6.S095 Notes

Lecturer: Peggy Yang

Andrew Liu

IAP 2023

My notes for 6.S095, "Probability Problem Solving". The instructor for this course (advanced track) was Peggy Yang.

Last updated on Saturday 27th May, 2023.

Contents

1 January 10, 2023

1.1 Basic Counting Principles

Lemma 1.1 (Addition Principle)
For n disjoint sets $S_1, S_2, \ldots, S_n$, the cardinality of their sum

$$|S_1 + S_2 + \ldots + S_n| = |S_1| + \ldots + |S_n|.$$

Definition 1.2
Let $S_1, \ldots, S_n$ be finite sets. Their **Cartesian Product** is defined

$$S_1 \times \ldots \times S_n = \{(s_1, \ldots, s_n) | s_1 \in S_1, \ldots, s_n \in S_n\}.$$

Lemma 1.3 (Multiplication Principle)
For n disjoint sets $S_1, S_2, \ldots, S_n$, the cardinality of their cartesian product

$$|S_1 \times \ldots \times S_n| = |S_1| \cdot \ldots \cdot |S_n|.$$

Definition 1.4
A function $f : A \to B$ between finite sets A and B is called a **bijection** if $f(a) = f(b) \implies a = b$ (injectivity) and for all $b \in B$, there exists $a \in A$ such that $f(a) = b$ (surjectivity).

Theorem 1.5 (Pascal's identity)
For nonnegative integers n, k,

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}.$$

Proof. The left hand side represents the number of ways to choose $k+1$ elements from a set of size $n+1$. Another way to count this is to consider whether or not to include the last element in the set. If this element is included, this contributes $\binom{n}{k}$. If this element is not included, this contributes $\binom{n}{k+1}$. Together, this forms the right hand side, so we are done. □

1.2 PIE

Lemma 1.6 (PIE)
Let $A_1, A_2, \ldots, A_n$ be finite sets. Then

$$|A_1 \cup \ldots \cup A_n| = \sum_{j=1}^{n} (-1)^{j-1} \sum_{\{i_1,\ldots,i_j\} \subseteq [n]} |A_{i_1} \cap \ldots \cap A_{i_n}|.$$

Example 1.7 (Derangements)
A permutation on n elements π_n is called a **derangement** if $\pi(i) \neq i$ for all $i \in [n]$. Let $D(n)$ be the number of derangements in S_n. Then

$$D(n) = \sum_{k=0}^{n} (-1)^k \frac{n!}{k!}.$$

Proof. $D(n)$ is equal to $n!$ minus the total number of permutations that have at least one fixed point. Let A_i be the set of all permutations which fixes i. Then

$$D(n) = n! - |A_1 \cup \ldots \cup A_n|.$$

For any set of k points which are fixed, there are $(n-k)!$ ways to permutate the remaining $(n-k)$ elements. Therefore, by PIE,

$$\begin{aligned} |A_1 \cup \ldots \cup A_n| &= \sum_{j=1}^{n} (-1)^{j-1} \sum_{\{i_1,\ldots,i_j\} \subseteq [n]} |A_{i_1} \cap \ldots \cap A_{i_n}| \\ &= \sum_{j=1}^{n} (-1)^{j-1} \binom{n}{j} (n-j)! = \sum_{j=1}^{n} (-1)^{j-1} \frac{n!}{j!}. \end{aligned}$$

Substituting this into our expression for $D(n)$ gives us the desired result. □

Using the taylor series for e, it can be proven that $D(n) = \lfloor n!/e + 1/2 \rfloor$.

Example 1.8 (Euler Totient)
For any positive integer m, the euler totient $\varphi(m)$ is defined as the number of positive integers between 1 and m inclusive that are coprime to n. If the prime factorization of m is $p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$, then

$$\varphi(m) = m \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right).$$

Proof. Consider instead the number of positive integers that are not coprime to n. These positive integers have at least 1 prime power in common with m. Let A_i denote the set of positive integers $\leq m$ with prime power p_i. Then

$$\begin{aligned}
\varphi(m) &= m - |A_1 \cup \dots \cup A_k| \\
&= m - \left(\sum_{j=1}^{k} (-1)^{j-1} \sum_{\{i_1,\dots,i_j\} \subseteq [m]} \frac{m}{p_{i_1} \cdot \dots \cdot p_{i_j}} \right) \\
&= m \left(1 + \sum_{j=1}^{k} \sum_{\{i_1,\dots,i_j\} \subseteq [m]} \left(\frac{-1}{p_{i_1}}\right)\left(\frac{-1}{p_{i_2}}\right) \dots \left(\frac{-1}{p_{i_j}}\right) \right) \\
&= m \prod_{i=1}^{k} \left(1 - \frac{1}{p_i}\right),
\end{aligned}$$

where the last line follows by polynomial expansion. □

2 January 10, 2023

2.1 Probability Measure

Definition 2.1
A **sample space** Ω is a set of individual outcomes. An **event space** $\mathcal{F}$ is a family of subsets of Ω.

Technical jargon: $\mathcal{F}$ must form a σ-algebra over Ω, meaning that Ω is in $\mathcal{F}$, and so are complements and countable intersections, unions of elements in Ω. When

Ω is finite, we assume that $\mathcal{F}$ is just the power set of Ω. We won't talk more about the technical details for σ-algebras.

Example 2.2
Roll 2 fair die.

The sample space is $\Omega = \{(i,j), 1 \le i,j \le 6\}$ recording the pair of rolls. The event space $\mathcal{F} = \mathcal{P}(\Omega)$ is the power set of Ω. If we let E be the event that the sum of the rolls is 7, then

$$E = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} \in \mathcal{F}.$$

Definition 2.3
A **probability measure** is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ satisfying

- $\mathbb{P}[\Omega] = 1$.
- For disjoint events $A_1, \ldots, A_n$,

$$\mathbb{P}\left[\bigcup_{i \ge 1} A_i\right] = \sum_{i \ge 1} \mathbb{P}[A_i]$$

The principle of inclusion-exclusion holds for probability measures.

Theorem 2.4
For events $A, B \in \mathcal{F}$,

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B],$$

and appropriate generalizations hold for an arbitrary number of events.

Proof. Use set operations. □

The simplest probability measure that can be defined on discrete sample space is the **counting measure**, also called the **uniform measure**.

Definition 2.5

The **counting measure** $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is defined by

$$\mathbb{P}[E] = \frac{|E|}{|\Omega|},$$

for all $E \in \mathcal{F}$.

Intuitively, this measure defines $\mathbb{P}[E]$ as the number of satisfying cases $|E|$ divided by the total number of cases $|\Omega|$. Another probability measure can be defined via the probability mass function (pmf).

Definition 2.6

If $\Omega = \{\omega_1, \ldots, \omega_n\}$ is a finite set and $p(\omega_1), \ldots, p(\omega_n)$ are nonnegative real numbers that sum to 1,

$$\mathbb{P}[E] = \sum_{\omega \in E} p(\omega)$$

defines a probability measure on $\mathcal{P}(\Omega)$ (i.e., $\mathcal{F}$). The function p is called the **probability mass function** (pmf) of $\mathbb{P}$.

2.2 Continuous Probability Spaces

How do we pick a random number from $[0,1]$? We want a uniform measure $\mathbb{P}$ on $[0,1]$ that satisfies $\mathbb{P}[(a,b)] = b - a$. This is not possible for $\mathcal{F} = \mathcal{P}(\Omega)$ (which is not really even well-defined). The solution is to restrict events to a smaller set $\mathcal{B}_{[0,1]}$, called the **Borel sets** of $[0,1]$. As before, the event space (in this case, the Borel sets) are a σ-algebra over Ω, meaning that they are formed by countable unions and intersections of elements in Ω. Borel sets are formed specifically by unions and intersections of open intervals. $\mathcal{B}_{[0,1]}$ can also contains closed intervals. For example,

$$[0.1, 0.2] = \bigcap_{i \geq 5} (0.1 - 2^{-i}, 0.2 + 2^{-i}),$$

which is a closed interval formed by a union of countably infinite open intervals that approach the upper and lower bounds infinitely close.

When $\mathbb{P}$ is uniform on $[0,1]$, what is $\mathbb{P}[0.5]$? If $\mathbb{P}[0.5] = c > 0$, every $\mathbb{P}[\{x\}] = c$, so $\mathbb{P}$ on any set with more than $1/c$ elements exceeds 1. Therefore, $\mathbb{P}$ of any singleton

is zero. This leads to a natural question: is this a contradiction?

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[\cup_{x\in[0,1]}\{x\}] = \sum_{x\in[0,1]} \mathbb{P}(\{x\}) = 0.$$

The answer is no. Our event space is $\mathcal{B}_{[0,1]}$, which has a countably infinite number of elements. On the other hand, the above is summing over all real number between 0 and 1, which is an uncountably infinite set. Over this set, $\mathbb{P}$ is not necessarily a probability measure, i.e., the sum of the probabilities of disjoint events need not strictly equal the sum of their union.

2.3 Random Variables

Random variables correspond to observations on random experiments. For example,

- Ω is the set of people in Cambridge.
- Experiment: Pick random person
- Observation: height H of person chosen.

$H : \Omega \to \mathbb{R}$ is called a **random variable**.

Definition 2.7
Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is a function $\Omega \to \mathbb{R}$.

Definition 2.8
Given random variable X on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider the function $\mathbb{P}_X : \mathcal{B}_{\mathbb{R}} \to [0,1]$ defined by

$$\mathbb{P}_X[B] = \mathbb{P}[X^{-1}(B)] = \mathbb{P}[\{\omega \in \Omega | X(\omega) \in B\}].$$

$\mathbb{P}_X$ is called the **pushforward** of X and determines a probability measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

$\mathbb{P}$ takes events as input, whereas $\mathbb{P}_X$ takes as input a subset of $\mathbb{R}$. This is important!

