= \left(\frac{p(e^\alpha - 1) + 1}{e^{\alpha(p+\varepsilon)}}\right)^n.
\end{aligned}$$

Notice that $1 + x \leq e^x$, so $1 + p(e^\alpha - 1) \leq e^{p(e^\alpha)-1}$. Hence,

$$\mathbb{P}\left(\frac{S_n}{n} > p + \varepsilon\right) \leq \left(e^{p(e^\alpha - 1) - \alpha(p+\varepsilon)}\right)^n = e^{n(p(e^\alpha - 1) - \alpha(p+\varepsilon))}.$$

Now let $\alpha = \frac{\varepsilon}{2}$. Then

$$\underbrace{\left(pe^{\frac{\varepsilon}{2}} - 1 - \frac{\varepsilon}{2}\right)}_{\leq \left(\frac{\varepsilon}{2}\right)^2} - \frac{\varepsilon^2}{2} \leq \frac{\varepsilon^2}{4} - \frac{\varepsilon^2}{2} = -\frac{\varepsilon^2}{4}.$$

Hence,

$$\mathbb{P}\left(\frac{S_n}{n} > p + \varepsilon\right) \leq e^{-\frac{n\varepsilon^2}{4}} = O(\exp(-n)).$$

This bound is called a **Chernoff bound**, and one exists for $\mathbb{P}\left(\frac{S_n}{n} < p - \varepsilon\right)$ as well.

The bound in the proof of Theorem 5.1 was

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\operatorname{Var}(X_i)}{n\varepsilon^2}.$$

This says something about the fluctuation. If $\varepsilon = \frac{c}{n^\alpha}$ for $0 < \alpha < \frac{1}{2}$, then $\frac{\operatorname{Var}(X_1)}{n\varepsilon^2} \xrightarrow{n\to\infty} 0$. However, once we take $\alpha = \frac{1}{2}$, i.e. $\varepsilon = \frac{c}{\sqrt{n}}$, then there is a constant (in fact, a bounded random variable X) in the limit as $n \to \infty$:

$$\frac{S_n}{n} \sim \mu + \frac{X}{\sqrt{n}}.$$

What's surprising is that X is actually the same for any distribution that we take the X_i from, and that X is Gaussian. Proving this leads us to the content of the central limit theorem.

6. Convergence in distribution

Definition 6.1. Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variable (not necessarily from the same probability space) and X some other random variable. We say X_n **converges in distribution** to X (denoted $X_n \Rightarrow X$ or $X_n \xrightarrow{d} X$) if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x where F is continuous.

Example 6.1 (Examples and non-examples of convergence in distribution)**.**

(a) Suppose $X_n \sim \text{Unif}[0, 1 + \frac{1}{n}]$ and $X \sim \text{Unif}[0, 1]$. Then

$$F_n(x) = \begin{cases} 0 & x < 0, \\ \frac{x}{1+\frac{1}{n}} & x \in [0, 1 + \frac{1}{n}], \\ 1 & x > 1 + \frac{1}{n}, \end{cases} \qquad F(x) = \begin{cases} 0 & x > 0, \\ x & x \in [0, 1], \\ 1 & x > 1. \end{cases}$$

Then F is continuous at all $x \in \mathbb{R}$ and $\lim_{n\to\infty} F_n(x) = F(x)$ for all x. So $X_n \xrightarrow{d} X$.

(b) Let $X_n = \frac{1}{n}$ and $X = 0$. Then

$$F_{X_n}(t) = \begin{cases} 0 & t < \frac{1}{n}, \\ 1 & t \geq \frac{1}{n}, \end{cases} \qquad F_X(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geq 0. \end{cases}$$

As $n \to \infty$,

$$F_{X_n}(0) \to 0 \neq 1 = F_X(0).$$

However, this does not bar us from convergence in distribution, because F_X is not continuous at $t = 0$.

(c) Convergence in distribution is a fairly weak condition. Indeed, let $X, Y \sim \text{Ber}(p)$ and $X_n = X$ and $Y_n = Y$ for all n. Then

$$X_n \xrightarrow{d} X, \qquad Y_n \xrightarrow{d} X,$$

but

$$X_n + Y_n \xrightarrow{d} X + Y \sim \text{Binom}(2, p),$$

which does not have the same distribution as

$$2X \sim 2 \cdot \text{Ber}(p).$$

(d) Let $X_i \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$ for integers $1 \leq i \leq n$. Let $M_n := \max\{X_1, \ldots, X_n\}$. We compute the CDF of M_n:

$$\mathbb{P}(M_n \leq t) = \begin{cases} 0 & t < 0, \\ \mathbb{P}(X_1 \leq t, \ldots, X_n \leq t) & t \in [0, 1], \\ 1 & t > 1. \end{cases}$$

Notice that $\mathbb{P}(X_1 \leq t, \ldots, X_n \leq t) = \mathbb{P}(X_1 \leq t)^n = t^n$ since X_i are i.i.d. So when $t \in [0, 1]$, we have

$$\sum_{n=1}^{\infty} \mathbb{P}(M_n \leq t) = \sum_{n=1}^{\infty} t^n = \frac{t}{1-t},$$

which is finite. By Theorem 5.8, $\mathbb{P}(M_n \leq t \text{ for infinitely many } n) = 0$. So $\mathbb{P}(M_n \geq t \text{ for some } n) = 1$. Hence, for all $\varepsilon > 0$, with $t = 1 - \varepsilon$, we have $M_n \xrightarrow{n\to\infty} 1$ almost surely, so $M_n \xrightarrow{d} 1$.

We now observe how "weak" this convergence is. Let's consider the random variable $M_n - 1$.

Claim 6.1. $n(1-M_n) \xrightarrow{d} Z \sim \text{Exp}(1)$. In other words, M_n is "approximately" $1 + \frac{1}{n}\text{Exp}(1)$.

Proof.

$$\begin{aligned}\mathbb{P}\left(n(1-M_n) \le t\right) &= \mathbb{P}\left(1 - M_n \le \frac{t}{n}\right) \\ &= \mathbb{P}\left(M_n \ge 1 - \frac{t}{n}\right) \\ &= 1 - \mathbb{P}\left(M_n < 1 - \frac{t}{n}\right) \\ &= \begin{cases} 0 & t < 0, \\ 1 - (1 - \frac{t}{n})^n & t \in [0, n], \\ 1 & t > n. \end{cases}\end{aligned}$$

Taking the limit as $n \to \infty$, we arrive at the distribution

$$\begin{cases} 0 & t > 0, \\ 1 - e^{-t} & t \ge 0, \end{cases}$$

which is drawn from a $\text{Exp}(1)$ distribution. ■

Theorem 6.2. $X_n \xrightarrow{d} X$ if and only if $\mathbb{E}[g(X_n)] \xrightarrow{n\to\infty} \mathbb{E}[g(X)]$ for any continuous bounded $g\colon \mathbb{R} \to \mathbb{R}$.

