MATH 307, STATISTICAL INFERENCE, FALL 2022

STEVEN HEILMAN

Contents

Date: November 1, 2022

Exercise 0.1. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. Show that

$$\cup_{y\in\mathbb{R}}\{x \in \mathbb{R}\colon f(x) = y\} = \mathbb{R}.$$

Also, show that the union on the left is disjoint. That is, if $y_1 \neq y_2$ and $y_1, y_2 \in \mathbb{R}$, then $\{x \in \mathbb{R}\colon f(x) = y_1\} \cap \{x \in \mathbb{R}\colon f(x) = y_2\} = \emptyset$.

0.1. **Probabilistic Models.**

Definition 0.2. A **probabilistic model** consists of

- A universal set Ω, which represents all possible outcomes of some random process.
- A **probability law P**. Given a set $A \subseteq \Omega$, the probability law assigns a number $\mathbf{P}(A)$ to the set A. A set $A \subseteq \Omega$ is also called an **event**. The number $\mathbf{P}(A)$ denotes the probability that the event A will occur. The probability law satisfies the axioms below.

Axioms for a Probability Law:

(i) For any $A \subseteq \Omega$, we have $\mathbf{P}(A) \geq 0$. (**Nonnegativity**)

(ii) For any $A, B \subseteq \Omega$ such that $A \cap B = \emptyset$, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

If $A_1, A_2, \ldots \subseteq \Omega$ and $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k). \qquad (\textbf{Additivity})$$

(iii) We have $\mathbf{P}(\Omega) = 1$. (**Normalization**)

Remark 0.3. Since $\mathbf{P}(A) + \mathbf{P}(A^c) = 1$, choosing $A = \emptyset$ shows that $\mathbf{P}(\emptyset) + \mathbf{P}(\Omega) = 1$, so that $\mathbf{P}(\emptyset) = 0$ by Axiom (iii). Consequently, suppose n is a positive integer, and let $A_1, \ldots, A_n \subseteq \Omega$ with $A_i \cap A_j = \emptyset$ whenever $i, j \in \{1, \ldots, n\}$ and $i \neq j$. For any $i > n$, let $A_i = \emptyset$. Then Axiom (ii) implies that

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

This identity also follows from the first part of Axiom (ii) and by induction on n.

Theorem 0.4 (**Total Probability Theorem**). *Let $A_1, \ldots, A_n$ be disjoint events in a sample space Ω. That is, $A_i \cap A_j = \emptyset$ whenever $i, j \in \{1, \ldots, n\}$ satisfy $i \neq j$. Assume also that $\cup_{i=1}^{n} A_i = \Omega$. Let $\mathbf{P}$ be a probability law on Ω. Then, for any event $B \subseteq \Omega$, we have*

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B \cap A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i)\mathbf{P}(B|A_i).$$

Example 0.5 (**Bernoulli Trials**). Let n be a positive integer. Let $\Omega = \{H, T\}^n$. Then Ω is a sample space representing n separate coin flips (H stands for heads, and T stands for tails). Let $0 < p < 1$. Let $\mathbf{P}$ be the probability law such that each coin toss occurs independently, and such that each coin has probability p of heads (H), and probability $1 - p$ of tails (T). That is, we are independently flipping n biased coins.

Let $1 \leq k \leq n$. Suppose the first k coins have landed as heads, and the rest of the coins are tails. By the definition of $\mathbf{P}$, this event occurs with probability $p^k(1-p)^{n-k}$. We

now ask: What is the probability that k of the coins are heads, and the remaining $n-k$ coins are tails? In order to answer this question, we need to compute $C_{n,k}$, the number of unordered lists of k copies of H, and $n-k$ copies of T. Equivalently, $C_{n,k}$ is the number of ways to place n coins on a table all showing tails, and then turn over k distinct coins to reveal exactly k heads. To compute the latter number, note that we can first turn over one of the n coins, and then we can turn over any of the remaining $n-1$ coins showing tails, and then we can turn over any of the remaining $n-2$ coins showing tails, and so on. So, there are $n(n-1)(n-2)\cdots(n-k+1)$ sequences of coin turns which can be made (while keeping track of their ordering). To make the same count of coin flips without keeping track of the ordering, we just divide by the number of orderings of the k heads coins, which is $k(k-1)\cdots(2)(1)$. In conclusion,

$$C_{n,k} = \binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots(2)(1)} = \frac{n!}{(n-k)!k!}.$$

Back to our original question, the probability that we have k heads and $n-k$ tails among n coin flips is

$$C_{n,k}\cdot p^k(1-p)^{n-k} = \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}.$$

Theorem 0.6 (**Binomial Theorem**). *Let $0<p<1$. Then*

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

More generally, for any real numbers x, y, we have

$$\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k} = (x+y)^n$$

Proof. We use the notation of Example 0.5. Let $0<p<1$. For any $0\le k\le n$, let A_k be the event that there are exactly k heads that resulted from flipping n coins. Then $A_i\cap A_j=\emptyset$ for all $i\neq j$ where $i,j\in\{0,\ldots,n\}$. Also, $\cup_{k=0}^n A_k=\Omega$. From Example 0.5, $\mathbf{P}(A_k)=\binom{n}{k}p^k(1-p)^k$. So, using Remark 0.3,

$$1=\mathbf{P}(\Omega)=\mathbf{P}(\cup_{k=0}^n A_k)=\sum_{k=0}^{n}\mathbf{P}(A_k)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}. \qquad (*)$$

Now, the right side is a polynomial in p, which is equal to 1 for all $0<p<1$. Therefore, the equality $(*)$ holds for all real p. (A polynomial which is equal to 1 on $[0,1]$ is also equal to 1 on the whole real line.) Assume temporarily that $x+y\neq 0$. Define $p=x/(x+y)$. Then $x=p(x+y)$, $y=(1-p)(x+y)$ and $1-p=y/(x+y)$. Using $(*)$, we have

$$1=\sum_{k=0}^{n}\binom{n}{k}\left(\frac{x}{x+y}\right)^k\left(\frac{y}{x+y}\right)^{n-k}=(x+y)^{-n}\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}.$$

That is, our desired equality holds whenever $x+y\neq 0$. Finally, the case $x+y=0$ follows by letting $t>0$ be a real parameter, using $\sum_{k=0}^n\binom{n}{k}x^k(y+t)^{n-k}=(x+y+t)^n$, and letting $t\to 0$, noting that both sides of the equality are continuous in t. □

1. Discrete Random Variables

So far we have discussed random events. Often it is also natural to describe random numbers. For example, the sum of two six-sided die is a random number. Or your score obtained by throwing a single dart at a standard dartboard is a random number. In probability, we call random numbers **random variables**.