Example 2.9

Let $\Omega = \{H, T\}$ and $\mathbb{P}$ be the counting measure on Ω. Let X be a random variable with $X(H) = 5$ and $X(T) = 10$ which represents how many dollars you win for flipping a head or a tail. Its pushforward measure $\mathbb{P}_X$ satisfies

$$\mathbb{P}_X[B] = \frac{1}{2}1_{5\in B} + \frac{1}{2}1_{10\in B},$$

where $1_{x\in B}$ is 1 when $x \in B$ and 0 otherwise.

For concreteness, $\mathbb{P}_X[\{5\}]$ is the image of the set of elements in Ω which satisfies $X(\Omega) = 5$ under $\mathbb{P}$, which is 1/2 (the only such element is H). Intuitively, $\mathbb{P}_X$ looks at the probabilities of X taking on certain values and not necessarily which specific elements cause $X(\omega)$ to take on those values. $\mathbb{P}_X$ is also referred to as the **distribution** or **law** of X.

Theorem 2.10

Pushforward $\mathbb{P}_X$ defines a probability measure on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$.

Proof. $\mathbb{P}_X[\Omega] = \mathbb{P}_X[\mathbb{R}] = \mathbb{P}[X \in \mathbb{R}] = 1$, since X is a real number by definition. Also, for disjoint $A_1, \ldots, A_n \in \mathcal{B}_\mathbb{R}$,

$$\begin{aligned}\mathbb{P}_X\left[\bigcup_i A_i\right] &= \mathbb{P}\left[\bigcup_i \{\omega \in \Omega | X(\omega) \in A_i\}\right] \\ &= \sum_i \mathbb{P}[\{\omega \in \Omega | X(\omega) \in A_i\}] = \sum_i \mathbb{P}_X[A_i],\end{aligned}$$

where the third equality follows from two facts: (1) X is a function, i.e., no element can belong to two disjoint images at the same time, so the huge expression corresponding to A_i are all disjoint, and (2) $\mathbb{P}$ itself satisfies additivity of disjoint events. □

Consider events of the form $\{X \leq a\}$ for real numbers a.

Definition 2.11

Given a random variable X, its **cumulative distribution function** (cdf) is a function $F_X : \mathbb{R} \to [0,1]$ defined by

$$F_X(a) = \mathbb{P}_X[(-\infty, a)] = \mathbb{P}[\{\omega \in \Omega | X(\omega) \leq a\}].$$

CDFs are good because they describe lots of things. For example,

$$\mathbb{P}[a < X \leq b] = \mathbb{P}_X[(a,b]] = \mathbb{P}_X[(-\infty, b]] - \mathbb{P}_X[(-\infty, a]] = F_X(b) - F_X(a).$$

(The inclusive/exclusive bounds here don't really matter).

Clarification on "random variable domain" vs "distribution".

Example 2.12

Let $\mathbb{P}$ be the uniform measure on $\Omega = [0,1]$. Define $X(x) = x^2$ for $x \in [0,1]$, and let $\mathbb{P}_X$ be the pushforward measure, or distribution of X.

In this example, $\mathbb{P}[[0, 1/4]] = 1/4$, since it is the uniform measure. On the other hand, $\mathbb{P}_X[[0, 1/4]] = 1/2$, since any $\omega \in \Omega$ satisfying $0 \leq \omega \leq 1/2$ satisfies $0 \leq \omega^2 \leq 1/4$.

Definition 2.13

Let X be a random variable and F_X its cdf. Then a function $f_X : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a **probability density function** (pdf) for X if for all a,

$$\int_{-\infty}^{a} f_X(x) dx = F_X(a).$$

The pdf is not always defined, for example, when the cdf is not differentiable. The fundamental theorem of calculus implies that $f_X(a) = F_X'(a)$.

Example 2.14

A random variable X is **Bernoulli** with parameter p if its domain is $\{0,1\}$ with $\mathbb{P}[X = 0] = 1 - p$ with $\mathbb{P}[X = 1] = p$. This we denote $X \sim \textsc{Bern}(p)$.

The cdf of $X \sim \text{Bern}(p)$ is given by

$$F_X(a) = \begin{cases} 0 & a < 0 \\ 1-p & 0 \le a < 1 \\ 1 & a \ge 1. \end{cases}$$

Example 2.15

A random variable X is **standard normal** if it has pdf

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

This we denote $X \sim N(0,1)$.

The cdf given by

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

does not have a nice closed form.

2.4 Multiple Random Variables

Definition 2.16

Given two continuous random variables X, Y defined on the same probability space, their **joint density** is a function $f_{X,Y}$ that satisfies

$$\mathbb{P}[a \le X \le b, c \le Y \le d] = \int_a^b \int_c^d f_{X,Y}(x,y)dydx.$$

For discrete random variables X, Y, an analogous quantity is

$$\mathbb{P}[X = x, Y = y] = f_{X,Y}(x,y).$$

Example 2.17

Let X, Y be independent, standard normal random variables on the same probability space. X and Y are said to be independent if their joint density factors as a product of their marginal distributions.

A few notes:

- Given a joint density function $f_{X,Y}$, the **marginal distributions** of X and Y are defined by integrating out the other variables, i.e.,

$$f_X(x) = \int f_{X,Y}(x,y)dy \qquad f_Y(y) = \int f_{X,Y}(x,y)dx.$$

- Integrating away other variables in order to obtain the marginal distrubtions for each individual random variable is a process called **marginalization**.
- In the context of Example 2.17, this definition of independence implies that

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) = \frac{1}{2\pi} e^{-x^2/2-y^2/2}.$$

Now let's deal with simple functions of multiple random variables.

Example 2.18
Let X, Y be independent and uniform on $[0,1]$ with joint density $f_{X,Y}(x,y) = 1$ for all $x, y \in [0,1]$. Define $Z = X + Y$. What is the distribution of Z?

$$F_Z(z) = \mathbb{P}[Z \le z] = \iint_{x+y\le z} f_{X,Y}(x,y)dxdy = \iint_{x+y\le z} 1\,dxdy.$$

This is the area of the intersection of the unit square with $x + y < z$ for constant z, so we find that $F_Z(z) = z^2/2$ when $0 \le z \le 1$, $F_Z(z) = 1 - (2-z)^2/2$ when $z \ge 1$, and $F_Z(z) = 1$ when $z \ge 2$. To calculate the pdf,

$$f_Z(z) = F_Z'(z) = \begin{cases} z & 0 \le z \le 1, \\ 2 - z & 1 \le z \le 2, \end{cases}$$

hence the name triangular distribution.

Example 2.19
Let X, Y be independent standard normal variables. Compute the distribution of $Z = X/Y$.

As before, let's compute the cdf.

$$F_Z(z) = \mathbb{P}[Z \le z] = \iint_{x/y \le z} f_{X,Y}(x,y)dxdy = \iint_{x/y \le z} \frac{1}{2\pi} e^{-x^2/2-y^2/2} dxdy.$$

Use the Jacobian to change variables to $a = x/y$ and $b = y$:

$$\frac{\partial(a,b)}{\partial(x,y)} = \left| \begin{pmatrix} \partial a/\partial x & \partial a/\partial y \\ \partial b/\partial x & \partial b/\partial y \end{pmatrix} \right| = \left| \begin{pmatrix} 1/y & -x/y^2 \\ 0 & 1 \end{pmatrix} \right| = 1/y,$$

so

$$\begin{aligned} F_Z(z) &= \iint_{a \le z} \frac{|b|}{2\pi} e^{-(ab)^2/2-b^2/2} dbda \\ &= \int_{-\infty}^{z} \int_{-\infty}^{\infty} \frac{|b|}{2\pi} e^{-(ab)^2/2-b^2/2} dbda \\ &= \frac{1}{2\pi} \int_{-\infty}^{z} \int_{0}^{\infty} \frac{2e^u}{a^2+1} dud a \\ &= \int_{-\infty}^{z} \frac{1}{\pi(a^2+1)} da. \end{aligned}$$

Note that this implies $f_Z(z) = 1/(\pi(z^2+1))$ is the pdf of Z. Finishing our computation,

$$F_Z(z) = \frac{1}{\pi} \arctan(z) + \frac{1}{2}.$$

It turns out that Z follows the **standard cauchy distribution**.

Definition 2.20

Given a nonnegative integer random variable X with pdf p_x, its **probability generating function** is

$$p(t) = \sum_{k=0}^{\infty} p_X(k) t^k = p_X(0) + t p_X(1) + t^2 p_X(2) + \dots$$

Lemma 2.21

If p, q, r are the probability generating functions of X, Y and $Z = X + Y$ respectively, then $r(t) = p(t) \cdot q(t)$.

Example 2.22

What positive integer labels can we give to two fair 6-sided dice such that the distribution of the sum of the rolls is the same as for two standard die?

For standard die C, D, their generating functions

$$p^C(t) = p^D(t) = \frac{1}{6}(t + t^2 + \ldots + t^6).$$

Therefore, the problem reduces to finding two polynomials $p^A(t)$, $p^B(t)$ satisfying

$$p^A(t)p^B(t) = p^C(t)p^D(t) = \frac{1}{36}t^2(t+1)^2(t^2-t+1)^2(t^2+t+1)^2.$$

2.5 Expectation

Definition 2.23

For a discrete random variable X, its **expectation** is

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

if the sum converges. For a continuous random variable X,

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

if it converges.