Proof (sketch). ($\Longleftarrow$) We can write $\mathbb{P}(X \le x) = \mathbb{E}[I_{X \le x}]$. Let $\widetilde{g}(x)$ be $I_{X \le x}$. Since $\widetilde{g}$ is not continuous, we can successively approximate it with continuous functions.

($\Longrightarrow$) Approximate g with a sum of indicators. Since each indicator will converge, their sum does as well. □

Remark 6.3. If we require that g is three times differentiable and g, g', g'', g''' bounded, then the above theorem still holds.

6.1. Normal distribution

April 16, 2024

Recall $\mathcal{N}(\mu, \sigma^2)$ is called the **Gaussian/normal distribution**. It has PDF

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

and CDF

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt.$$

We leave it as an integral because there is no elementary form of this CDF. We also omit the parameters μ, σ^2 from the subscript when $\mu = 0$ and $\sigma^2 = 1$. Notice that φ (with parameters $\mu = 0$, $\sigma = 1$) is symmetric about $x = 0$, so

$$\Phi(-x) = 1 - \Phi(x).$$

The normal distribution has the property that if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

Therefore, we can "normalize" any normal distribution X to have mean 0 and variance 1 by letting $\overline{X} = \frac{X-\mu}{\sigma}$.

6.2. Central limit theorem

Theorem 6.4 (Central limit theorem)**.** Suppose X_i are i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Hence, for $-\infty \le a < b \le \infty$,

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \mathbb{P}(a \le \mathcal{N}(0,1) \le b) = \Phi(b) - \Phi(a). \tag{6.1}$$

Historically, this was first proven with $X_i \sim \mathrm{Ber}(p)$ (so $S_n \sim \mathrm{Binom}(n,p)$). This is called the *deMoirve-Laplace theorem*, or the *normal approximation of the binomal distribution*.

A natural next question to ask is how large n has to be for us to see a Gaussian. This is a hard question, but there exists bounds.

Theorem 6.5 (Berry-Esseen)**.** For some constant C, the error of the central limit theorem is bounded by

$$\left|\mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{\sigma n^2}} - \Phi(x)\right)\right| \le \frac{C\mathbb{E}\left[|X-\mu|^3\right]}{\sigma^3\sqrt{n}}.$$

In 1941, Berry got the bound with $C = 188$, but Esseen discovered an error in his proof in 1942, which changed the constant to $C = 759$. In 2011, Shevtsova proved that $C = 0.47$ works.

April 18, 2024 *Sketch of proof of Theorem 6.4 for* $X_i \sim \mathrm{Ber}(p)$. Let $q = 1-p$. Since S_n is discrete, we can express the probability as sum:

$$\begin{aligned}
\mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) &= \mathbb{P}\left(a\sqrt{npq} + np \le S_n \le b\sqrt{npq} + np\right) \\
&= \sum_{\substack{a\sqrt{npq}+np \le k \le b\sqrt{npq}+np \\ k \text{ possible}}} \mathbb{P}(S_n = k) \\
&= \sum_{\substack{a\sqrt{npq}+np \le k \le b\sqrt{npq}+np \\ k \text{ possible}}} \binom{n}{k} p^k q^{n-k}.
\end{aligned}$$

The next main idea is to use **Stirling's approximation** for the factorial:

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

Following this, we can show

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k q^{n-k} \sim \varphi\left(\frac{k-np}{\sqrt{npq}}\right) \frac{1}{\sqrt{npq}}.$$

Then with a bound on Riemann sums, we can show the sum is approximately an integral, hence showing that $\mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right)$ looks vaguely like $\Phi(b) - \Phi(a)$. □

6.2.1. Lindenberg swapping

We now prove the general central limit theorem with new techniques. First, "normalize" X_i by setting

$$\overline{X_i} = \frac{X_i - \mu}{\sigma},$$

so $\overline{X_i}$ has mean 0 and variance 1. It suffices to show that

$$\frac{\overline{X_1} + \cdots + \overline{X_n}}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

We know that for $Y_i \sim \mathcal{N}(0,1)$ for $1 \le i \le n$,

$$\frac{Y_1 + \cdots + Y_n}{\sqrt{n}} \overset{d}{=} \mathcal{N}(0,1).$$

The idea with Lindenberg swapping is to carefully swap $\overline{X_i}$ and Y_i and show that the difference does not change too much. To prove convergence in distribution, we need to show

$$\lim_{n\to\infty} \mathbb{E}\left[f\left(\frac{\overline{X_1} + \cdots + \overline{X_n}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right)\right] = 0 \tag{6.2}$$

for all three-times differentiable functions $f\colon \mathbb{R} \to \mathbb{R}$ with f, f', f'', f''' bounded (this is by Remark 6.3). For the following proof, assume that X_i has **finite third moment** ($\mathbb{E}\left[|X_i|^3\right] < \infty$).

Proof of Theorem 6.4 with Lindenberg swapping. Normalize X_i to have mean 0 and variance 1 for $1 \le i \le n$. Define

$$\begin{aligned} S_{n,k} &\coloneqq Y_1 + \cdots + Y_k + X_{k+1} + \cdots + X_n = Z_{n,k} + X_{k+1} \\ Z_{n,k} &\coloneqq Y_1 + \cdots + Y_k + X_{k+2} + \cdots + X_n. \end{aligned}$$

Notice that $S_{n,0} = S_n$ and $S_{n,n} \sim \mathcal{N}(0,1)$. We can rewrite (6.2) as

$$\lim_{n\to\infty} \mathbb{E}\left[f\left(\frac{S_{n,0}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_{n,n}}{\sqrt{n}}\right)\right].$$

We can also use linearity of expectation and rewrite this as a telescoping sum

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \mathbb{E}\left[f\left(\frac{S_{n,k}}{\sqrt{n}}\right) - f\left(\frac{S_{n,k+1}}{\sqrt{n}}\right)\right].$$

Notice that $S_{n,k} - S_{n,k+1} = X_{k+1} - Y_{k+1}$, so we have the substitutions $S_{n,k} = Z_{n,k} + X_{k+1}$, and $S_{n,k+1} = Z_{n,k} + Y_{k+1}$, we rewrite the term inside the expectation above as

$$f\left(\frac{Z_{n,k} + X_{k+1}}{\sqrt{n}}\right) - f\left(\frac{Z_{n,k} + Y_{k+1}}{\sqrt{n}}\right).$$