Definition 1.1 (**Random Variable**)**.** Let Ω be a sample space. Let $\mathbf{P}$ be a probability law on Ω. A **random variable** X is a function $X\colon \Omega \to \mathbb{R}$. A **discrete random variable** is a random variable whose range is either finite or countably infinite.

Proposition 1.2 (**Properties of Random Variables**)**.**

- *If X and Y are random variables, then $X+Y$ is a random variable.*
- *If X is a random variable and if $f\colon \mathbb{R} \to \mathbb{R}$ is a function, then $f(X) = f \circ X$ is a random variable.*

A random variable is "just" a function. So, in some sense, from your preparation in calculus, you are already quite familiar with random variables. However, the new terminology of "random variable" carries a new perspective on functions as well. For example, in probability theory, we concern ourselves with the probability that the random variable takes various values.

Example 1.3. Let $\Omega = \{1,2,3,4,5,6\}^2$. Let $\mathbf{P}$ denote the uniform probability law on Ω. As usual, Ω and $\mathbf{P}$ denote the rolling of two distinct fair six-sided dice. We define random variables X, Y as follows. For any $(i,j) \in \Omega$, define $X(i,j) = i$, and define $Y(i,j) = j$. Then X and Y are random variables. Moreover, X is the roll of the first die, and Y is the roll of the second die. So, $X+Y$ is the sum of the rolls of the dice, and $X+Y$ is a random variable.

Example 1.4. Consider the following simplified version of a dartboard. Let $\Omega = \mathbb{R}^2$. For any set $A \subseteq \Omega$, define

$$\mathbf{P}(A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dxdy.$$

Let $(x,y) \in \Omega$. Define a random variable $X\colon \Omega \to \mathbb{R}$ so that

$$X(x,y) = \begin{cases} 1 & \text{, if } x^2+y^2 \leq 1 \\ 0 & \text{, if } x^2+y^2 > 1 \end{cases}.$$

That is, if you hit the dartboard $\{(x,y) \in \Omega\colon x^2+y^2 \leq 1\}$, then $X = 1$. Otherwise, $X = 0$. So, X is a random variable which represents your score after throwing a random dart according to the probability law $\mathbf{P}$.

Example 1.5. Consider the following model of a more complicated dartboard. Let $\Omega = (0,1)^2 \subseteq \mathbb{R}^2$. For any set $A \subseteq \Omega$, let $\mathbf{P}(A)$ denote the area of A. Let $(x,y) \in \Omega$. Define a random variable $X\colon \Omega \to \mathbb{R}$ so that $X(x,y)$ is the smallest integer j such that $x > 2^{-j}$ and $y > 2^{-j}$. For example, if $(x,y) = (1/3, 1/3)$, then $2^{-1} > x > 2^{-2}$ and $2^{-1} > y > 2^{-2}$, so $X(x,y) = 2$. Or if $(x,y) = (1/5, 1/3)$, then $2^{-2} > x > 2^{-3}$ and $2^{-1} > y > 2^{-2} > 2^{-3}$, so $X(x,y) = 3$. In this example, X is a random variable which represents your score after throwing a random dart according to the probability law $\mathbf{P}$. By the definition of X, if we would like to get a large score, we see that it is more beneficial to aim for the bottom left corner of the square, i.e. we want to get close to $(0,0)$.

If we have a random variable X, one of the first tasks in probability is to compute various quantities for X to better understand X. For example, we could ask, "What value does X typically take?" (What is the mean value or average value of X?) "Typically, how far is X from its mean value?" (What is the variance of X?) We will start to answer these questions in Section 2. For now, we need to get through some preliminary concepts.

1.1. Probability Mass Function (PMF).

Definition 1.6 (**Probability Mass Function**). Let X be a random variable on a sample space Ω, so that $X\colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$. Consider the event $\{\omega \in \Omega\colon X(\omega) = x\}$. This event is often denoted as $\{X = x\}$. The **probability mass function** of X, denote $p_X\colon \mathbb{R} \to [0,1]$ is defined by

$$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{X = x\}) = \mathbf{P}(\{\omega \in \Omega\colon X(\omega) = x\}), \qquad x \in \mathbb{R}.$$

Let $A \subseteq \mathbb{R}$. We denote $\{\omega \in \Omega\colon X(\omega) \in A\} = \{X \in A\}$.

Example 1.7. Let $\Omega = \{H,T\}^2$ and let $\mathbf{P}$ be the uniform probability measure on Ω. Then Ω and $\mathbf{P}$ represent the outcome of flipping two distinct fair coins. Let X be the number of heads that are rolled. That is, $X(T,T) = 0$, $X(H,T) = 1$, $X(T,H) = 1$ and $X(H,H) = 2$. Therefore,

$$p_X(x) = \begin{cases} 1/4 & \text{, if } x = 0 \\ 1/2 & \text{, if } x = 1 \\ 1/4 & \text{, if } x = 2 \\ 0 & \text{, otherwise.} \end{cases}$$

Note that $\mathbf{P}(X > 0) = 1/2 + 1/4 = 3/4$. That is, with probability $3/4$, at least one head is rolled.