Theorem 2.24 (Linearity of Expectation)

Given random variables X_1, X_2, not necessarily independent, and constants c_1, c_2,

$$\mathbb{E}[c_1X_1 + c_2X_2] = c_1\mathbb{E}[X_1] + c_2\mathbb{E}[X_2].$$

Definition 2.25

Given a random variable X, its **variance** is defined as

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

In practice,

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

is an easier formula to use.

Definition 2.26

Given a random variable X, its **standard deviation** is defined as

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

Standard deviation is often easier to interpret than variance, because it has the same units as the original quantity X.

Lemma 2.27

For nonnegative X,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x]dx = \int_0^\infty 1 - F_X(x)dx.$$

Proof.

$$\begin{aligned}\int_0^\infty \mathbb{P}[X > x]dx &= \int_{x=0}^\infty \int_{y=x}^\infty f_X(y)dydx \\ &= \int_{y=0}^\infty \int_{x=0}^y f_X(y)dxdy \\ &= \int_{y=0}^\infty y f_X(y)dy = \mathbb{E}[X].\end{aligned}$$

□

This makes it easier to calculate expectation (in some cases).

Example 2.28

The **exponential distribution** with rate λ is given by the pdf

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Expectation calculation using integration by parts:

$$\mathbb{E}[X] = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = \left(-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right)\Big|_0^\infty = \frac{1}{\lambda}.$$

Expectation calculation using the trick from above:

$$\mathbb{E}[X] = \int_0^\infty \left(1 - \int_0^x \lambda e^{-\lambda x} dx\right) dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Variance calculation using integration by parts:

$$\mathbb{E}[X^2] = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x} dx = \left(-x^2 e^{-\lambda x} - \frac{2x}{\lambda} e^{-\lambda x} - \frac{2}{\lambda^2} e^{-\lambda x}\right)\Big|_0^\infty = \frac{2}{\lambda^2},$$

so $\mathrm{Var}[X] = 1/\lambda^2$.

3 January 17, 2023

3.1 Independence

Definition 3.1 (Independence)

Two events $A, B \in \mathcal{F}$ are independent if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B].$$

Alternatively, two random variables X, Y defined on the same probability space are said to be independent if, for any $A, B \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A] \cdot \mathbb{P}[Y \in A].$$

This definition is the same as the joint density definition that we applied last lecture.

Theorem 3.2

For random variables X, Y defined on the same probability space, the following are equivalent:

- $\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$ for every $A, B \in \mathcal{B}_{\mathbb{R}}$.
- $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for every $x, y \in \mathbb{R}$.

Proof. First assume $X \in A$ and $Y \in B$. Then set $A = [x + dx]$ and $B = [y + dy]$ for some small enough dx, dy such that $f_X, f_Y, f_{X,Y}$ are constant over A, B, and $A \times B$. Then,

$$\mathbb{P}[X \in A]\mathbb{P}[Y \in B] = \left(\int_A f_X(a)\mathrm{d}a\right)\left(\int_B f_Y(b)\mathrm{d}b\right) = (f_X(x)\mathrm{d}x)(f_Y(y)\mathrm{d}y).$$

But also,

$$\mathbb{P}[X \in A, Y \in B] = \iint_{(a,b)\in A\times B} f_{X,Y}(a,b)\mathrm{d}a\mathrm{d}b = f_{X,Y}\mathrm{d}x\mathrm{d}y,$$

which is enough to imply that $f_X f_Y = f_{X,Y}$. Conversely, assume that $f_X f_Y = f_{X,Y}$. Writing out the integral for $\mathbb{P}[X \in A, Y \in B]$, we can separate $f_{X,Y}$ into its two marginal distributions and then integrate separately, which gives $\mathbb{P}[X \in A, Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$. □

Now we introduce mutual independence, which is a way to deal with independence between more than just two random variables.

Definition 3.3

Random variables $X_1, \ldots, X_n$ are **mutually independent** if for any sets $A_1, \ldots, A_n \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}\left[\bigcap_i (X_i \in A_i)\right] = \prod_i \mathbb{P}[X_i \in A_i].$$

Example 3.4

Mutual independence implies pairwise independence.

Proof. WLOG, we show that X_1 and X_2 are pairwise independent. For all $i \geq 3$, let $A_i = \mathbb{R}$. Since $X_i \in A_i$ is always true for $i \geq 3$, this implies

$$\mathbb{P}\left[\bigcap_i (X_i \in A_i)\right] = \mathbb{P}[X_1 \in A_1, X_2 \in A_2].$$

Also, $\mathbb{P}[X_i \in A_i] = 1$ for all $i \geq 3$, so the product on the right is the same as $\mathbb{P}[X_1 \in A_1]\mathbb{P}[X_2 \in A_2]$. Thus, X_1 and X_2 are pairwise independent. □

Example 3.5

Pairwise independence does not imply mutual independence.

Let $X_1, X_2, X_3 \in \{0, 1\}$ be the results of three independent flips of a fair coin. Let $Y_1 = 0$ when $X_2 = X_3$ and $Y_1 = 1$ otherwise. Define Y_2, Y_3 analogously.

Y_i are not mutually independent, since $\mathbb{P}[Y_1 = Y_2 = Y_3 = 1] = 0$, while $\mathbb{P}[Y_1 = 1]\mathbb{P}[Y_2 = 1]\mathbb{P}[Y_3 = 1] = 1/8$.

On the other hand, Y_i are pairwise independent, since each pair of Y_i occurs with probability $1/4 = 1/2 \cdot 1/2$, which is also the product of their marginal distributions.

3.2 Conditional Probability

Definition 3.6

Given two events A, B, the conditional probability of A given B is defined as

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

Conditional probabilities when A and B are random variables is defined analogously.

Note that $\mathbb{P}[A|B] = \mathbb{P}[A]$ when A and B are independent.

Lemma 3.7

Conditional probabilities are probability measures.

Proof. Conditional probability satisfies the two laws governing probability mea-

sures:

- $0 \leq \mathbb{P}[A|B] \leq 1$.
- For disjoint events A_i,

$$\mathbb{P}\left[\left(\bigcup_i A_i\right)|B\right] = \frac{\mathbb{P}[(\cup_i A_i) \cap B]}{\mathbb{P}[B]} = \frac{\sum_i \mathbb{P}[A_i \cap B]}{\mathbb{P}[B]} = \sum_i \mathbb{P}[A_i|B].$$

□

Theorem 3.8 (Law of total probability)
Let Y be a random variable taking discrete values $y_1, \ldots, y_n$. For any event A,

$$P[X \in A] = \sum_{i=1}^{n} \mathbb{P}[Y = y_i]\mathbb{P}[X \in A|Y = y_i].$$

Proof. The events $\{Y = y_i\}$ partition the sample space, so

$$\sum_{i=1}^{n} \mathbb{P}[(X \in A) \cap (Y = y_i)] = \mathbb{P}[X \in A|(\cap_{i=1}^{n}\{Y = a_i\})] = \mathbb{P}[X \in A].$$

□

Theorem 3.9 (Bayes' rule)
For random variables X, Y,

$$\mathbb{P}[X|Y] = \frac{\mathbb{P}[Y|X]\mathbb{P}[X]}{\mathbb{P}[Y]}.$$

3.3 Inference

Bayes' rule is important to the field of **inference**, which is important for probability statistics, information theory, machine learning, etc. Inference involves inferring properties of some random variable (Y) via data (X).

Example 3.10

Most people have never had the same Uber driver twice. How can we use this observation to estimate the number of Uber drivers in Boston?

In this case,

- X is the observation that I have never had the same uber driver twice.
- Y is the number of Uber drivers in Boston.

Let's try guessing values for Y and seeing what happens.

Definition 3.11

Given data $X = x$ and random variable Y to be estimated, the **maximum likelihood estimation** for the value of Y is

$$\hat{y}_{MLE} = \arg\max_y p_{X|Y}(x|y).$$

Intuitively, we're calculating the probability that we observe the data x (i.e., I have never had the same Uber driver twice), over all possible values for the random variable Y (i.e., the number of Uber drivers in Boston). Then, the value of Y that gives us the greatest probability is our MLE.

If we're equally likely to get any driver on every ride, the probability of our observation assuming $Y = y$ is

$$p_{X|Y}(x|y) = \frac{y(y-1)\cdots(y-n+1)}{y^n}.$$

This function is increasing in y, so $\hat{y}_{MLE} = \infty$, which is not useful. To fix this, we can further impose a **prior** on Y.

Definition 3.12

Given data $X = x$ and random variable Y to be estimated, the **maximum a posteriori** (MAP) estimation for the value of Y is

$$\hat{y}_{MAP} = \arg\max_y \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)} = \arg\max_y p_{X|Y}(x|y)p_Y(y).$$

The prior is represented by the distribution $p_Y(y)$, and represents some prior belief that we have about the distribution of Y. For example, we know that Y must be less than the total number of people in Boston, which is something that can be captured by this prior distribution. The way our modified formula works is by Bayes' rule; we wish to maximize the probability of Y given X over all values of y, which is the quantity inside of the arg max.

One reasonable prior we might use for this scenario is the **log-normal distribution**, which has pdf

$$p_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right).$$

Another example to consider:

Example 3.13

Let $X_1, \ldots, X_n$ drawn independently from $\mathcal{N}(\mu, 1)$. Also, suppose that we know that μ should be concentrated around zero, so we can impose the modeling assumption $\mu \sim \mathcal{N}(0, 1)$. What is the posterior distribution $p_{\mu|X_1,\ldots,X_n}$?