Expanding this as a Taylor series about $\frac{Z_{n,k}}{\sqrt{n}}$, we find that for some $\frac{Z_{n,k}}{\sqrt{n}} \le \xi_{n,k}, \zeta_{n,k} \le \frac{Z_{n,k}+X_{k+1}}{\sqrt{n}}$,

$$\begin{aligned} &f\left(\frac{Z_{n,k} + X_{k+1}}{\sqrt{n}}\right) - f\left(\frac{Z_{n,k} + Y_{k+1}}{\sqrt{n}}\right) \\ &\quad = f\left(\frac{Z_{n,k}}{\sqrt{n}}\right) + f'\left(\frac{Z_{n,k}}{\sqrt{n}}\right)\frac{X_{k+1}}{\sqrt{n}} + \frac{1}{2!}f''\left(\frac{Z_{n,k}}{\sqrt{n}}\right)\left(\frac{X_{k+1}}{\sqrt{n}}\right)^2 + \frac{1}{3!}f'''\left(\frac{\xi_{n,k}}{\sqrt{n}}\right)\left(\frac{X_{k+1}}{\sqrt{n}}\right)^3 \\ &\qquad - \left(f\left(\frac{Z_{n,k}}{\sqrt{n}}\right) + f'\left(\frac{Z_{n,k}}{\sqrt{n}}\right)\frac{Y_{k+1}}{\sqrt{n}} + \frac{1}{2!}f''\left(\frac{Z_{n,k}}{\sqrt{n}}\right)\left(\frac{Y_{k+1}}{\sqrt{n}}\right)^2 + \frac{1}{3!}f'''\left(\frac{\zeta_{n,k}}{\sqrt{n}}\right)\left(\frac{Y_{k+1}}{\sqrt{n}}\right)^3\right). \end{aligned}$$

Note that $Z_{n,k}$ is independent of X_{k+1} and Y_{k+1}, so expectation will be multiplicative. Moreover, $\mathbb{E}\left[\frac{X_{k+1}}{\sqrt{n}}\right] = \mathbb{E}\left[\frac{Y_{k+1}}{\sqrt{n}}\right] = \frac{1}{\sqrt{n}}$, and $\mathbb{E}\left[\left(\frac{X_{k+1}}{\sqrt{n}}\right)^2\right] = \mathbb{E}\left[\left(\frac{Y_{k+1}}{\sqrt{n}}\right)^2\right] = \operatorname{Var}\left(\frac{X_{k+1}}{\sqrt{n}}\right) + \mathbb{E}\left[\frac{X_{k+1}}{\sqrt{n}}\right]^2 = \operatorname{Var}\left(\frac{Y_{k+1}}{\sqrt{n}}\right) + \mathbb{E}\left[\frac{Y_{k+1}}{\sqrt{n}}\right]^2 = \frac{1}{n}$ so everything but the third power terms will cancel:

$$\begin{aligned}
\mathbb{E}\left[f\left(\frac{Z_{n,k}+X_{k+1}}{\sqrt{n}}\right) - f\left(\frac{Z_{n,k}+Y_{k+1}}{\sqrt{n}}\right)\right] &= \mathbb{E}\left[\frac{1}{6}f'''\left(\frac{\xi_{n,k}}{\sqrt{n}}\right)\left(\frac{X_{k+1}}{\sqrt{n}}\right)^3 - \frac{1}{6}f'''\left(\frac{\zeta_{n,k}}{\sqrt{n}}\right)\left(\frac{Y_{k+1}}{\sqrt{n}}\right)^3\right] \\
&\le \mathbb{E}\left[\left|\frac{1}{6}f'''\left(\frac{\xi_{n,k}}{\sqrt{n}}\right)\left(\frac{X_{k+1}}{\sqrt{n}}\right)^3\right| + \left|\frac{1}{6}f'''\left(\frac{\zeta_{n,k}}{\sqrt{n}}\right)\left(\frac{Y_{k+1}}{\sqrt{n}}\right)^3\right|\right] \\
&\le C\mathbb{E}\left[\frac{|X_{n+1}|^3}{n^{\frac{3}{2}}} + \frac{|Y_{n+1}|^3}{n^{\frac{3}{2}}}\right] \qquad (f''' \text{ finite}) \\
&\le \frac{C'}{n^{\frac{3}{2}}} = O(n^{-\frac{3}{2}}) \xrightarrow{n\to\infty} 0. \\
& \qquad\qquad (\text{finite third moment})
\end{aligned}$$

□

6.3. Applications

The general rule for applying the central limit theorem for Bernoulli random variables is that if $np(1-p) > 10$, Theorem 6.4 will work well. Otherwise, we will use the *Poisson estimation of the binomial*, which we will discuss in subsection 6.4.

Example 6.6. Flip a coin 10^4 times. Let S be the random variable representing the total number of heads. Estimate $\mathbb{P}(S \in [4850, 5100])$.

Solution. Notice that $\mathbb{E}[S] = \frac{1}{2}\cdot 10^4 = 5000$ and $\operatorname{Var}(S) = 10^4 \cdot \frac{1}{2}\cdot\left(1-\frac{1}{2}\right) = 2500$, so $\sqrt{\operatorname{Var}(S)} = 50$. Then

$$\mathbb{P}(4850 \le S \le 5100) = \mathbb{P}\left(-3 \le \frac{S-5000}{50} \le 2\right) \overset{(6.4)}{\approx} \Phi(2) - \Phi(-3) = \Phi(2) + \Phi(3) - 1 \approx 0.9759.$$

The actual value is 0.9765..., so this is a good approximation. □

6.3.1. Continuity correction

We use **continuity correction** when we want to approximate a "small" interval with the central limit theorem.

Example 6.7. Roll a fair die 720 times. What is the probability that there are exactly 113 sixes?

Solution. Let S be the number of sixes. We have $S \sim \text{Binom}(720, \frac{1}{6})$, so the exact value is

$$\mathbb{P}(S = 113) = \binom{720}{113}\left(\frac{1}{6}\right)^{113}\left(\frac{5}{6}\right)^{720-113}.$$

To apply Theorem 6.4, we take an interval around the value we want (noting that $\mathbb{E}[S] = \frac{720}{6} = 120$ and $\sqrt{\operatorname{Var}(S)} = \sqrt{720\cdot\frac{1}{6}\cdot\frac{5}{6}} = 10$):

$$\begin{aligned}
\mathbb{P}(S = 113) &= \mathbb{P}(112.5 < S < 113.5) \\
&\overset{(6.4)}{\approx} \Phi\left(\frac{112.5-120}{10} < \frac{S-120}{10} < \frac{113.5-120}{10}\right) \\
&\approx 0.0312.
\end{aligned}$$

□

Remark 6.8. Moving the bounds by ± 0.5 (i.e. changing from 113 to $(112.5, 113.5)$) might seem arbitrary, but it turns out this is the best value to choose so that the normal is closest to the actual binomial value. Of course, this value changes if X_i are not taken from a Bernoulli distribution.