Proposition 1.8. *Let X be a discrete random variable on a sample space Ω. Then*

$$\sum_{x \in \mathbb{R}} p_X(x) = 1.$$

Proof. For each $x \in \mathbb{R}$, let B_x be the event that $X = x$. If $x \neq y$, then $B_x \cap B_y = \emptyset$. Also, $\cup_{x \in \mathbb{R}} B_x = \Omega$. So, using Axiom (ii) for probability laws in Definition 0.2,

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(\cup_{x \in \mathbb{R}} B_x) = \sum_{x \in \mathbb{R}} \mathbf{P}(B_x) = \sum_{x \in \mathbb{R}} p_X(x).$$

□

We now give descriptions of some commonly encountered random variables.

Definition 1.9 (**Bernoulli Random Variable**). Let $0 < p < 1$. A random variable X is called a **Bernoulli random variable with parameter** p if $X = 1$ with probability p, and $X = 0$ with probability $1 - p$. Put another way, $X = 1$ when a single flipped biased coin lands heads, and $X = 0$ when the coin lands tails. The PMF is given by

$$p_X(x) = \begin{cases} p & \text{, if } x = 1 \\ 1 - p & \text{, if } x = 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Remark 1.10. Note that we defined the random variable X without specifying any sample space Ω. This de-emphasis on the domain is one aspect of probability that we mentioned above. For example, we could choose $\Omega = \{0, 1\}$ and define $\mathbf{P}$ on Ω such that $\mathbf{P}(0) = 1 - p$ and $\mathbf{P}(1) = p$. Then define $X \colon \Omega \to \mathbb{R}$ so that $X(\omega) = \omega$ for all $\omega \in \Omega$. Then X is a Bernoulli random variable.

Alternatively, we could choose $\Omega = [0, 5]$, and define $\mathbf{P}$ on Ω such that $\mathbf{P}[a, b] = \frac{1}{5}(b - a)$ whenever $0 \leq a < b \leq 5$. Then, we could define $Y \colon \Omega \to \mathbb{R}$ by

$$Y(\omega) = \begin{cases} 1 & \text{, if } \omega < 5p \\ 0 & \text{, if } \omega \geq 5p. \end{cases}$$

Then Y is also a Bernoulli random variable. As we can see, the sample spaces of X and Y are very different.

Definition 1.11 (**Binomial Random Variable**)**.** Let $0 < p < 1$ and let n be a positive integer. A random variable X is called a **binomial random variable with parameters** n **and** p if X has the following PMF. If k is an integer with $0 \leq k \leq n$, then

$$p_X(k) = \mathbf{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

For any other x, we have $p_X(x) = 0$. In Example 0.5, we showed that this probability distribution arises from flipping n biased coins. In particular, X is the number of heads that arise when flipping n biased coins. In Theorem 0.6, we verified that

$$\sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = 1.$$

Definition 1.12 (**Geometric Random Variable**)**.** Let $0 < p < 1$. A random variable X is called a **geometric random variable with parameter** p if X has the following PMF. If k is a positive integer, then

$$p_X(k) = \mathbf{P}(X = k) = (1 - p)^{k-1} p.$$

For any other x, we have $p_X(x) = 0$. Note that X is the number of times that are needed to flip a biased coin in order to get a heads (if the coin has probability p of landing heads). Also, using the summation of geometric series, we verify

$$\begin{aligned} \sum_{k=1}^{\infty} p_X(k) &= \sum_{k=1}^{\infty} (1 - p)^{k-1} p = p \sum_{k=1}^{\infty} (1 - p)^{k-1} = p \lim_{n \to \infty} \sum_{k=1}^{n} (1 - p)^{k-1} \\ &= p \lim_{n \to \infty} \frac{1 - (1 - p)^{n+1}}{1 - (1 - p)} = \frac{p}{p} = 1. \end{aligned}$$

Definition 1.13 (**Poisson Random Variable**)**.** Let $\lambda > 0$. A random variable X is called a **Poisson random variable with parameter** λ if X has the following PMF. If k is a nonnegative integer, then

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

For any other x, we have $p_X(x) = 0$. Using the Taylor expansion for the exponential function, we verify

$$\sum_{k=0}^{\infty} p_X(k) = e^{-\lambda} \sum_{k=0} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

The Poisson random variable provides a nice approximation to the binomial distribution, as we now demonstrate.

Proposition 1.14 (**Poisson Approximation to the Binomial**). *Let $\lambda > 0$. For each positive integer n, let $0 < p_n < 1$, and let X_n be a binomial distributed random variable with parameters n and p_n. Assume that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$. Then, for any nonnegative integer k, we have*

$$\lim_{n\to\infty} \mathbf{P}(X_n = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Lemma 1.15. *Let $\lambda > 0$. For each positive integer n, let $\lambda_n > 0$. Assume that $\lim_{n\to\infty} \lambda_n = \lambda$. Then*

$$\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda}$$

Proof. Let log denote the natural logarithm. For any $x < 1$, define $f(x) = \log(1 - x)$. From L'Hôpital's Rule,

$$\lim_{x\to 0} \frac{f(x)}{x} = \lim_{x\to 0} f'(x) = \lim_{x\to 0} \frac{-1}{1-x} = -1. \qquad (*)$$

So, using $\lim_{n\to\infty} \lambda_n/n = 0$ we can apply $(*)$ and then $\lim_{n\to\infty} \lambda_n = \lambda$, so

$$\begin{aligned}\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^n &= \lim_{n\to\infty} \exp\left(\log\left(1 - \frac{\lambda_n}{n}\right)^n\right) \\ &= \exp\left(\lim_{n\to\infty} \frac{\log\left(1 - \frac{\lambda_n}{n}\right)}{\lambda_n/n} \lambda_n\right) = \exp((-1)(\lambda)) = e^{-\lambda}.\end{aligned}$$

□

Proof of Proposition 1.14. For any positive integer n, let $\lambda_n = np_n$. Then $\lim_{n\to\infty} \lambda_n = \lambda$ and $\lim_{n\to\infty} \lambda_n/n = 0$. And if k is a nonnegative integer,

$$\begin{aligned}\mathbf{P}(X_n = k) &= \binom{n}{k} \left(\frac{\lambda_n}{n}\right)^k \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \\ &= \left(\prod_{i=1}^{k} \frac{n-i+1}{n}\right) \frac{\lambda_n^k}{k!} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-k}\end{aligned}$$

So, using Lemma 1.15, $\lim_{n\to\infty} \mathbf{P}(X_n = k) = 1 \cdot \frac{\lambda^k}{k!} e^{-\lambda} \cdot 1$. □

Remark 1.16. A Poisson random variable is often used as an approximation for counting the number of some random occurrences. For example, the Poisson distribution can model the number of typos per page in a book, the number of magnetic defects in a hard drive, the number of traffic accidents in a day, etc.