Let $x_1, \ldots, x_n$, and μ_0, denote the data that we observe. Each X_i is drawn independently, so we have:

$$\begin{aligned} p_{X_1,\ldots,X_n|\mu}(x_1, \ldots, x_n|\mu_0) &= \prod_{i=1}^{n} p_{X_i|\mu}(x_i|\mu_0) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu_0)^2/2}. \end{aligned}$$

Our prior distribution is given by $p_\mu(\mu_0) \sim \mathcal{N}(0, 1)$. Also, the posterior function is a function of μ_0 and $x_1, \ldots, x_n$ only. Since each X_i represents our data, we can treat them like constant values with respect to our posterior distribution, and therefore ignore terms like $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n)$ (this value is constant for fixed data). Applying

Bayes' rule, we now have

$$\begin{aligned} p_{\mu|X_1,\dots,X_n}(\mu_0|x_1,\dots,x_n) &\propto p_{X_1,\dots X_n|\mu}(x_1,\dots,x_n|\mu_0)p_\mu(\mu_0) \\ &\propto \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2-\frac{1}{2}\mu_0^2\right) \\ &\propto \exp\left(\frac{n+1}{2}\left(\mu_0-\frac{\sum_{i=1}^{n}x_i}{n+1}\right)^2\right), \end{aligned}$$

which comes from expanding and completing the square with respect to μ_0 (and absorbing terms into the proportionality). This implies

$$p_{\mu|X_1,\dots,X_n}(\mu_0|x_1,\dots,x_n) \propto \mathcal{N}\left(\frac{n}{n+1}\overline{x},\frac{1}{n+1}\right).$$

Intuitively, this makes sense. With no prior, we expect the distribution of μ to be centered around $\overline{x}$, since this is the only data we are given. Given the prior, i.e., the expectation that μ is actually distributed normally around 0, the center of the posterior distribution is pulled closer to zero.

3.4 Discrete Conditioning

Definition 3.14

Two events A, B are conditionally independent given C if

$$\mathbb{P}[A\cap B|C]=\mathbb{P}[A|C]\cdot\mathbb{P}[B|C].$$

Analogously, let X, Y, Z be random variables. X and Y are conditionally independent given Z if for any $A, B, C \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}[X\in A, Y\in B|Z\in C]=\mathbb{P}[X\in A|Z\in C]\cdot\mathbb{P}[Y\in B|Z\in C].$$

Example 3.15

Consider the same coin flipping example as previously.

Y_1, Y_2 are conditionally independent given X_3, since knowing the result of X_3 provides no information in deducing Y_1 or Y_2. However, Y_1 and Y_2 are not conditionally independent given Y_3, since given $Y_3 = 1$, it is not possible for $Y_1 = 1$ and

$Y_2 = 1$.

Example 3.16
Conditional independence does not imply marginal independence.

Consider two coins, one fair and the other with two heads. Let Z denote a random choice of either coin. Then, let X, Y denote two flips of this coin. Given Z, X and Y are independent, so X and Y are conditionally independent on Z. On the other hand, they are not marginally independent; given that X is heads, the probability that the coin with two heads was chosen is 2/3, and therefore the probability that Y is also heads increases.

Example 3.17
Marginal independence does not imply conditional independence.

Let X and Y be the outcomes of two flips of a fair coin. If $Y = H$, Let $Z = X$; otherwise, let $Z =!X$. X and Z are marginally independent with no information about Y. On the other hand, given Y, X and Z are no longer independent, so X and Z are not conditionally independent.

Definition 3.18
Given discrete random variables X, Y, the **conditional expectation** of X given Y is

$$\mathbb{E}[X|Y] = \sum_{x \in X(\Omega)} x\mathbb{P}[X = x|Y].$$

The analogous statement for continuous random variables is

$$\mathbb{E}[Y|X = x] = \int_{\mathbb{R}} y f_{Y|X}(y, x)\mathrm{d}y,$$

where $f_{Y|X}(y, x)$ is the **conditional density** of Y given $X = x$.

The conditional density satisfies

$$f_{X,Y}(x, y) = f_{Y|X}(y, x)f_X(x) = f_{X|Y}(x, y)f_Y(y),$$

where $f_{X,Y}(x, y)$ is the joint density of X and Y. Also note that $\mathbb{E}[X|Y]$ is a function

of Y, so this quantity is itself also a random variable.

Definition 3.19
The **conditional variance** of X given Y is

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

Lemma 3.20 (Self-conditioning)
Given random variable X and deterministic function f, then

$$\mathbb{E}[f(X)|X] = f(X).$$

Proof. Assume $X = y$. Then

$$\mathbb{E}[f(X)|X = y] = \sum_{x \in f(X)(\Omega)} x\mathbb{P}[f(X) = x|X = y].$$

All probabilities are zero unless $x = f(y)$, in which case the probability is 1, so

$$\mathbb{E}[f(X)|X = y] = f(y).$$

Since this is true for any possible value of X, $\mathbb{E}[f(X)|X] = f(X)$, as desired. □

Proposition 3.21 (Conditional Linearity of Expectation)
For any real constants c_1, c_2 and random variables X_1, X_2, Y, linearity of conditional expectation holds. That is,

$$\mathbb{E}[c_1X_1 + c_2X_2|Y] = c_1\mathbb{E}[X_1|Y] + c_2\mathbb{E}[X_2|Y].$$

Example 3.22
Let X, Y be the results of two independent rolls of a fair 6-sided die, and let $Z = X + Y$. Compute $\mathbb{E}[X|Z]$ and $\mathrm{Var}[X|Z]$.

By symmetry, $\mathbb{E}[X|Z] = Z/2$. Now the variance:

$$\mathrm{Var}[X|Z] = \mathbb{E}[(X - \mathbb{E}[X|Z])^2|Z] = \mathbb{E}[(X - Z/2)^2|Z].$$

OK, now this is ugly. Given Z, X ranges on an interval from $\max(1, Z-6)$ to $\min(6, Z-1)$, inclusive. Say that this interval has start a and end $a+\ell-1$. Then $Z = 2a+\ell-1$, and summing $(X-Z/2)^2$ over this interval gives us

$$\sum_x (x-Z/2)^2 = \sum_{i=1}^{\ell} \left(\frac{-\ell+(2i-1)}{2}\right)^2 = 2\sum_{i=1}^{\ell-(2i-1)>0} \left(\frac{l-(2i-1)}{2}\right)^2.$$

If ℓ is odd, then we're just summing the first $((\ell-1)/2)$ squares, in which case

$$\sum_x (x-Z/2)^2 = 2\frac{(\ell-1)/2 \cdot \ell \cdot (\ell+1)/2}{6} = \frac{1}{12}\ell(\ell^2-1).$$

If ℓ is even, then

$$2\sum_{i=1}^{\ell-(2i-1)>0} \left(\frac{l-(2i-1)}{2}\right)^2 = \frac{1}{2}\sum_{i=1}^{\ell-(2i-1)>0} (\ell-(2i-1))^2,$$

which is the sum of the odd squares from 1 to $(\ell-1)$, which is also the sum of the first $(\ell-1)$ squares, minus the sum of the even squares from 2 to $(\ell-2)$. Therefore,

$$\sum_x (x-Z/2)^2 = \frac{1}{2}\left(\frac{\ell(\ell-1)(2\ell-1)}{6} - 4\cdot\frac{(\ell-2)/2\cdot(\ell/2)\cdot(\ell-1)}{6}\right) = \frac{1}{12}\ell(\ell^2-1).$$

The sum turns out to be the same in both cases. Since the length of the interval is ℓ, the expected value is $(\ell^2-1)/12$. Therefore,

$$\mathrm{Var}[X|Z] = \mathbb{E}[X-Z/2|Z] = ((\min(6,Z-1)-\max(1,Z-6)+1)^2-1)/12.$$

3.5 Continuous Conditioning

Polar coordinates are a thing. Let (X,Y) be drawn from a probability distribution on the plane, and define (R,Θ) so that

$$X = R\cos\Theta, Y = R\sin\theta, 0 \le R, 0 \le \Theta < 2\pi.$$

By the Jacobian,

$$p_{R,\Theta}(r,\theta) = rp_{X,Y}(x,y).$$

Example 3.23
Let (X, Y) be a randomly chosen point in the interior of the unit disc. Compute $\mathbb{E}[X^2 + Y^2]$.

We know $p_{X,Y}(x, y) = 1/\pi$ is uniform, so $p_{R,\Theta}(r, \theta) = r p_{X,Y}(r\cos\theta, r\sin\theta) = r/\pi$. Notice that this factors:

$$p_R(r) = 2r \quad \text{and} \quad p_\Theta(\theta) = \frac{1}{2\pi},$$

hence R and Θ are independent. Now, our expectation is

$$\mathbb{E}[X^2 + Y^2] = \mathbb{E}[R^2] = \int_0^1 r^2(2r\mathrm{d}r) = \frac{1}{2}.$$

Example 3.24
In the notation of the previous example, evaluate $\mathbb{E}[X^2 + Y^2|X]$.

By the self-conditioning Lemma, $\mathbb{E}[X^2|X] = X^2$, so it suffices to compute $\mathbb{E}[Y^2|X]$. This can be computed by first computing the density $f_{Y|X}(y, x) = f_{X,Y}(x, y) \cdot f_X(x)$, and then integrating

$$\mathbb{E}[Y^2|X] = \int y^2 f_{Y|X}(y, x)\mathrm{d}y.$$

Example 3.25
Compute the Beta integral

$$I_{a,b} = \int_0^1 x^a(1-x)^b\mathrm{d}x.$$

Let $X_1, \ldots, X_{a+b+1}$ be independent and identically distributed random variables in $[0, 1]$, all uniform. Let E be the event that X_1 is the $(a+1)$th smallest among the X_i. Then,

$$\mathbb{P}[E|X_1 = x] = \binom{a+b}{a} x^a(1-x)^b,$$

since $\mathbb{P}[X_i \le x] = x$ for uniform variables, and we need to choose a to be less than X_1. Also, the pdf of X_1 is constant, i.e., $p_{X_1}(x) = 1$, since X_1 is uniformly distributed.