6.3.2. Confidence intervals

Recall from the law of large numbers, if $X_i \sim \text{Ber}(p)$, then

$$\frac{S_n}{n} \to p, \quad \text{a.s.}$$

so $\frac{S_n}{n} = \widehat{p}$ gives a good estimate of p. We can compute how good this approximation is using Theorem 6.4:

$$\begin{aligned}
\mathbb{P}\left(|\widehat{p} - p| < \varepsilon\right) &= \mathbb{P}\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = \mathbb{P}\left(-\varepsilon n < S_n - np < \varepsilon n\right) \\
&= \mathbb{P}\left(\frac{-\varepsilon n}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{\varepsilon n}{\sqrt{np(1-p)}}\right) \\
&\overset{(6.4)}{\approx} \Phi\left(\frac{\varepsilon n}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-\varepsilon n}{\sqrt{np(1-p)}}\right) \\
&= 2\Phi\left(\frac{\varepsilon n}{\sqrt{np(1-p)}}\right) - 1 \\
&\geq 2\Phi(2\varepsilon\sqrt{n}) - 1. \qquad (p(1-p) \leq 0.25 \implies \sqrt{p(1-p)} \leq 0.5)
\end{aligned}$$

Example 6.9. How many times do we need to flip a coin so that $\widehat{p} = \frac{S_n}{n}$ is within 0.05 from p with probability at least 0.99?

Solution. Using the approximation from above,

$$\mathbb{P}\left(|\widehat{p} - p < \varepsilon|\right) \geq 2\Phi(2 \cdot (0.05) \cdot \sqrt{n}) - 1 \geq 0.99.$$

Using a table, we find that $\Phi(0.1\sqrt{n}) \geq 0.995 \implies n = 666$. □

Definition 6.2. Let $r \in [0, 1]$. The $100r\%$ **confidence interval** for unknown p is $(\widehat{p} - \varepsilon, \widehat{p} + \varepsilon)$, where $\varepsilon > 0$ is chosen such that

$$\mathbb{P}\left(|\widehat{p} - p| < \varepsilon\right) \geq r.$$

Example 6.10. Suppose we run a random Bernoulli trial 1000 times, and we record 450 successes. Find the 95% confidence interval for success probability p.

$$\mathbb{P}\left(|\widehat{p} - p| < \varepsilon\right) \geq 2\Phi(2\varepsilon\sqrt{n}) - 1 \geq 0.95 \implies \varepsilon \approx 0.03.$$

So $p \in (0.42, 0.48)$ is our 95% confidence interval.

Example 6.11 (Polls)**.** Let $p \in [0, 1]$. Suppose $100p\%$ of people like a brand. Let S_n be the number of people that like the brand. So $\widehat{p} = \frac{S_n}{n}$. While this is sampling without replacement, for large n, S_n is approximately $\text{Binom}(n, p)$. For instance, if we want the 90% confidence interval, we need $\varepsilon = 0.082$.

6.4. Poisson estimation of the binomial

It turns out that the Poisson distribution is a good estimation of the binomial when $2np^2$ is very small. This is because of the following:

Theorem 6.12. Let $S_n \sim \text{Binom}(n, p)$ and $Y \sim \text{Poiss}(np)$. Then

$$\sum_{k=0}^{\infty} |\mathbb{P}(S_n = k) - \mathbb{P}(Y = k)| \leq 2np^2.$$

Example 6.13. Let $S \sim \text{Binom}(10, \frac{1}{10})$. Then

$$\mathbb{P}(S \leq 1) = \mathbb{P}(S = 0) + \mathbb{P}(S = 1) = \left(\frac{9}{10}\right)^{10} + \binom{10}{1}\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^{9} = 0.7361...$$

We now approximate this value with the Poisson distribution. The maximum error we expect is $2np^2 = \frac{2}{10}$. Since $np = 1$, take $Y \sim \text{Poiss}(1)$. Then $\mathbb{P}(Y = k) = \frac{1}{k!}e^{-1}$ for $k = 0, 1, \ldots$. Hence,

$$\mathbb{P}(S \leq 1) \approx \mathbb{P}(Y \leq 1) = \mathbb{P}(Y = 0) + \mathbb{P}(Y = 1) = e^{-1} + e^{-1} = 2e^{-1} = 0.7358...$$

So this was a good estimation, and didn't require calculating any binomials.

We will show that the normal approximation does badly. We find that $\mathbb{E}[S] = 1$ and $\text{Var}(S) = np(1-p) = \frac{9}{10} < 10$. So

$$\mathbb{P}(S \leq 1) = \mathbb{P}\left(\frac{S-1}{\sqrt{\frac{9}{10}}} < \frac{1-1}{\sqrt{\frac{9}{10}}}\right) \overset{(6.4)}{\approx} \Phi(0) = \frac{1}{2}.$$

This is quite far off. Even when we add continuity correction (on the top estimate), we get

$$\mathbb{P}(S \leq 1) = \mathbb{P}(S \leq 1.5) = \mathbb{P}\left(\frac{S-1}{\sqrt{\frac{9}{10}}} \leq \frac{0.5}{\sqrt{\frac{9}{10}}}\right) \overset{(6.4)}{\approx} \Phi\left(\frac{0.5}{\sqrt{\frac{9}{10}}}\right) = 0.7019...,$$

which is still not as good as the Poisson estimation.

7. Generating functions in probability

There are three types of generating functions we will go over. Let X be a random variable.

Definition 7.1. The **probability generating function** of X is defined as

$$G_X(s) \coloneqq \mathbb{E}\left[s^X\right], \qquad \forall s \geq 0.$$

Example 7.1. Let $X \geq 0$ and take integer values. Then

$$G_X(s) = \sum_{k=0}^{\infty} s^k \mathbb{P}(X = k).$$

Definition 7.2. The **moment generating function** of X is defined as

$$M_X(s) \coloneqq \mathbb{E}\left[e^{tX}\right], \qquad \text{for all } t \text{ where it exists.}$$

Notice that $M_X(t) = G_X(e^t)$.

Definition 7.3. The **characteristic function** of X is defined as

$$\varphi_X(s) \coloneqq \mathbb{E}\left[e^{itX}\right] = \mathbb{E}\left[\cos(tX) + i\sin(tX)\right], \qquad \forall t \in \mathbb{R}.$$

The characteristic function is particularly nice because it is always defined for all t. It is the one most mathematicians would prefer using, but for this section, we will focus on the moment generating function.