Exercise 1.17. The Wheel of Fortune involves the repeated spinning of a wheel with 72 possible stopping points. We assume that each time the wheel is spun, any stopping point is equally likely. Exactly one stopping point on the wheel rewards a contestant with $\$1,000,000$. Suppose the wheel is spun 24 times. Let X be the number of times that someone wins $\$1,000,000$. Using the Poisson Approximation the Binomial, estimate the following probabilities: $\mathbf{P}(X = 0)$, $\mathbf{P}(X = 1)$, $\mathbf{P}(X = 2)$. (Hint: consider the binomial distribution with $p = 1/72$.)

Remark 1.18. The Bernoulli, binomial, geometric and Poisson random variables are all examples of the following general construction of a random variable. Let $a_0, a_1, a_2, \ldots \geq 0$ such that $\sum_{i=0}^{\infty} a_i = 1$. Then define a random variable X such that $\mathbf{P}(X = i) = a_i$ for all nonnegative integers i.

There are many other random variables we will encounter in this class as well, but these will be enough for now.

1.2. **Functions of Random Variables.**

Proposition 1.19. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω, and let $f\colon \mathbb{R} \to \mathbb{R}$. Then $f(X)$ has PMF*

$$p_{f(X)}(y) = \sum_{x \in \mathbb{R}\colon\, f(x)=y} p_X(x), \qquad \forall\, y \in \mathbb{R}.$$

Proof. Let $x, y, z \in \mathbb{R}$. Let A_x be the event that $X = x$. If $z \neq x$, then $A_x \cap A_z = \emptyset$. Also, $\cup_{x \in \mathbb{R}} A_x = \Omega$. So, using Axiom (ii) of Definition 0.2,

$$p_{f(X)}(y) = \mathbf{P}(f(X) = y) = \mathbf{P}(\cup_{x \in \mathbb{R}} \{f(X) = y\} \cap A_x) = \sum_{x \in \mathbb{R}} \mathbf{P}(\{f(X) = y\} \cap A_x)$$
$$= \sum_{x \in \mathbb{R}\colon\, f(x)=y} \mathbf{P}(X = x) \;=\; \sum_{x \in \mathbb{R}\colon\, f(x)=y} p_X(x).$$

□

Exercise 1.20. Let $\Omega = \{-3, -2, -1, 0, 1, 2, 3\}$. Suppose $X(\omega) = \omega$ for all $\omega \in \Omega$. Let $f\colon \mathbb{R} \to \mathbb{R}$ so that $f(x) = x^2$ for any $x \in \mathbb{R}$. Compute the PMF of $f(X)$.

2. Expectation

Now that we understand random variables a bit more, we can finally start to answer some of the fundamental questions of probability, such as:

What is the average value of a random variable?

Put another way, what is the mean value of a random variable? Or, what value should we expect a particular random variable to have? Answering this question is of interest in many applications of probability. For example, if I can figure out a way to gain $\$1$ from a stock transaction with probability .51, while losing $\$1$ from a stock transaction with probability .49, and if I keep performing this transaction many times, I should probably expect to gain money over time.

2.1. **Expectation, Variance.**

Definition 2.1 (**Expected Value**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Assume that $X\colon \Omega \to [0,\infty)$. We define the **expected value** of X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \sum_{x\in\mathbb{R}} x p_X(x).$$

For a discrete random variable with $X\colon \Omega \to \mathbb{R}$, if $\mathbf{E}\,|X| < \infty$, we then define $\mathbf{E}(X) = \sum_{x\in\mathbb{R}} x p_X(x)$ as above. The expected value of X is also referred to as the **mean** of X, or the **first moment** of X. More generally, if n is a positive integer, we define the n^{th} **moment** of X to be $\mathbf{E}(X^n)$.

Example 2.2. If X takes the values $\{1,2,3,4,5,6\}$ each with probability $1/6$, then we have

$$\mathbf{E}(X) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = \frac{21}{6} = \frac{7}{2}.$$

That is, on average, the result of the roll of one fair six-sided die will be around $7/2$. We can also compute

$$\mathbf{E}(X^2) = \frac{1}{6}(1^2) + \frac{1}{6}(2^2) + \frac{1}{6}(3^2) + \frac{1}{6}(4^2) + \frac{1}{6}(5^2) + \frac{1}{6}(6^2) = \frac{91}{6}.$$

Remark 2.3. Suppose X takes the value $(-2)^k$ with probability 2^{-k} for every positive integer k. Then $|X|$ takes the value 2^k with probability 2^{-k} for every positive integer k. So, $\mathbf{E}\,|X| = \sum_{k\geq 1} 1 = \infty$. So, $\mathbf{E}(X)$ is undefined.