Therefore, by the law of total probability,

$$\mathbb{P}[E] = \int_0^1 \mathbb{P}[E|X_1 = x] p_{X_1}(x) \mathrm{d}x = \binom{a+b}{a} I_{a,b}.$$

On the other hand, by symmetry, $\mathbb{P}[E] = 1/(a+b+1)$. Therefore,

$$I_{a,b} = \frac{1}{(a+b+1)\binom{a+b}{a}} = \frac{a!b!}{(a+b+1)!}.$$

4 January 19, 2023

4.1 Correlation

Definition 4.1

The covariance of two random variables X, Y is given by

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Approximately, covariance measures the strength of the linear relationship between X and Y, e.g., whether or not greater values of X corresponds to greater values of Y and vice versa. When $X = Y$, $\mathrm{Cov}(X,Y) = \mathrm{Var}(X)$. In some sense, it can be thought of as a weaker version of independence, since it captures linear dependence, but not dependence in full generality.

Lemma 4.2

If X, Y are independent, then $\mathrm{Cov}(X,Y) = 0$.

Proof. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ when X and Y are independent. □

Lemma 4.3

The converse is not true.

Proof. Let $X \sim U[-1,1]$ and $Y = X^2$ so that X, Y are not independent. On the other hand,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X^3] - \mathbb{E}[X]\mathbb{E}[X^2] = 0.$$

□

Covariance allows us to capture linear sums of variances:

Lemma 4.4

For random variables X, Y,

$$\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}(X,Y).$$

Proof.

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

□

Properties of covariance:

- $\text{Cov}(X,a) = 0$
- $\text{Cov}(X,X) = \text{Var}[X]$
- $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
- Covariance is bilinear: $\text{Cov}(aX+bY, cZ+dW) = ac\,\text{Cov}(X,Z) + ad\,\text{Cov}(X,W) + bc\,\text{Cov}(Y,Z) + bd\,\text{Cov}(Y,W)$.

Example 4.5

Let S be drawn uniformly at random among all subsets of size k of $[n]$. For each $1 \le i \le n$, $X_i = 1$ if $i \in S$ and $X_i = 0$ otherwise. Find $\text{Cov}(X_1, X_2)$.

By symmetry, $\text{Cov}(X_i, X_j)$ are equal for all $i \neq j$. Also, $X_i \sim \text{Bern}(k/n)$, so $\text{Var}[X_i] = k/n - (k/n)^2 = k(n-k)/n^2$. Using $\text{Var}[a] = 0$,

$$0 = \text{Var}[k] = \text{Var}[X_1 + \ldots + X_n] = n\,\text{Var}[X_1] + 2\binom{n}{2}\text{Cov}(X_1, X_2).$$

Therefore,

$$\text{Cov}(X_1, X_2) = \frac{-k(n-k)}{n^2(n-1)}.$$

Intuitively, it makes sense that this quantity is negative, since, if we are given that $X_1 = 1$, the probability that X_2 is also 1 decreases.

Definition 4.6

The **correlation** between X, Y is defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X] \cdot \mathrm{Var}[Y]}}.$$

Correlation can be thought of as a normalized version of covariance, in the sense that correlation is immune to scaling: $\rho(aX, bY) = \rho(X,Y)$ for constants a, b. Correlation is also dimensionless.

Lemma 4.7 (Translation invariance)

For constants a, b, $\rho(X + a, Y + b) = \rho(X,Y)$.

Lemma 4.8

For any random variables X, Y, $|\rho(X,Y)| \le 1$.

Proof. By translation invariance, shift everything so that $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Then $\mathrm{Cov}(X,Y) = \mathbb{E}[XY]$, $\mathrm{Var}[X] = \mathbb{E}[X^2]$, and $\mathrm{Var}[Y] = \mathbb{E}[Y^2]$, so we want to show $\mathbb{E}[XY] \le \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. To do so, we can apply Cauchy-Schwarz to the pair of functions $(x\sqrt{p_{X,Y}(x,y)}, y\sqrt{p_{X,Y}(x,y)})$:

$$\begin{aligned}
\mathbb{E}[XY] &= \iint_{\mathbb{R}^2} xy\, p_{X,Y}(x,y) \\
&\le \sqrt{\left(\iint_{\mathbb{R}^2} x^2 p_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x\right)\left(\iint_{\mathbb{R}^2} y^2 p_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x\right)} \\
&= \sqrt{\left(\iint_{\mathbb{R}} x^2 p_X(x)\mathrm{d}x\right)\left(\iint_{\mathbb{R}} y^2 p_Y(y)\mathrm{d}y\right)} \\
&= \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.
\end{aligned}$$

□

Example 4.9

Let A, B, C be random variables with $\rho(A,B) = \rho(A,C) = 0.5$. What are the possible values of $\rho(B,C)$?

Shift and scale so that $\sigma_A = \sigma_B = \sigma_C = 0$ and $\mathrm{Var}(A) = \mathrm{Var}(B) = \mathrm{Var}(C) = 1$. When $B = C$, $\rho(B,C) = \mathrm{Cov}(B,C)/\mathrm{Var}(B) = 1$. Also, if we consider the random variable $B + C - A$:

$$\begin{aligned}\mathrm{Var}[B+C-A] &= \mathrm{Var}[B] + \mathrm{Var}[C] + \mathrm{Var}[-A] + \\ &\qquad 2(\mathrm{Cov}(B,C) + \mathrm{Cov}(B,-A) + \mathrm{Cov}(C,-A)) \\ &= 3 + 2(\rho(B,C) - \rho(A,B) - \rho(A,c)) \\ &= 1 + 2\rho(B,C).\end{aligned}$$

Since $\mathrm{Var}[B+C-A] \geq 0$, $\rho(B,C) \geq -1/2$. To show that all values in between are attainable, let $A \sim N(0,1)$, $X \sim N(0,t)$, and $Y \sim N(0,1-t)$ to be independent for $t \leq 1$. Then, let

$$B = \frac{A}{2} + \frac{\sqrt{3}}{2}(X+Y), \quad C = \frac{A}{2} + \frac{\sqrt{3}}{2}(X-Y).$$

Note that $\mathbb{E}[B] = \mathbb{E}[C] = 0$. Also, since every variable is chosen independnetly, linearity of variance implies $\mathrm{Var}(B) = \mathrm{Var}(C) = (1/2)^2 + (\sqrt{3}/2)^2 = 1$. Now,

$$\begin{aligned}\rho(B,C) &= \mathrm{Cov}(B,C) \\ &= \mathbb{E}\left[\left(\frac{A}{2} + \frac{\sqrt{3}}{2}(X+Y)\right)\left(\frac{A}{2} + \frac{\sqrt{3}}{2}(X-Y)\right)\right] \\ &= \frac{3t-1}{2}.\end{aligned}$$

By construction, this works for any $0 \leq t \leq 1$, so we're done.

Definition 4.10

Given random variables $X_1, \ldots, X_n$, their **covariance matrix** is defined as the $n \times n$ matrix K, whose entry K_{ij} is defined to be $\mathrm{Cov}(X_i, X_j)$.

- Diagonal entries are the variances of each X_i.
- The covariance matrix is positive semidefinite.

Definition 4.11

A real $n \times n$ matrix A is positive-semidefinite if, for any vector x, the quantity $x^t Ax$ is nonnegative. This is equivalent to A having nonnegative eigenvalues.

4.2 The indicator method

Theorem 4.12 (Linearity of Expectation)

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

Proof.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \iint (x+y) p_{x,y}(x,y) \mathrm{d}x\mathrm{d}y \\
&= \iint x p_{x,y}(x,y) \mathrm{d}x\mathrm{d}y + \iint y p_{x,y}(x,y) \mathrm{d}x\mathrm{d}y \\
&= \int x p_x \mathrm{d}x + \int y p_y \mathrm{d}y \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

□

Example 4.13

n people put their hats into a bag. They take turns drawing a random hat from a bag. Compute the expectation and variance of the number of people who get their original hat back.

To compute the variance, we need $\mathbb{E}[X^2]$. We can compute this using a decomposition trick.

$$\mathbb{E}[(X_1 + \ldots + X_n)^2] = \sum_{1 \le i,j \le n} \mathbb{E}[X_i X_j].$$

When $i \neq j$, $\mathbb{P}[X_i X_j] = 1$ is $1/(n \cdot (n-1))$. When $i = j$, $\mathbb{P}[X_i X_j] = 1/n$. Therefore, $\mathbb{E}[X^2] = n(n-1)/(n(n-1)) + n/n = 2$, so $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1$.

Example 4.14 (Coupon collector's problem)
Consider a cereal box contest in which each box of cereal contains one of n different types of coupons, and one must collect one of every coupon to win. Let T be the number of boxes we must open to win. Compute $\mathbb{E}[T]$.

Let Z_k be the amount of time it takes to aquire the $(k+1)^{\text{th}}$ new coupon. Note that $\mathbb{E}[Z_k] = n/(n-k)$, since on each draw the probability that you get a new type, having already collected k types, is $(n-k)/n$. Also, $T = \sum Z_k$, so we can use linearity of expectation to get

$$\mathbb{E}[T] = \mathbb{E}\left[\sum_{k=1}^{n-1} Z_k\right] = n\sum_{k=1}^{n-1}\frac{1}{k} \approx n\log n.$$

4.3 Results on conditional expectations

Lemma 4.15
For independent random variables X, Y and any deterministic function f,

$$\mathbb{E}[Yf(X)|X] = f(X)\mathbb{E}[Y].$$

Proof. When X and Y are independent, $f_{Y|X}(y|x) = f_Y(y)$. So,

$$\begin{aligned}\mathbb{E}[Yf(X)|X = x] &= \int_{\mathbb{R}} yf(x)f_Y(y)\mathrm{d}y \\ &= f(x)\mathbb{E}[Y].\end{aligned}$$

Since this holds for any $x \in X(\Omega)$, we are done. □

Lemma 4.16
For random variables X, Y and deterministic function f,

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

Remember that $\mathbb{E}[X|Y]$ is a function of Y, so it is itself a random variable, and we may therefore compute the expectation and variance of this quantity.