7.1. Properties of moment generating function

Proposition 7.2 (Properties of M_X)**.**

(a) $M_X(0) = 1$ (but can be ∞ everywhere else)

(b) $\frac{d}{dt}\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right] = \mathbb{E}\left[Xe^{tX}\right]$. At $t = 0$, the derivative is $\mathbb{E}[X]$

(c) The nth derivative of M_X evaluated at $t = 0$ is $\mathbb{E}[X^n]$ (provided that M_X *has* derivatives at 0)

(d) If X, Y are independent, then

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX} \cdot e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t).$$

(e) If $M_X = M_Y$ holds on some interval around zero, and is finite, then $X \overset{d}{=} Y$.

Example 7.3 (Some moment generating functions of known distributions)**.**

1. Let $X \sim \text{Geom}(p)$. Let $q = 1 - p$. Then

$$G_X(s) = \sum_{k=1}^{\infty} s^k \mathbb{P}(X = k) = ps \sum_{\ell=0}^{\infty} (qs)^\ell = \begin{cases} \frac{ps}{1-qs} & |qs| < 1, \\ \infty & |qs| \geq 1. \end{cases}$$

$$M_X(t) = \begin{cases} \frac{pe^t}{1-qe^t} & |qe^t| < 1 \quad (\implies t \leq \log\frac{1}{q}), \\ \infty & |qe^t| \geq 1. \end{cases}$$

April 23, 2024

2. Let $X \sim \text{Poiss}(\lambda)$. Recall that $\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$ for $k = 0, 1, \ldots$. Then

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!}e^{-\lambda} \\ &= e^{-\lambda}\sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

3. Computing the normal distribution moment generating function will be very useful for us later. Let $Z \sim \mathcal{N}(0,1)$. Then $f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. So

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} e^{ts}f(s)\,ds \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ts - \frac{s^2}{2}}\,ds \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(s-t)^2+\frac{1}{2}t^2}\,ds \\ &= \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(s-t)^2}\,ds \\ &= e^{\frac{t^2}{2}}. \end{aligned}$$

The expression inside the inegral is just a shifted Gaussian PDF (without the $\frac{1}{\sqrt{2\pi}}$ constant), which evaluates to $\sqrt{2\pi}$.

4. Suppose we know that a random variable X has moment generating function $M_X(t) = \frac{1}{5}e^{-17t} + \frac{1}{4} + \frac{11}{20}e^{2t}$. First notice that $\frac{1}{4} = \frac{1}{4}e^{0\cdot t}$. Since the random variable $\widetilde{X}$ that is -17 with probability $\frac{1}{5}$, 0 with probability $\frac{1}{4}$ and 2 with probability $\frac{11}{20}$ has the same moment generating function (do you see how we found $\widetilde{X}$?), a future result on uniqueness (Proposition 7.7) tells us that $X \overset{d}{=} \widetilde{X}$.

We can use the sum property to compute moment generating functions of sums of random variables.

Example 7.4. Let $X_i \sim \text{Ber}(p)$ be i.i.d. Then $M_{X_i}(t) = 1 - p + pe^t$. Let $S_n = X_1 + \cdots + X_n \sim \text{Binom}(n, p)$. Then

$$M_{S_n}(t) = M_{X_1}(t)\cdots M_{X_n}(t) = (1 - p + pe^t)^n.$$

Moment generating functions also allow us to compute s-moments of random variables.

Example 7.5. We showed that if $Z \sim \mathcal{N}(0,1)$, $M_Z(t) = e^{\frac{t^2}{2}}$. Suppose we want the n-moment of Z. We could compute this directly:

$$\mathbb{E}[Z^n] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t^n e^{-\frac{t^2}{2}}\,dt.$$

On the other hand, moment generating functions give us

$$\mathbb{E}[Z^n] = M_Z^{(n)}(0) = \frac{d^n}{dx^n}e^{\frac{t^2}{2}}\bigg|_{t=0}.$$

If we express $e^{\frac{t^2}{2}}$ around $t = 0$ as a Taylor series then we directly get information about all the derivatives at zero. Recall the Taylor series expansion of M_Z about $t = 0$ is

$$M_Z(t) = \sum_{n=0}^{\infty} \frac{M_Z^{(n)}(t)}{n!} t^n.$$

We know the Taylor expansion of $e^{\frac{t^2}{2}}$:

$$M_Z(t) = \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!}.$$

Now we compare the coefficients of each series to get the values for $M_Z^{(n)}(0)$, which turn out to be

$$\mathbb{E}[Z^n] = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ \frac{n!}{2^{\frac{n}{2}}(\frac{n}{2})!} & \text{if } n \text{ is even.} \end{cases}$$

The moment generating function also has a nice linearity property:

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = e^{tb} M_X(at).$$

Hence, for example, if $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$M_X(t) = e^{t\mu + \frac{\sigma^2 t^2}{2}}.$$

Example 7.6. Suppose a random variable X has $M_X(t) = \left(\frac{2}{3} + \frac{1}{3}e^t\right)^n$. Consider

$$\widehat{X} = \sum_{i=1}^{n} X_i,$$

where X_i are i.i.d random variables that are 0 with probability $\frac{2}{3}$ and 1 with probability $\frac{1}{3}$. Notice that $\widehat{X}$ has the same moment generating function as X, so $X \overset{d}{=} \widehat{X}$.

Proposition 7.7. Let X, Y be random variables. If there exists δ such that for any $t \in (-\delta, \delta)$, $M_X(t) = M_Y(t)$, then X and Y are equal in distribution.

Moreover, $M_{X_n} \to M_X$ around zero if and only if $X_N \xrightarrow{d} X$.

The proof is omitted.

Example 7.8. Suppose $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then their moment generating functions are

$$M_X(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}}, \qquad M_Y(t) = e^{\mu_2 t + \frac{\sigma_2^2 t^2}{2}}.$$

So

$$M_{X+Y}(t) = e^{\mu_1 t + \frac{\sigma_1^2 t^2}{2}} \cdot e^{\mu_2 t + \frac{\sigma_2^2 t^2}{2}} = e^{(\mu_1+\mu_2)t + \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}.$$

By Proposition 7.7, $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Proof of Theorem 6.4 with moment generating functions. Suppose X_i are i.i.d. with mean 0 and variance 1. We want to show that for $S_n = X_1 + \cdots + X_n$,

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Then

$$M_{\frac{S_n}{\sqrt{n}}}(t) = M_{S_n}\left(\frac{t}{\sqrt{n}}\right) = \left(M_{X_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

We can now compute each M_{X_i}:

$$\begin{aligned} M_{X_i}\left(\frac{t}{\sqrt{n}}\right) &= \mathbb{E}\left[e^{\frac{t}{\sqrt{n}}X}\right] \\ &= \mathbb{E}\left[1 + \frac{t}{\sqrt{n}}X + \frac{t^2}{2n}X^2 + O\left(\frac{1}{n^{\frac{3}{2}}}\right)\right] \\ &= 1 + \frac{t^2}{2n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right). \end{aligned}$$

Now one can show with analysis that

$$M_{\frac{S_n}{\sqrt{n}}}(t) = \left(1 + \frac{t^2}{2n} + O\left(\frac{1}{n^{\frac{3}{2}}}\right)\right)^n \xrightarrow{n\to\infty} e^{\frac{t^2}{2}} = M_Z(t),$$

where $Z \sim \mathcal{N}(0,1)$. So Proposition 7.7 tells us $\frac{S_n}{\sqrt{n}} \xrightarrow{d} Z$. □

8. Conditional expectation

The conditional expectation of one variable, X, given another, Y, is a random variable in Y that gives the "best estimate" of X once knowledge is gained about Y.