Example 2.4. In a recent Powerball lottery, one ticket costs \$2, and the jackpot was around $\$(1/2)10^9$ (after deducting taxes). The number of people winning the jackpot shares the jackpot. Let X be your profit from buying one lottery ticket. Consider the following simplified version of the lottery. Suppose you either are the only winner of the jackpot, or you lose. There were around $(1/3)10^9$ tickets sold, and around $(1/3)10^9$ distinct possible ticket numbers. Assume that every ticket is chosen uniformly at random among all possible ticket numbers, and whether or not someone wins or loses is independent of everyone else. Let $p = 3\cdot 10^{-9}$. Then the probability that you win and everyone else loses is $p(1-p)^{1/p} \approx p/e \approx p/3$. That is, $\mathbf{P}(X = -2) \approx 1 - p/3$ and $\mathbf{P}(X = (1/2)10^9 - 2) \approx p/3$. So,

$$\mathbf{E}X = -2(1-p) + (1/2)10^9 p \approx -2 + 3/2 = -.5.$$

Since the expected value is negative, it was not sensible to buy a lottery ticket. Also, let N be the number of people who get the winning number. Using the Poisson Approximation to the Binomial with $\lambda = 1$, we have $\mathbf{P}(N = k) \approx \frac{1}{ek!}$ for any positive integer k. So, $\mathbf{P}(N = 0) \approx 1/e$, $\mathbf{P}(N = 1) \approx 1/e$, $\mathbf{P}(N = 2) \approx 1/(2e) \approx 1/6$, $\mathbf{P}(N = 3) \approx 1/(6e) \approx 1/18$, and so on. So, having two or three winners is not so unexpected.

Proposition 2.5 (**Expected Value Rule**)**.** *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Let $f\colon \mathbb{R}\to\mathbb{R}$ be a function. Then*

$$\mathbf{E}(f(X)) = \sum_{x\in\mathbb{R}} f(x) p_X(x).$$

In particular, if n is a positive integer, we can choose $f(x) = x^n$ to get

$$\mathbf{E}(X^n) = \sum_{x \in \mathbb{R}} x^n p_X(x).$$

Also, if a, b are constants, we can choose $f(x) = ax + b$ to get

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$$

Proof. From Proposition 1.19, $p_{f(X)}(y) = \sum_{x \in \mathbb{R}:\ f(x)=y} p_X(x)$. So,

$$\begin{aligned}\mathbf{E}(f(X)) &= \sum_{y \in \mathbb{R}} y p_{f(X)}(y) = \sum_{y \in \mathbb{R}} \sum_{x \in \mathbb{R}:\ f(x)=y} y p_X(x) \\ &= \sum_{y \in \mathbb{R}} \sum_{x \in \mathbb{R}:\ f(x)=y} f(x) p_X(x) = \sum_{x \in \mathbb{R}} f(x) p_X(x).\end{aligned}$$

In the last equality, we used Exercise 0.1.

Now, let a, b be constants. Using Proposition 2.5 and then Proposition 1.8,

$$\mathbf{E}(aX + b) = \sum_{x \in \mathbb{R}} (ax + b) p_X(x) = a \sum_{x \in \mathbb{R}} x p_X(x) + b \sum_{x \in \mathbb{R}} p_X(x) = a\mathbf{E}(X) + b.$$

□

Definition 2.6 (Variance). Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. We define the **variance** of X, denoted $\mathrm{var}(X)$, by

$$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

We define the **standard deviation** of X, denoted σ_X, by

$$\sigma_X = \sqrt{\mathrm{var}(X)}.$$

The notation $\mathbf{E}(X - \mathbf{E}(X))^2$ is a shorthand for $\mathbf{E}[(X - \mathbf{E}(X))^2]$.

Proposition 2.7 (Properties of Variance). *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a discrete random variable on Ω. Let a, b be constants. Then*

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Moreover,

$$\mathrm{var}(aX + b) = a^2 \mathrm{var}(X).$$

Proof. Using Proposition 2.5 and then Propositions 2.5 and 1.8,

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}(X - (\mathbf{E}(X))^2) = \sum_{x \in \mathbb{R}} (x - \mathbf{E}(X))^2 p_X(x) \\ &= \sum_{x \in \mathbb{R}} x^2 p_X(x) - 2\mathbf{E}(X) \sum_{x \in \mathbb{R}} x p_X(x) + (\mathbf{E}(X))^2 \sum_{x \in \mathbb{R}} p_X(x) \\ &= \mathbf{E}(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + (\mathbf{E}(X))^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.\end{aligned}$$

From Proposition 2.5, $\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$. So, using Proposition 2.5,

$$\begin{aligned}\mathrm{var}(aX + b) &= \mathbf{E}(aX + b - (a\mathbf{E}(X) + b))^2 = \mathbf{E}(aX - a\mathbf{E}(X))^2 = \mathbf{E}(a^2(X - \mathbf{E}(X))^2) \\ &= a^2 \mathbf{E}(X - \mathbf{E}(X))^2 = a^2 \mathrm{var}(X).\end{aligned}$$

□

Exercise 2.8. Let $a, b \in \mathbb{R}$ and let $X\colon \Omega \to [-\infty, \infty]$ be a random variable with $\mathbf{E}X^2 < \infty$. Show that

$$\operatorname{var}(aX + b) = a^2 \operatorname{var}(X).$$

Then, let X be a standard Gaussian. Show that $\mathbf{E}X = 0$ and $\operatorname{var}(X) = 1$.

Finally, show that the quantity $\mathbf{E}(X - t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbf{E}X$.

Example 2.9. Returning again to Example 2.2, suppose X takes the values $\{1, 2, 3, 4, 5, 6\}$ each with probability $1/6$. We computed $\mathbf{E}(X) = 7/2$, so

$$\begin{aligned}\operatorname{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2 \\ &= \frac{1}{6}(1 - \frac{7}{2})^2 + \frac{1}{6}(2 - \frac{7}{2})^2 + \frac{1}{6}(3 - \frac{7}{2})^2 + \frac{1}{6}(4 - \frac{7}{2})^2 + \frac{1}{6}(5 - \frac{7}{2})^2 + \frac{1}{6}(6 - \frac{7}{2})^2 = \frac{35}{12}.\end{aligned}$$

Alternatively, we computed in Example 2.2 that $\mathbf{E}(X^2) = 91/6$. So, by Proposition 2.7, $\operatorname{var}(X) = 91/6 - (7/2)^2 = 182/12 - 147/12 = 35/12$. Lastly, the standard deviation of X is $\sigma_X = \sqrt{35/12} \approx 1.7078$. So, the value of X is typically in the interval $(\mathbf{E}(X) - \sigma_X, \mathbf{E}(X) + \sigma_X) = (3.5 - 1.7078, 3.5 + 1.7078)$.