The next two theorems are dubbed the **tower laws**.

Theorem 4.17 (Law of total expectation)
For random variables X, Y where $\mathbb{E}[X]$ is finite,

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

This can be generalized:

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

for deterministic function f.

Recall that $\mathbb{E}[X|Y]$ is the expected value of X given a prior Y. The law of total expectation says that the best prediction that we can make for X across all possible priors is $\mathbb{E}[X]$, which is the same as the value that we predict X to have with no priors.

Theorem 4.18 (Law of total variance)
For random variables X, Y with finite $\text{Var}[X]$,

$$\text{Var}[X] = \text{Var}_Y[\mathbb{E}_X[X|Y]] + \mathbb{E}_Y[\text{Var}_X[X|Y]].$$

(Subscripts denote what to take expectation/variance over).

Example 4.19
Start with a distribution X with mean μ and standard deviation σ. Then, raise μ by 20% with probability 0.5. Compute the expected value and variance of X after this takes place.

Let Y be the distribution of X after applying the changes.

$$\mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y|X]] = \mathbb{E}_X[1.1X] = 1.1\mu.$$

Using the law of total variance,

$$\text{Var}[Y] = 0.01\mu^2 + 1.22\sigma^2.$$

4.4 Moment generating functions

The expression $\mathbb{E}[X^k]$ is called the **k-th moment** of X. The first moment of X is its mean. The second moment is the variance (when the mean is normalized). The third and fourth moments are called skew and kurtosis. Each moment is significant in some way.

Definition 4.20

The moment generating function (mgf) of a random variable X is defined to be the function

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

for values of t where the expectation is defined. If the only such value is $t = 0$, then X does not have an mgf.

Lemma 4.21

For each k,

$$\mathbb{E}[X^k] = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M(t)\bigg|_{t=0} = M^{(k)}(0).$$

Proof.

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{i\geq 0} \frac{(tX)^i}{i!}\right].$$

Taking k derivatives and setting $t = 0$, everything dies except for X^k, hence done. □

Lemma 4.22

For independent X, Y,

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

for all t.

Proof. X, Y independent implies that e^{tX} and e^{tY} are also independent, given fixed t. Therefore,

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

□

Example 4.23

Let X follow the geometric distribution with parameter p, i.e.,

$$\mathbb{P}[X = k] = (1-p)^{k-1}p$$

for $k \geq 1$. Compute $\text{Var}(X)$.

The key here is to know how to cleanly evaluate $\mathbb{E}[X^2]$. One approach is to compute the moment generating function $M_X(t)$ as follows:

$$M_X(t) = \mathbb{E}[e^{tX}] = \sum_{k\geq 1}(1-p)^{k-1}pe^{kt} = \frac{pe^t}{1-(1-p)e^t}.$$

Then, take some derivatives. The other approach is to use a nested geometric series. Both are ugly.

5 January 24, 2023

5.1 Stochastic Processes

Definition 5.1

A **stochastic process** $\{X_t\}_{t\in T}$ is a collection of random variables X_i, where the index t is some element of an index set T. For continuous stochastic processes, T is often $\mathbb{R}_{\geq 0}$, and for the discrete processes, T is often $\mathbb{Z}_{\geq 0}$.

Definition 5.2 (Sojourn time)

For some stochastic process $\{X_t\}_{t\in T}$, let $S = \{W_1, W_2, \ldots\}$ be the set of indices at which X_{W_i} is equal to a predetermined value K. Each W_m is called a **waiting time**, and represents the duration of time between the beginning of the process and the mth success. Each gap between waiting times $S_m = W_m - W_{m-1}$ is called a **Sojourn time**.

5.2 Counting distributions

Definition 5.3

A **Bernoulli random variable** $X \sim \text{Bern}(p)$, with $0 \leq p \leq 1$, has discrete pmf $f_X(0) = 1 - p$, $f_X(1) = p$.

In other words, a beroulli random variable with paramater p represents a binary outcome with probability p of resolving successfully. Naturally, $\mathbb{E}[X] = p$ and $\text{Var}[X] = p - p^2$.

Example 5.4

Two weighted coins have probability p, q of landing heads. After the first coin is flipped, the second coin is flipped only if the first coin was heads; otherwise, it is not flipped. Compute the variance in the outcome of the second coin.

The outcome of the second coin being heads or not can be modelled by a Bernoulli random variable with parameter pq. Therefore, the variance is $pq(1 - pq)$.

Definition 5.5

A **Bernoulli process** is a discrete-time stochastic process of finite or infinite i.i.d (independent identically distributed) Bernoulli random variables.

For example, a Bernoulli process may be a sequence of coin flips. The famous binomial distribution measures the number of successes in a Bernoulli process.

Definition 5.6

A **binomial random variable** $X \sim B(n, p)$ has discrete pmf

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $k \in \{0, \ldots, n\}$.

Binomial random variables measure Bernoulli processes of length n, whose individual Bernoulli trials all have parameter p. Since each Bernoulli trial is independent, linearity of expectation and linearity of variance implies that $\mathbb{E}[X] = np$,

$\text{Var}[X] = np(1-p)$.

Lemma 5.7

Let $X \sim B(n,p)$ and $Y \sim B(m,p)$ be independent binomial random variables. Then $Z = X + Y$ is distributed as $B(n+m,p)$.

Proof.

$$\begin{aligned} f_Z(k) &= \sum_{i=0}^{k} f_X(i) f_Y(k-i) \\ &= \sum_{i=0}^{k} \left(\binom{n}{i} p^i (1-p)^{n-i} \right) \left(\binom{m}{k-i} p^{k-i} (1-p)^{m-k+i} \right) \\ &= p^k (1-p)^{n+m-k} \sum_{i=0}^{k} \binom{n}{i} \binom{m}{k-i} \\ &= \binom{n+m}{k} p^k (1-p)^{n+m-k}. \end{aligned}$$

□

Lemma 5.8

Let $X \sim B(n,p)$ and $Y \sim B(X,q)$ be independent binomial random variables. Then $Y \sim B(n,pq)$.

Proof. Intuitively, this setup is the same as the following: flip a coin n times with probability p of getting heads. For each head, flip another coin with probability q of getting heads. Since you only get both heads with probability pq, it makes sense that Y is the same as $B(n,pq)$. More rigorously,

$$\begin{aligned}
\mathbb{P}[Y=m] &= \sum_{k=0}^{n} \mathbb{P}[X=k]\cdot\mathbb{P}[Y=m|X=k] \\
&= \sum_{k=m}^{n} \binom{n}{k} p^k (1-p)^{n-k} \binom{k}{m} q^m (1-q)^{k-m} \\
&= \sum_{k=m}^{n} \binom{n}{m}\binom{n-m}{k-m} p^k (1-p)^{n-k} q^m (1-q)^{k-m} \\
&= \dots \\
&= \binom{n}{m} (pq)^m (1-pq)^n.
\end{aligned}$$

□

Definition 5.9

A **hypergeometric** random variable $X \sim \text{Hypergeom}(N, K, n)$ models a sequence of Bernoulli trials that takes place *without replacement*. This is in contrast to binomial random variables, which model Bernoulli trials *with replacement*. The parameters represent the drawing of n objects out of total possible N different objects, where K of them represent "successful" objects.

The discrete pmf is given by

$$f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

This expression represents the probability that you select k successful objects out of K possible successful objects and $n-k$ non-successful objects out of $N-K$ possible non-successful objects.

We also have that

$$\mathbb{E}[X] = n\cdot\frac{K}{N} \quad \text{and} \quad \text{Var}[X] = \frac{nK(N-K)(N-n)}{N^2(N-1)}.$$

Example 5.10

Acceptance sampling is a techniqued used for quality control. The process of acceptance sampling involves drawing smaller samples from a larger pool of objects, and accepting or rejecting the entire pool of objects based on the result of the smaller sample. For example, choosing to accept or reject a lot of 1000 toy trucks based on a smaller sample of 100. Given that 50 of these trucks are defective, the probability that we see more than 5 in our sample can be modeled by a hypergeometric distribution.

5.3 Waiting Times

Definition 5.11

Geometric distrubtions model the **Sojourn times** in a Bernoulli process. That is, they measure the times between consecutive successes.

Definition 5.12

Let $X \sim \text{Geom}(p)$. Then the discrete pmf

$$f_X(k) = (1-p)^{k-1}p.$$

The discrete cdf is given by

$$F_X(k) = \sum_{i=1}^{k}(1-p)^i p = 1-(1-p)^k.$$

Note that this definition is inclusive on the first successful trial itself, i.e., $f_X(k)$ represents the probability that the first $k-1$ trials fail, and the kth trial is a success.

Lemma 5.13 (Memorylessness)

For geometric random variable X,

$$\mathbb{P}[X > n+m | X > n] = \mathbb{P}[X > m].$$

Lemma 5.14

Let $X_1, \ldots, X_n$ be geometric random variables distributed as $X_i \sim \textsc{Geom}(p_i)$. If $Y = \min_{1 \le k \le n} X_k$, then

$$Y \sim \textsc{Geom}\left(1 - \prod_{i=1}^{n} (1 - p_i)\right)$$

Proof. The probability that at least one event happens is equal to the complement of none of them happening. □

Definition 5.15

Let $X \sim \mathrm{NB}(r, p)$ be a **negative binomial** random variable. Then the discrete pmf

$$f_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}.$$

Negative binomial distributions measure the value of W_r, i.e., the waiting time for the r-th success.