Much like for expectation, we defined conditional expectation for "nice" random variables (in this case, discrete random variables) before getting into the general construction for any random variables.

We went into less detail with the general construction, and I think it is better to motivate it with the discrete example, so I will put the general construction after.

8.1. Discrete random variables

Definition 8.1. Let X, Y be *discrete* random variables on the same space. The **conditional probability mass function** of X given $Y = y$ is

$$p_{X|Y}(x \mid y) \coloneqq \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}.$$

The **conditional expectation of** X **given** $Y = y$ is

$$\mathbb{E}[X \mid Y = y] \coloneqq \sum_{x \text{ possible}} x\mathbb{P}(X = x \mid Y = y).$$

Notice that $\mathbb{E}[X \mid Y = y]$ is a function of y. Let's call it $v(y)$. The **conditional expectation of** X **given** Y is the random variable

$$\mathbb{E}[X \mid Y] \coloneqq v(Y).$$

Example 8.1.

1. Let X and Y be $\{0, 1\}$-valued random variables such that

$$(X, Y) = (0, 0) \text{ with probability } \frac{3}{10} \qquad (X, Y) = (1, 0) \text{ with probability } \frac{2}{10}$$
$$(X, Y) = (0, 1) \text{ with probability } \frac{1}{10} \qquad (X, Y) = (1, 1) \text{ with probability } \frac{4}{10}$$

   What is $\mathbb{E}[X \mid Y]$?

   Solution. We have,

$$\mathbb{E}[X \mid Y = 0] = 0 \cdot \mathbb{P}(X = 0 \mid Y = 0) + 1 \cdot \mathbb{P}(X = 1 \mid Y = 0) = \frac{\frac{1}{10}}{\frac{4}{10}} = \frac{1}{4},$$

   and

$$\mathbb{E}[X \mid Y = 1] = 0 \cdot \mathbb{P}(X = 0 \mid Y = 1) + 1 \cdot \mathbb{P}(X = 1 \mid Y = 1) = \frac{\frac{4}{10}}{\frac{6}{10}} = \frac{2}{3}.$$

   So $\mathbb{E}[X \mid Y] = \begin{cases} \frac{1}{4} & Y = 0, \\ \frac{2}{3} & Y = 1. \end{cases}$ □

2. Let $X \sim \text{Poiss}(\lambda)$ and $Y \sim \text{Poiss}(\mu)$. Then $Z = X + Y \sim \text{Poiss}(\lambda + \mu)$. Find $\mathbb{E}[X \mid Z = \ell]$ for $\ell = 0, 1, \ldots$

Solution.

$$\begin{aligned} p_{X|Z}(k \mid \ell) &= \frac{\mathbb{P}(X=k, Z=\ell)}{\mathbb{P}(Z=\ell)} = \frac{\mathbb{P}(X=k)\,\mathbb{P}(Y=\ell-k)}{\mathbb{P}(Z=\ell)} \\ &= \begin{cases} 0 & k<0,\ k>\ell, \\ \frac{\ell!}{k!(\ell-k)!}\cdot\frac{\lambda^k\mu^{\ell-k}}{(\ell+\mu)^\ell} & k=0,\dots,\ell, \end{cases} \\ &= \begin{cases} 0 & k<0,\ k>\ell, \\ \binom{\ell}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{\ell-k} & k=0,\dots,\ell. \end{cases} \end{aligned}$$

This is a Binom$(\ell, \frac{\lambda}{\lambda+\mu})$ distribution! So

$$\mathbb{E}[X \mid Z=\ell] = \frac{\lambda}{\lambda+\mu}\ell \implies \mathbb{E}[X \mid Z] = \frac{\lambda}{\lambda+\mu}Z. \qquad \square$$

Proposition 8.2 (Properties of conditional expectation)**.** Let X and Y be random variables.

In general, there is no "linearity" property for Y.

(a) (linearity) $\mathbb{E}[aX_1 + bX_2 \mid Y] = a\mathbb{E}[X_1 \mid Y] + b\mathbb{E}[X_2 \mid Y]$

(b) Suppose X and Y are independent. Then $\mathbb{E}[X \mid Y] = \mathbb{E}[X]$.

(c) Let f be a function. $\mathbb{E}[Xf(Y) \mid Y = f(Y)\mathbb{E}[X \mid Y]]$. In particular, $\mathbb{E}[f(Y) \mid Y] = f(Y)$, and $\mathbb{E}[Y \mid Y] = Y$.

(d) $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$.

Property (d) will show up often in this section. There are many questions about a random variable X that "happens after" Y, which is taken from a known distribution. We can use property (d) to compute the expectation of X as some nice function of the expectation of Y, whose distribution we know.

Proof of property (d). Suppose $\mathbb{E}[X \mid Y] = v(Y)$. Then

$$\begin{aligned} \mathbb{E}[v(Y)] &= \sum_{y \text{ possible}} v(y)\mathbb{P}(Y=y) \\ &= \sum_{y \text{ possible}} \left(\sum_{x \text{ possible}} x\mathbb{P}(X=x \mid Y=y) \right) \mathbb{P}(Y=y) \\ &= \sum_{x \text{ possible}} x \left(\sum_{y \text{ possible}} \mathbb{P}(X=x, Y=y) \right) \\ &= \sum_{x \text{ possible}} x\mathbb{P}(X=x) \\ &= \mathbb{E}[X] \end{aligned}$$

$\square$

8.2. General construction

Definition 8.2. Let X and Y be random variables on the same space. Suppose $\mathbb{E}[|X|]$ is finite. Then the **conditional expectation** of X given Y is a function $v(Y)$ such that

$$\mathbb{E}[v(Y)h(Y)] = \mathbb{E}[Xh(Y)]$$

for all bounded and (Borel) measurable functions h.

We will run into some sneaky issues from this definition. Namely:

1. Does such a function $\nu(Y)$ exist?
2. Is it unique (i.e. can we call it *the* conditional expectation)?
3. Does the discrete definition match the general case?

(1) and (2) are more fit for a measure theory course. We will prove (3).