Example 2.10. Let X be a Poisson random variable with parameter $\lambda > 0$. Then $p_X(k) = e^{-\lambda}\lambda^k/k!$ when k is a nonnegative integer. We then compute

$$\begin{aligned}\mathbf{E}(X) &= \sum_{k=0}^{\infty} k p_X(k) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.\end{aligned}$$

Exercise 2.11. Let X be a discrete random variable taking a finite number of values. Let $t \in \mathbb{R}$. Consider the function $f\colon \mathbb{R} \to \mathbb{R}$ defined by $f(t) = \mathbf{E}(X - t)^2$. Show that the function f takes its minimum value when $t = \mathbf{E}X$. Moreover, if X takes at least two different values, each with some positive probability, then f is uniquely minimized when $t = \mathbf{E}X$.

Exercise 2.12. Let $0 < p < 1$ and let n be a positive integer. Compute the mean of a binomial random variable with parameter p. Then, compute the mean of a Poisson random variable with parameter $\lambda > 0$.

Exercise 2.13. Let X be a nonnegative random variable on a sample space Ω. Assume that X only takes integer values. Prove that

$$\mathbf{E}(X) = \sum_{n=1}^{\infty} \mathbf{P}(X \geq n).$$

Exercise 2.14. As we will see later in the course, the expectation is very closely related to integrals. This exercise gives a hint toward this relation. Let $\Omega = [0, 1]$. Let $\mathbf{P}$ be the probability law on Ω such that $\mathbf{P}([a, b]) = \int_a^b dt = b - a$ whenever $0 \leq a < b \leq 1$. Let n be a positive integer. Let $X\colon \Omega \to \mathbb{R}$ be such that X is constant on any interval of the form $[i/n, (i + 1)/n)$, whenever $0 \leq i \leq n - 1$. Show that

$$\mathbf{E}(X) = \int_0^1 X(t) dt$$

Now, consider a different probability law, where $\mathbf{P}([a,b]) = \int_a^b \frac{1}{2\sqrt{t}} dt$ whenever $0 \leq a < b \leq 1$. Show that

$$\mathbf{E}(X) = \int_0^1 X(t) \frac{1}{2\sqrt{t}} dt.$$

Exercise 2.15. Let $a_1, \ldots, a_n$ be distinct numbers, representing the quality of n people. Suppose n people arrive to interview for a job, one at a time, in a random order. That is, every possible arrival order of these people is equally likely. For each $i \in \{1, \ldots, n\}$, upon interviewing the i^{th} person, if $a_i > a_j$ for all $1 \leq j < i$, then the i^{th} person is hired. That is, if the person currently being interviewed is better than the previous candidates, she will be hired. What is the expected number of hirings that will be made? (Hint: let $X_i = 1$ if the i^{th} person to arrive is hired, and let $X_i = 0$ otherwise. Consider $\sum_{i=1}^n X_i$.)

2.2. **Joint Mass Function, Covariance.**

Definition 2.16 (**Joint PMF**). Let X, Y be two discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Let $x, y \in \mathbb{R}$. Define the **joint probability mass function** of X and Y by

$$p_{X,Y}(x,y) = \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(X = x \text{ and } Y = y) = \mathbf{P}(X = x, Y = y).$$

Let A be a subset of $\mathbb{R}^2$. We define

$$\mathbf{P}((X,Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x,y).$$

Proposition 2.17. *Let X, Y be two discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Then for any $x, y \in \mathbb{R}$,*

$$p_X(x) = \sum_{t \in \mathbb{R}} p_{X,Y}(x,t), \qquad p_Y(y) = \sum_{t \in \mathbb{R}} p_{X,Y}(t,y).$$

Proof. We prove the first equality, since the second one is proven similarly. For any $t \in \mathbb{R}$, let A_t be the event that $Y = t$. If $t_1 \neq t_2$, then $A_{t_1} \cap A_{t_2} = \emptyset$. And $\cup_{t \in \mathbb{R}} A_t = \Omega$. So, from Axiom (ii) in Definition 0.2,

$$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\cup_{t \in \mathbb{R}} \{X = x\} \cap \{Y = t\}) = \sum_{t \in \mathbb{R}} \mathbf{P}(X = x, Y = t) = \sum_{t \in \mathbb{R}} p_{X,Y}(x,t).$$

□

Remark 2.18. We refer to p_X as the **marginal** PMF of X, and we refer to p_Y as the marginal PMF of Y, to distinguish these PMFs from the joint PMF $p_{X,Y}$.

Proposition 2.19. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X and Y be discrete random variables on Ω taking a finite number of values. Let c be a constant. Let $f : \mathbb{R}^2 \to \mathbb{R}$. Then*

$$\mathbf{E} f(X,Y) = \sum_{(x,y) \in \mathbb{R}^2} f(x,y) p_{X,Y}(x,y).$$

Consequently, choosing $f(x,y) = x + y$, or $f(x,y) = cx$ where c is a constant,

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y), \qquad \mathbf{E}(cX) = c\mathbf{E}(X).$$

So, in linear algebraic terms, $\mathbf{E}$ is a linear transformation.