Lemma 5.16

$$\mathrm{NB}(r, p) \sim \sum_{i=1}^{r} \textsc{Geom}(p).$$

Proof. Negative binomial distributions measures total waiting times, while geometric distributions measure the time between each waiting time. By the memorylessness property of geometric distributions, Sojourn waiting times are independent. □

By linearity of expectation and variance (for independent random variables), this implies

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

when $X \sim \mathrm{NB}(r, p)$.

5.4 Continuous Time Distributions

Here we explore the continuous analogues of binomial and geometric distributions: Poisson and Exponential distributions, respectively.

Definition 5.17

A **poisson** random variable $X \sim \text{Pois}(\lambda)$ is said to have "rate" $\lambda > 0$ and discrete pmf

$$f_X(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Suppose $X \sim \text{B}(n,p)$. Now fix $\lambda = np$ (recall that this is the expected number of successes over some number of trials), and increase n arbitrarily large; in the limit, $n \to \infty$ and $p \to 0$. Now,

$$\begin{aligned} f_X(k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{k!} p^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \end{aligned}$$

hence we recover the Poisson distribution. This is considered to be the continuous time limit of the binomial distribution in the sense that we maintain the mean number of successful trials in a given time period, but the Poisson distribution is running infinitely many trials. Given $X \sim \text{Pois}(\lambda)$, we also have

$$\mathbb{E}[X] = \text{Var}[X] = \lambda,$$

which makes sense intuitively, since $\mathbb{E}[X] = np = \lambda$ and $\text{Var}[X] = np(1-p) = \lambda$ in the limit when $n \to \infty$ and $p \to 0$.

Example 5.18 (binomial approximation with poisson)

When n is large, p is small, and $\lambda = np$ is medium-sized, using $Y \sim \text{Pois}(np)$ can be a good approximation for $X \sim \text{B}(n,p)$. In general, it's a lot easier to calculate $f_Y(k)$ vs. $f_X(k)$ for any particular k.

Lemma 5.19

Let $X_i \sim \text{Pois}(\lambda_i)$ be independent. Then $Y = \sum_{i=1}^n X_i \sim \text{Pois}(\sum_{i=1}^n \lambda_i)$.

Proof.

$$\begin{aligned}\mathbb{P}[X_1 + \ldots + X_n = k] &= \sum_{x_1+\ldots+x_n=k} \prod_{i=1}^n \left(\frac{\lambda_i^{x_i} e^{-\lambda_i}}{x_i!}\right) \\ &= \sum_{x_1+\ldots+x_n=k} \frac{e^{-\lambda_1-\ldots-\lambda_n}}{k!} \binom{k}{x_1,\ldots,x_n} \prod_{i=1}^n \lambda_i^{x_i} \\ &= \frac{e^{-\lambda_1-\ldots-\lambda_n}}{k!} (\lambda_1 + \ldots \lambda_n)^k.\end{aligned}$$

□

Definition 5.20

A **Poisson process** with rate $\lambda > 0$ is defined over the positive reals satisfying:

(1) For any strictly increasing non-negative $t_0, t_1, \ldots, t_m$, $X(t_{i+1}) - X(t_i)$ are all independent random variables.

(2) The random variable $X(s+t) - X(s) \sim \text{Pois}(\lambda t)$ (for $t > 0$). In words, the distribution of the number of events along any interval only depends on the length of the interval.

(3) $X(0) = 0$.

This definition is equivalent to the following reformulation:

Definition 5.21

Let $N(a, b]$ be a random variable that counts the number of events along the interval $(a, b]$. N is a **Poisson point process** with rate λ if:

(1) For any strictly increasing non-negative $t_0, \ldots, t_m$, $N(t_i, t_{i+1}]$ are all independent.

(2) $N(s, t] \sim \text{Pois}(\lambda(t-s))$ (for $t > 0$).

Some key facts about Poisson processes:

- Given the total number of events that occur along an interval, the distribution of those events along the interval is uniformly distributed.
- Due to the first fact, given the number of events that occur along an interval, the number of events that occur along subintervals of that interval follows a binomial distribution. For example, if you know that 10 events occurred 4 times in one hour, the probability that 3 of them occurred in the first 15 minutes is given by $\binom{10}{3}(.25)^3(.75)^7$.
- **Poisson Splitting**: If $\{X(t)\}$ is a Poisson process with rate λ, say we "split" each successful event into a type 1 event with probability p, and a type 2 event with probability $(1-p)$. Then, the processes that contains only events of type 1 is also a Poisson process with rate λp. The process that contains only events of type 2 is a Poisson process with rate $\lambda(1-p)$
- **Poisson Superposition**: Let $\{X_1(t),\ldots,X_n(t)\}$ be Poisson processes with rates $\lambda_1,\ldots,\lambda_n$, respectively. The union of all of these processes is also a Poisson process with rate $\lambda_1+\ldots+\lambda_n$.

Finally, the exponential distribution is the last distribution we cover. Like the Poisson distribution is the continuous analogue of the binomial distribution, the exponential distribution is the continuous analogue of the geometric distribution.

Definition 5.22

An exponential random variable $X \sim \text{Exp}(\lambda)$ has continuous pdf

$$f_X(k) = \lambda e^{-\lambda k}.$$

It has cdf

$$F_X(k) = 1 - e^{-\lambda k}.$$

6 January 26, 2023

7 January 31, 2023

The goal of asymptotics is to study the limiting behavior of random variables. For example, does the mean of independent observations of the same random variable

converge? What is the approximate distribution of sample means if we have a lot of different samples?

7.1 Modes of Convergence

Definition 7.1 (Convergence in Distribution)
Consider a sequence of random variables $X_1, \ldots$, and another random variable X. Let $F(x)$ denote the cdf of X, and $F_n(x)$ for X_n. We say that X_n converges in distribution to X if, for every x at which F is continuous,

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

Usually, this is denoted

$$X_n \xrightarrow[n\to\infty]{d} X.$$

Example 7.2
Let $Y_1, \ldots$ be a sequence of independent random variables distributed uniformly on $[0,1]$. Let $X_n = \max_{1\le i\le n} Y_i$. Then the random variable $n(1-X_n)$ converges in distribution to a random variable with distribution $\text{Exp}(1)$.

To show that this is true, we first compute the cdf of the random variable that is converging.

$$\begin{aligned} F_n(x) &= \mathbb{P}[n(1-X_n) \le x] \\ &= \mathbb{P}[X_n \ge 1 - x/n] \\ &= 1 - \mathbb{P}[X_n \le 1 - x/n] \\ &= 1 - \left(1 - \frac{x}{n}\right)^n. \end{aligned}$$

In the limit, this is equal to $1 - e^{-x}$, which is the cdf for a random variable with distribution $\text{Exp}(1)$, so we are done.

Convergence in distribution was only concerned about the long-run behavior of the cdf. Convergence in probability is stronger in the sense that the random variables must also get close to the target random variable.

Definition 7.3 (Convergence in Probability)

Consider a sequence of random variables $X_1, \ldots$, and another random variable X. We say that X_n converges in probability to X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

Usually, this is denoted

$$X_n \xrightarrow[n\to\infty]{p} X.$$

Example 7.4

Consider the sequence of random variables $X_1, \ldots$ where $X_n \sim \text{Exp}(n)$. Then, $X_n \xrightarrow[n\to\infty]{p} 0$.

Note that X_n is nonnegative, so

$$\mathbb{P}[|X_n - 0| > \varepsilon] = \int_{\epsilon}^{\infty} ne^{-nx} dx = e^{-n\varepsilon}.$$

Taking the limit as n goes to infinity, this approaches zero.

Example 7.5

Convergence in distribution does not imply convergence in probability. Define X_n with $X_0 = \text{Unif}[0,1]$, and

$$X_n = \begin{cases} X_0 & n \text{ even} \\ 1 - X_0 & n \text{ odd.} \end{cases}$$

This sequence converges in distribution to X_0, since they're all uniformly distributed. On the other hand, they don't converge in probability to any random variable X. Intuitively, this is because the sequence oscillates, whereas any random variable represents a function that outputs one value.

Definition 7.6 (Almost sure convergence)

A sequence of random variables $X_1, \ldots,$ converges almost surely to a random variable X if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

Usually, this is denoted

$$X_n \xrightarrow[n\to\infty]{a.s.} X.$$

This is the strongest sense of convergence. Almost sure convergence implies convergence in probability, which implies convergence in distribution.

Example 7.7

Consider the sequence of random variables $X_1, \ldots$ where $X_n = 0$ with probability $1/n^2$ and $X_n = 1$ otherwise. Then, $X_n \xrightarrow[n\to\infty]{a.s.} 1$.

The Borel-Cantelli Lemma gives us that: if sum of the probabilities of an infinite number of events is infinite, then the probability of infinitely many of them occurring is 1. On the other hand, if their sum is finite, then the probability of infinitely many of them *not* occuring is 1. In this case, since $\sum 1/n^2$ is finite, the probability that infinitely many X_i are 0 is 0. This implies that, with probability 1, there are infinitely many $X_i = 1$, so X_n converges almost surely to 1.

Example 7.8

Almost sure convergence does not imply convergence in probability. Consider the previous example, modified so that $X_n = 0$ with probability $1/n$ and $X_n = 1$ otherwise.

X_n converges in probability to 0, since $\lim_{n\to\infty} 1/n = 0$. On the other hand, the Borel-Cantelli Lemma implies that X_n does not converge almost surely to 0.