Proof of issue (3).

$$\begin{aligned}\mathbb{E}[\nu(Y)h(Y)] &= \sum_{y \text{ possible}} \nu(y)h(y)\mathbb{P}(Y=y)\\ &= \sum_{x,y \text{ possible}} xh(y)\mathbb{P}(X=x \mid Y=y)\,\mathbb{P}(Y=y)\\ &= \sum_{x,y \text{ possible}} xh(y)\mathbb{P}(X=x \mid Y=y)\\ &= \mathbb{E}[Xh(Y)]\end{aligned}$$

□

Remark 8.3. This definition requires checking a large space of functions. It actually suffices to show that $\mathbb{E}[\nu(Y)h(Y)] = \mathbb{E}[Xh(Y)]$ holds for all h that are indicator functions.

Thankfully, everything from Proposition 8.2 still applies for this definition. For example, we can prove linearity:

$$\begin{aligned}\mathbb{E}[(aX_1 + bX_2)h(Y)] &= a\mathbb{E}[X_1h(Y)] + b\mathbb{E}[X_2h(Y)]\\ &= a\mathbb{E}[\mathbb{E}[X_1 \mid Y]h(Y)] + b\mathbb{E}[\mathbb{E}[X_2 \mid Y]h(Y)]\\ &= \mathbb{E}[(a\mathbb{E}[X_1 \mid Y] + b\mathbb{E}[X_2 \mid Y])h(Y)].\end{aligned}$$

8.2.1. Absolutely continuous conditional expectation

As expected, we can also do conditional expectation with absolutely continuous random variables. Suppose X and Y have PDF f. Then we can compute $\mathbb{E}[X \mid Y]$ as follows. Let

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

Then

$$\mathbb{E}[X \mid Y=y] = \int_{-\infty}^{\infty} xf_{X|Y}(x,y)\,dx = \nu(y),$$

for some function ν. Then

$$\mathbb{E}[X \mid Y] = \nu(Y).$$

8.3. Examples of conditional expectation

April 30, 2024

Example 8.4. Cut a stick of length 1 once. Take the first portion of the stick after the cut and cut it again. What is the expected length of the first portion of the stick after both cuts? What about the variance?

Solution. We can reframe this as a conditional expectation problem. Let Y be the length of the first portion after the first cut. Let X be the length of the first portion after the second cut. Notice that once we know $Y = y \in (0,1)$, $X \sim \text{Unif}(0,y)$. Then

$$\mathbb{E}[X \mid Y=y] = \frac{y}{2} \implies \mathbb{E}[X \mid Y] = \frac{Y}{2}.$$

Hence,

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}\left[\frac{Y}{2}\right] = \frac{1}{4}.$$

We can also compute $\operatorname{Var}(X)$:

$$\mathbb{E}\left[X^2 \mid Y = y\right] = \int_0^y x^2 \frac{1}{y}\,dx = \frac{1}{y}\left[\frac{x^3}{3}\right]_{x=0}^{y} = \frac{y^2}{3} \implies \mathbb{E}\left[X^2 \mid Y\right] = \frac{Y^2}{3}.$$

So

$$\mathbb{E}\left[X^2\right] = \mathbb{E}\left[\mathbb{E}\left[X^2 \mid Y = y\right]\right] = \mathbb{E}\left[\frac{Y^2}{3}\right] = \frac{1}{3}\int_0^1 y^2\,dy = \frac{1}{9}.$$

So

$$\operatorname{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{1}{9} - \frac{1}{16} = \frac{7}{144}. \qquad \square$$

Let's continue with this example. Suppose we want to find f_X. We have

$$f_{X|Y}(x \mid y) = \begin{cases} \frac{1}{y} & 0 < x < y \\ 0 & \text{otherwise.} \end{cases}, \qquad f_Y(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y) f_Y(y) = \begin{cases} \frac{1}{y} & 0 < x < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

With the joint PDF, we can compute the marginal PDF f_X by integrating:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \begin{cases} \int_x^1 \frac{1}{y}\,dy = -\log x & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Notice now that $\mathbb{E}[X]$ and $\mathbb{E}\left[X^2\right]$ are now computable as integrals. However, they are much harder than the solution we gave before.

Example 8.5. Suppose n people apply for a job that has two screening tests. The first test has a probability p of success. Only people who passe the first test may take the second one, which has probability r of success.

Let L be the number of people who passed the second test. What is $\mathbb{E}[L]$? What is the distribution of L (i.e. p_L)?

Solution. Let M be the number of people who passed the first test. Given $M = m$, we have $L \sim \operatorname{Binom}(m, r)$. Then

$$\mathbb{E}[L \mid M = m] = mr \implies \mathbb{E}[L \mid M] = Mr.$$

Then

$$\mathbb{E}[L] = \mathbb{E}[\mathbb{E}[L \mid M]] = \mathbb{E}[Mr] = r\mathbb{E}[M] = npr.$$

We know

$$p_{L|M}(\ell \mid m) = \binom{m}{\ell} r^\ell (1-r)^{m-\ell}, \qquad (\ell = 0, 1, \ldots, m)$$

and

$$p_M(m) = \binom{n}{m} p^m (1-p)^{n-m}. \qquad (m = 0, 1, \ldots, n)$$

So

$$\begin{aligned} p_{L,M}(\ell, m) &= p_{L|M}(\ell \mid m) p_M(m) \\ &= \binom{m}{\ell} r^\ell (1-r)^{m-\ell} \binom{n}{m} p^m (1-p)^{n-m}. \qquad (0 \le \ell < m \le n) \end{aligned}$$

Giving us (for $0 \leq \ell \leq n$),

$$\begin{aligned}
p_L(\ell) &= \sum_{m=\ell}^{n} p_{L,M}(\ell, m) \\
&= \sum_{m=\ell}^{n} \frac{m!}{\ell!(m-\ell)!} \cdot \frac{n!}{m!(n-m)!} r^{\ell}(1-r)^{m-\ell} p^{m}(1-p)^{n-m} \\
&= \frac{n! r^{\ell}}{\ell!} \sum_{m=\ell}^{n} \frac{1}{(m-\ell)!(n-m)!} (1-r)^{m-\ell} p^{m}(1-p)^{n-m} \\
&= \frac{n! r^{\ell} p^{\ell}}{\ell!(n-\ell)!} \sum_{m=\ell}^{n} \frac{(n-\ell)!}{(m-\ell)!(n-m)!} (1-r)^{m-\ell} p^{m-\ell}(1-p)^{n-m} \\
&= \binom{n}{\ell} r^{\ell} p^{\ell} \sum_{m=\ell}^{n} \binom{n-\ell}{m-\ell} ((1-r)p)^{m-\ell}(1-p)^{n-m} \\
&= \underbrace{\binom{n}{\ell} (rp)^{\ell}(1-rp)^{n-\ell}}_{\sim \text{Binom}(n, rp)}.
\end{aligned}$$

We want this to "look" binomial.