Proof. Let $z \in \mathbb{R}$. Then $p_{f(X,Y)}(z) = \mathbf{P}(f(X,Y) = z)$. Let $x, y \in \mathbb{R}$. Let $A_{x,y}$ be the event $\{X = x\} \cap \{Y = Y\}$. If $(x_1, y_1) \neq (x_2, y_2)$, then $A_{x_1,y_1} \cap A_{x_2,y_2} = \emptyset$. And $\cup_{(x,y)\in\mathbb{R}^2} A_{x,y} = \Omega$. So, from Axiom (ii) of Definition 0.2,

$$\mathbf{P}(f(X,Y) = z) = \mathbf{P}(\cup_{(x,y)\in\mathbb{R}^2}\{f(X,Y) = z\} \cap A_{x,y})$$
$$= \sum_{(x,y)\in\mathbb{R}^2} \mathbf{P}(\{f(X,Y) = z\} \cap \{X = x\} \cap \{Y = y\}) = \sum_{(x,y)\in\mathbb{R}^2 :\, f(x,y)=z} \mathbf{P}(X = x, Y = y).$$

Note that $\mathbb{R}^2 = \cup_{z\in\mathbb{R}}\{(x,y) \in \mathbb{R}^2 : f(x,y) = z\}$, where the union is disjoint. So,

$$\mathbf{E}(f(X,Y)) = \sum_{z\in\mathbb{R}} z p_{f(X,Y)}(z) = \sum_{z\in\mathbb{R}} z \sum_{(x,y)\in\mathbb{R}^2 :\, f(x,y)=z} \mathbf{P}(X = x, Y = y)$$
$$= \sum_{(x,y)\in\mathbb{R}^2} f(x,y)\mathbf{P}(X = x, Y = y)$$

The first equality is proven. We now consider $f(x,y) = x + y$. We have

$$\mathbf{E}(X+Y) = \sum_{x\in\mathbb{R}} x \sum_{y\in\mathbb{R}} \mathbf{P}(X = x, Y = y) + \sum_{y\in\mathbb{R}} y \sum_{x\in\mathbb{R}} \mathbf{P}(X = x, Y = y)$$
$$= \sum_{x\in\mathbb{R}} x\mathbf{P}(X = x) + \sum_{y\in\mathbb{R}} y\mathbf{P}(Y = y) = \mathbf{E}(X) + \mathbf{E}(Y).$$

In the last line, we used Proposition 2.17 to get $\sum_{y\in\mathbb{R}} \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(X = x)$, and $\sum_{x\in\mathbb{R}} \mathbf{P}(\{X = x\} \cap \{Y = y\}) = \mathbf{P}(Y = y)$. Finally, the equality $\mathbf{E}(cX) = c\mathbf{E}(X)$ was proven in Proposition 2.5. □

Exercise 2.20. Suppose there are ten separate bins. You first randomly place a sphere randomly in one of the bins, where each bin has an equal probability of getting the sphere. Once again, you randomly place another sphere uniformly at random in one of the bins. This process occurs twenty times, so that twenty spheres have been placed in bins. What is the expected number of empty bins at the end?

Exercise 2.21. You want to complete a set of 100 baseball cards. Cards are sold in packs of ten. Assume that each card is equally likely to be contained in any pack of cards. How many packs of cards should you buy in order to get a complete set of cards?

Exercise 2.22. Suppose we are drawing cards out of a standard 52 card deck without replacing them. How many cards should we expect to draw out of the deck before we find (a) a King? (b) a Heart?

Exercise 2.23. Let $f : \mathbb{R} \to \mathbb{R}$ be twice differentiable function. Assume that f is convex. That is, $f''(x) \geq 0$, or equivalently, the graph of f lies above any of its tangent lines. That is, for any $x, y \in \mathbb{R}$,

$$f(x) \geq f(y) + f'(y)(x - y).$$

(In Calculus class, you may have referred to these functions as "concave up.") Let X be a discrete random variable. By setting $y = \mathbf{E}(X)$, prove **Jensen's inequality**:

$$\mathbf{E}f(X) \geq f(\mathbf{E}(X)).$$

In particular, choosing $f(x) = x^2$, we have $\mathbf{E}(X^2) \geq (\mathbf{E}(X))^2$.

Definition 2.24 (**Covariance**). Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X and Y be discrete random variables on Ω taking a finite number of values. We define the **covariance** of X and Y, denoted $\operatorname{cov}(X, Y)$, by

$$\operatorname{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

Remark 2.25.

$$\operatorname{cov}(X, X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \operatorname{var}(X).$$

Remark 2.26. By the Cauchy-Schwarz inequality (see Theorem 2.27), we have

$$|\operatorname{cov}(X, Y)| \leq (\mathbf{E}(X - \mathbf{E}X)^2)^{1/2}(\mathbf{E}(Y - \mathbf{E}Y)^2)^{1/2}.$$

So, the covariance is well defined if X, Y both have finite variance. Note that

$$\operatorname{cov}(X, X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \operatorname{var}(X).$$

Theorem 2.27 (**Cauchy-Schwarz Inequality**). *Let $X, Y \colon \Omega \to \mathbb{R}$ be random variables. Then*

$$\mathbf{E}\,|XY| \leq (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Proof. By scaling, we may assume $\mathbf{E}X^2 = \mathbf{E}Y^2 = 1$ (zeros and infinities being trivial). From concavity of the log function, we have the pointwise inequality

$$|X(\omega)Y(\omega)| = (|X(\omega)|^2)^{1/2}(|Y(\omega)|^2)^{1/2} \leq \frac{1}{2}\,|X(\omega)|^2 + \frac{1}{2}\,|Y(\omega)|^2\,, \qquad \forall\, \omega \in \Omega$$

which upon integration gives the result. □

The covariance of X and Y is meant to measure whether or not X and Y are related somehow. We will discuss the meaning of covariance a bit more further below. For now, we make the following observation.