Definition 7.9 (Convergence in L^p norm)
A sequence of random variables $X_1, \ldots$ converges in the L^p-norm to a random variable X, for some $p \geq 1$, if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0,$$

given that their respective p-th absolute moments exist.

Lemma 7.10
Convergence in the L^q norm implies convergence in the L^p norm for $q > p$.

Proof. Note that $x^{p/q}$ is concave down. Therefore, by Jensen's,

$$\mathbb{E}[|X_n - X|^p] = \mathbb{E}[(|X_n - X|^q)^{p/q}] \leq \mathbb{E}[(|X_n - X|)^q]^{p/q}.$$

Taking the limit on both sides finishes. □

Lemma 7.11
Convergence in the L^p norm implies convergence in probability.

Proof. By Markov's inequality,

$$\mathbb{P}[|X_n - X| > \varepsilon] = \mathbb{P}[|X_n - X|^p > \varepsilon^p] \leq \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p}.$$

Taking the limit finishes. □

There isn't a good way to compare convergence in the L^p norm and convergence almost surely. (?)

7.2 Law of Large Numbers

Intuitively, when we are trying to determine the mean of a random variable X, taking more samples will generally give a better estimate for the mean. Why is this the case?

Theorem 7.12 (Weak Law of Large Numbers)
Consider a sequence of independent and identically distributed random variables $X_1, \ldots$ with finite expectation $\mathbb{E}[X]$. Denote the sample mean

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_n.$$

Then, $\overline{X_n} \xrightarrow[n\to\infty]{p} \mathbb{E}[X]$. This is equivalent to saying that for any $\varepsilon, \delta > 0$, there exists n where

$$\mathbb{P}[|\overline{X_n} - \mathbb{E}[X]| > \varepsilon] < \delta.$$

Example 7.13
Let $X_1, \ldots,$ be a sequence of independent random variables with $\mathbb{E}[X_i] = \mu$ and the variance of each term is finite. Let $Y_k = X_k X_{k+1}$ for every k. Prove that

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow[n\to\infty]{p} \mu^2.$$

By the weak law of large numbers,

$$\frac{2}{n}(Y_1 + Y_3 + \ldots) \xrightarrow[n\to\infty]{p} \mu^2$$
$$\frac{2}{n}(Y_2 + Y_4 + \ldots) \xrightarrow[n\to\infty]{p} \mu^2.$$

Averaging gives the desired result. Note that splitting the sum like this is necessary, since we require every "sample" to be independent.

Theorem 7.14 (Strong Law of Large Numbers)
Consider a sequence of independent and identically distributed random variables $X_1, \ldots$ with finite expectation $\mathbb{E}[X]$. Then $\overline{X_n} \xrightarrow[n\to\infty]{a.s.} \mathbb{E}[X]$.

With probability one, the sample mean approaches the actual mean.

7.3 Central Limit Theorem

Theorem 7.15 (Central Limit Theorem)
For a sequence of independent, identically distributed random variables $X_1, \ldots$, with finite expectation $\mathbb{E}[X]$ and variance σ^2, it holds that

$$\frac{\sqrt{n}(\overline{X_n} - \mathbb{E}[X])}{\sigma} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1).$$

Important: this holds for any random variable. The sample means are normally distributed, always. Intuitively, the more samples you have, the variance decreases from the mean ($\sigma^2 \to \sigma^2/n$).

Example 7.16
We roll a fair die 100 times. What is the probability that the average roll is at most 3?

By the central limit theorem, we can the sample means to be normally distributed. The variance of a dice roll is $35/12$, so $\sigma = \sqrt{35/12}$. The z-statistic is therefore given by

$$z = \frac{\overline{X_n} - \mathbb{E}[X]}{\sigma/\sqrt{n}} \approx -2.9.$$

Therefore,

$$\mathbb{P}[\overline{X} \le 3] = \mathbb{P}[z \le -2.9] = 0.17\%.$$

Example 7.17
Consider 10 numbers $X_1, \ldots, X_{10}$ drawn independently and uniformly at random from $[0, 100]$. Let A be their sum, rounded. Let B be the sum of their rounded values. Use the central limit theorem to approximate the probability that $A = B$.

Write $X_i = K_i + \varepsilon_i$, where K_i is X_i rounded to the nearest integer. ε_i is uniform, so

$$A = B \iff \varepsilon_1 + \ldots + \varepsilon_{10} \in [-0.5, 0.5].$$

By the central limit theorem,

$$\sum_{i=1}^{n} \varepsilon_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0, n\sigma^2),$$

so we may approximate our probability as

$$\mathbb{P}[-0.5 \le \mathcal{N}(0, 5/6) \le 0.5] \approx 0.42.$$

7.4 Slutsky and Borel-Cantelli

Theorem 7.18 (Slutsky's Theorem)

Consider two sequences of random variables X_n and Y_n, where $X_n \xrightarrow[n\to\infty]{d} X$ and $Y_n \xrightarrow[n\to\infty]{p} c$, where c is a constant. Then,

(1) $X_n + Y_n \xrightarrow[n\to\infty]{d} X + c$

(2) $X_n Y_n \xrightarrow[n\to\infty]{d} cX$

(3) $X_n / Y_n \xrightarrow[n\to\infty]{d} X_n / c$, if $c \neq 0$.

These results should feel intuitive.

Example 7.19

Let $X_i \sim \text{Unif}[-1, 1]$. Let

$$Z_n = \frac{\sqrt{n}\sum_{k=1}^{n} X_k}{\sum_{k=1}^{n}(X_k^2 + X_k^3)}.$$

Prove that $Z_n \xrightarrow[n\to\infty]{d} \mathcal{N}(0, 3)$.

By the central limit theorem,

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2).$$

By the weak law of large numbers,

$$\frac{1}{n}\sum_{k=1^n}(X_k^2+X_k^3)\xrightarrow[n\to\infty]{p}\mathbb{E}[X_k^2+X_k^3]=1/3.$$

Therefore, by Slutsky,

$$Z_n\xrightarrow[n\to\infty]{d}\mathcal{N}(0,9\sigma^2)=\mathcal{N}(0,3).$$

Theorem 7.20 (Borel-Cantelli)
Let $E_1,\ldots,$ be a sequence of events in some probability space. If

$$\sum_{i=1}^{\infty}\mathbb{P}[E_i]<\infty,$$

then, with probability 1, only a finite number of events will occur. This is the same as saying

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty}\bigcup_{j=1}^{\infty}E_j\right)=0.$$

7.5 Bounding Methods

Lemma 7.21 (Union Bound)
For events $A_1,\ldots,A_n$,

$$\mathbb{P}[A_1\cup\ldots\cup A_n]\le\sum_{i=1}^{n}\mathbb{P}[A_i].$$

Proof. We show that this is true by inducting on n. For $n=2$, the result follows

from PIE. Now, assume that it is true for $n = k - 1$.

$$\begin{aligned}\mathbb{P}\left[\bigcup_{i=1}^{k} A_i\right] &= \mathbb{P}\left[\left(\bigcup_{i=1}^{k-1} A_i\right) \cup A_k\right] \\ &\leq \mathbb{P}\left[\bigcup_{i=1}^{k-1} A_i\right] + \mathbb{P}[A_k] \\ &\leq \sum_{i=1}^{k} \mathbb{P}[A_i],\end{aligned}$$

by our inductive hypothesis. □

Lemma 7.22 (Markov's Inequality)
Consider a nonnegative random variable X with finite expectation $\mathbb{E}[X]$. Then,

$$\mathbb{P}[X \geq k \cdot \mathbb{E}[X]] \leq \frac{1}{k}.$$

This is the same as saying

$$\mathbb{P}[X \geq k] \leq \frac{\mathbb{E}[X]}{k}.$$

As stated, this bound is usually not very good. The proof follows directly from the law of total expectation.

Lemma 7.23 (Chebyshev's inequality)
Consider a random variable X with finite expectation $\mathbb{E}[X]$ and variance σ^2. Then for any $k > 0$,

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

Proof. This is a direct application of Markov's Inequality to the random variable $(X - \mathbb{E}[X])^2$. □

The next bounds are very popular in Computer Science.

Lemma 7.24 (Generic Chernoff Bounds)

Consider a random variable X and any $k \in \mathbb{R}$, $t \in \mathbb{R}_{\geq 0}$. Then,

$$\mathbb{P}[X \geq k] \leq \frac{\mathbb{E}[e^{tX}]}{e^{tk}}.$$

Proof. This is a direct application of Markov's Inequality to the random variable e^{tX}. □

Generally speaking, Chernoff bounds are stronger than bounds using Chebyshev, which are stronger than bounds using Markov.

Example 7.25

Let $X \sim \mathrm{B}(n, 1/2)$. Compute upper bounds on $\mathbb{P}[X > 3n/4]$.

Using Markov, $\mathbb{P}[X > 3n/4] = 2/3$, which is really weak. Using Chebyshev, $\mathbb{P}[X > 3n/4] = 4/n$, which is weak, but decays. Noting that $\mathbb{E}[e^{tX}] = \sum_{i=0}^{n} \binom{n}{i} \cdot 2^{-n} e^{ti} = (1 + e^t)^n \cdot 2^{-n}$, the generic Chernoff bound gives us:

$$\mathbb{P}[X > 3n/4] \leq \exp\left(n \log\left(\frac{1 + e^t}{2}\right) - \frac{3nt}{4}\right).$$

This holds for any $t \in \mathbb{R}_{\geq 0}$ (hence is the "generic" bound), so we may choose a specific t to make the bound as tight as possible. Differentiating, the expression on the RHS is minimized when $t = \log 3$, giving us

$$\mathbb{P}[X > 3n/4] \leq \left(\frac{16}{27}\right)^{n/4}.$$

This is the tightest upper bound by far, since it decays exponentially.