□

Example 8.6. Roll a die infinitely many times. Let N be the number of trials to get the first 6 and Y the number of 5's in the first N trials. Compute $\mathbb{E}[Y]$.

Solution. Given $N = n$, $Y \sim \text{Binom}(n-1, \frac{1}{5})$.[2] Then

$$\mathbb{E}[Y \mid N = n] = \frac{n-1}{5} \implies \mathbb{E}[Y \mid N] = \frac{N-1}{5}.$$

Hence,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid N]] = \mathbb{E}\left[\frac{N-1}{5}\right] = \frac{1}{5}\mathbb{E}[N] - \frac{1}{5} = 1.$$

□

Solution with symmetry. Let Y_i be the number of i's in the first N trials for $i = 1, 2, \ldots, 5$. We want to compute $\mathbb{E}[Y_5]$. Notice that $N = Y_1 + \cdots + Y_5 + 1$. Hence,

$$6 = \mathbb{E}[N] = \mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_5] + 1 \overset{\text{(symm.)}}{\implies} 5\mathbb{E}[Y_5] + 1 = 6 \implies \mathbb{E}[Y_5] = 1.$$

□

Proposition 8.7 (Wald's identity)**.** Let N be a non-negative integer random variable with $\mathbb{E}[N] = \lambda$. Let X_i be i.i.d. random variables that are independent of N with $\mathbb{E}[X_i] = \mu$. Then $\mathbb{E}[X_1 + \cdots + X_N] = \mu\lambda$.

Linearity does not apply here because the sum is taken to X_N!

Proof.

$$\begin{aligned}
\mathbb{E}[X_1 + \cdots + X_N \mid N = n] &= \mathbb{E}[X_1 + \cdots + X_n \mid N = n] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i \mid N = n] \overset{\text{(indep.)}}{=} \sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \mu n \\
\implies \mathbb{E}[X_1 + \cdots + X_N \mid N] &= \mu N.
\end{aligned}$$

So

$$\mathbb{E}[X_1 + \cdots + X_N] = \mathbb{E}[\mathbb{E}[X_1 + \cdots + X_N \mid N]] = \mathbb{E}[\mu N] = \mu\lambda.$$

□

[2] Noticing this is subtle. It follows from the fact that the nth roll is already determined, and because we know that none of the first $n-1$ rolls are 6.

Example 8.8. We cannot use Proposition 8.7 without independence. Consider Example 8.6 and let I_j indicate if the jth roll was 5. Then

$$Y = I_1 + \cdots + I_{N-1},$$

but

$$\mathbb{E}[Y] \neq \mathbb{E}[N-1] \cdot \mathbb{E}\left[I_j\right] = 5 \cdot \frac{1}{6}.$$

This is because I_j is *not* independent of N.

Example 8.9. Let $X \sim \text{Poiss}(\lambda)$ represent the number of customers in a store. Suppose we have three types of coupons with probability p_1, p_2, p_3. Let X_i be the number of customers with coupon i ($i = 1, 2, 3$). Then if $X = n$,

$$(X_1, X_2, X_3) \sim \text{Multinom}(n, 3, p_1, p_2, p_3).$$

Suppose we have k_1, k_2, k_3 such that $k_1 + k_2 + k_3 = n$. Then

$$\begin{aligned}
\mathbb{P}(X_1 = k_1, X_2 = k_2, X_3 = k_3) &= \mathbb{P}(X_1 = k_1, X_2 = k_2, X_3 = k_3 \mid X = n)\,\mathbb{P}(X = n) \\
&= \frac{n!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \frac{\lambda^n}{n!} e^{-\lambda} \\
&= \frac{\lambda^{k_1} p_1^{k_1} \lambda^{k_2} p_2^{k_2} \lambda^{k_3} p_3^{k_3}}{k_1!k_2!k_3!} e^{-\lambda(p_1+p_2+p_3)}.
\end{aligned}$$

Hence, X_i are independent, and $X_i \sim \text{Poiss}(\lambda p_i)$. This is a *Poisson process* (see MATH623).

8.4. What does "best guess" mean?

We considered $\mathbb{E}[X \mid Y]$ as the "best guess" of X given Y. We will show this formally.

Proposition 8.10. Let X have $\mathbb{E}[X] = \mu$. Then

$$\min_a \mathbb{E}\left[(X-a)^2\right] = \text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right].$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[(X-a)^2\right] &= \mathbb{E}\left[((X-\mu)-(a-\mu))^2\right] \\
&= \mathbb{E}\left[(X-\mu)^2\right] + (a-\mu)^2 - 2(a-\mu)\mathbb{E}[(X-\mu)] \\
&= \text{Var}(X) + (a-\mu)^2.
\end{aligned}$$

This attains its minimum at $a = \mu$. □

Theorem 8.11. Let $\mathbb{E}\left[X^2\right]$ and $\mathbb{E}\left[Y^2\right]$ be finite. Then

$$\inf_h \mathbb{E}\left[(X-h(Y))^2\right] = \mathbb{E}\left[(X-\mathbb{E}[X \mid Y])^2\right],$$

where the infimum is taken over all measurable functions.

We can think of $\mathbb{E}[X \mid Y]$ as the "projection" of X onto the function Y.

Proof.

$$\begin{aligned}
\mathbb{E}\left[(X-h(Y))^2\right] &= \mathbb{E}\left[((X-\mathbb{E}[X \mid Y]) - (h(Y)-\mathbb{E}[X \mid Y]))^2\right] \\
&= \mathbb{E}\left[(X-\mathbb{E}[X \mid Y])^2\right] + \mathbb{E}\left[(h(Y)-\mathbb{E}[X \mid Y])^2\right] \\
&\quad - \underbrace{\mathbb{E}[(X-\mathbb{E}[X \mid Y])(\mathbb{E}[X \mid Y]-h(Y))]}_{=0}.
\end{aligned}$$

By the general definition of conditional expectation.

The term $\mathbb{E}\left[(h(Y) - \mathbb{E}[X \mid Y])^2\right]$ is non-negative, so we have that

$$\mathbb{E}\left[(X - h(Y))^2\right] \geq \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2\right], \tag{8.1}$$

as desired. To show this inequality is strict, notice that

$$\mathbb{E}\left[(h(Y) - \mathbb{E}[X \mid Y])^2\right] = \int_{-\infty}^{\infty} (\mathbb{E}[X \mid Y = y] - h(y))^2 f_Y(y)\, dy.$$

If $h(y) \neq \mathbb{E}[X \mid Y = y]$ for all y such that $f_Y(y) > 0$, then this term is positive, so the bound in (8.1) becomes strict. □