Lemma 2.28. *Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let $X_1, \ldots, X_n$ be discrete random variables on Ω taking a finite number of values. Then*

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j).$$

Proof. From Proposition 2.19, $\mathbf{E}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{E}(X_i)$. So,

$$\begin{aligned}
\operatorname{var}(\sum_{i=1}^{n} X_i) &= \mathbf{E}(\sum_{i=1}^{n} X_i - \mathbf{E}(\sum_{i=1}^{n} X_i))^2 = \mathbf{E}(\sum_{i=1}^{n} (X_i - \mathbf{E}(X_i)))^2 \\
&= \mathbf{E}\left(\sum_{i=1}^{n} (X_i - \mathbf{E}(X_i))^2\right) + 2\mathbf{E}\left(\sum_{1 \leq i < j \leq n} (X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\right) \\
&= \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \operatorname{cov}(X_i, X_j).
\end{aligned}$$

□

Exercise 2.29. Let n be a positive integer, and let $0 < p < 1$. Let $\Omega = \{0,1\}^n$. Any $\omega \in \Omega$ can then be written as $\omega = (\omega_1, \ldots, \omega_n)$ with $\omega_i \in \{0,1\}$ for each $i \in \{1, \ldots, n\}$. Let $\mathbf{P}$ be the probability law described in Example 0.5. That is, for any $\omega \in \Omega$, we have

$$\mathbf{P}(\omega) = \prod_{i=1}^{n} p^{\omega_i}(1-p)^{1-\omega_i} = p^{\sum_{i=1}^n \omega_i}(1-p)^{n-\sum_{i=1}^n \omega_i}.$$

For each $i \in \{1, \ldots, n\}$, define $X_i \colon \Omega \to \mathbb{R}$ so that $X_i(\omega) = \omega_i$ for any $\omega \in \Omega$. That is, if Ω and $\mathbf{P}$ model the flipping of n distinct biased coins, then $X_i = 1$ when the i^{th} coin is heads, and $X_i = 0$ when the i^{th} coin is tails.

First, show that $\mathbf{P}(\Omega) = 1$. Then, compute the expected value of X_i for each $i \in \{1, \ldots, n\}$. Next, compute the expected value of $Y = \sum_{i=1}^n X_i$. Finally, prove that Y is a binomial random variable with parameters n and p.

Exercise 2.30 (**Inclusion-Exclusion Formula**)**.** This Exercise gives an alternate proof of the following identity, which is known as the Inclusion-Exclusion Formula: Let $A_1, \ldots, A_n \subseteq \Omega$. Then:

$$\mathbf{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{1 \le i < j \le n} \mathbf{P}(A_i \cap A_j) + \sum_{1 \le i < j < k \le n} \mathbf{P}(A_i \cap A_j \cap A_k) \\ \cdots + (-1)^{n+1} \mathbf{P}(A_1 \cap \cdots \cap A_n).$$

Let Y be a random variable such that $Y = 1$ on $\cup_{i=1}^n A_i$, and such that $Y = 0$ otherwise.

- Show that $Y = 1 - \prod_{i=1}^n (1 - X_i)$.
- Expand out the product in the previous item, and take the expected value of both sides of the result. Deduce the Inclusion-Exclusion formula.

2.2.1. *More than Two Random Variables.* Our results on the joint PMF can be easily extended to any number of random variables. For example, if $X_1, \ldots, X_n$ are discrete random variables, and if $x_1, \ldots, x_n \in \mathbb{R}$, the joint PMF of $X_1, \ldots, X_n$ is defined as

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n).$$

Then

$$p_{X_1}(x_1) = \sum_{x_2,\ldots,x_n \in \mathbb{R}} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n),$$

$$p_{X_1,X_2}(x_1, x_2) = \sum_{x_3,\ldots,x_n \in \mathbb{R}} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n), \qquad \text{etc.}$$

Also, if $f \colon \mathbb{R}^n \to \mathbb{R}$ is a function, we have

$$\mathbf{E} f(X_1, \ldots, X_n) = \sum_{x_1,\ldots,x_n \in \mathbb{R}} f(x_1, \ldots, x_n) p_{X_1,\ldots,X_n}(x_1, \ldots, x_n).$$

2.3. **Independence of Random Variables.** Recall that sets A, B are independent when $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. The independence of random variables is a bit more involved than the independence of sets, since we will require many equalities to hold.

Definition 2.31 (**Independence of a Random Variable and a Set**)**.** Let X be a discrete random variable on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$. We say that X **is independent of** A if

$$\mathbf{P}(\{X = x\} \cap A) = \mathbf{P}(X = x)\mathbf{P}(A), \qquad \forall\, x \in \mathbb{R}.$$

That is, $\{X = x\}$ is independent of A, for all $x \in \mathbb{R}$. That is, knowing that A has occurred does not changed our knowledge of any value of X.

Example 2.32. Let $\Omega = \{0, 1\}^2$ and let $\mathbf{P}$ be the uniform probability measure on Ω. Then $\mathbf{P}$ models the toss of two distinct fair coins. For any $\omega = (\omega_1, \omega_2) \in \{0, 1\}^2$, define $X(\omega) = \omega_1$. That is, $X = 1$ when the first coin toss is heads (1), and $X = 0$ when the first coin toss is tails (0). Let A be the event that the second coin toss is heads. That is, $A = \{(0, 1), (1, 1)\}$. We will show that X and A are independent.

$$\mathbf{P}(\{X = 1\} \cap A) = \mathbf{P}(\{(1, 0), (1, 1)\} \cap A) = \mathbf{P}(1, 1) = 1/4 = (1/2)(1/2) = \mathbf{P}(X = 1)\mathbf{P}(A).$$
$$\mathbf{P}(\{X = 0\} \cap A) = \mathbf{P}(\{(0, 0), (0, 1)\} \cap A) = \mathbf{P}(0, 1) = 1/4 = (1/2)(1/2) = \mathbf{P}(X = 0)\mathbf{P}(A).$$

Therefore, X and A are independent.

Definition 2.33 (**Independence of a Random Variable from another**)**.** Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that X **is independent of** Y if

$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x)\mathbf{P}(Y = y), \qquad \forall\, x, y \in \mathbb{R}.$$

That is, $\{X = x\}$ is independent of $\{Y = y\}$, for all $x, y \in \mathbb{R}$. That is, knowing the values of Y does not changed our knowledge of any value of X. Written another way,

$$p_{X,Y}(x, y) = p_X(x)p_Y(y), \qquad \forall\, x, y \in \mathbb{R}.$$

When two random variables are independent, they satisfy many nice properties. For example,

Theorem 2.34. *Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that X and Y are independent. Assume that X and Y take a finite number of values. Then*

$$\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$$

Proof. Using Proposition 2.19 and the equality $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$,

$$\mathbf{E}(XY) = \sum_{x,y \in \mathbb{R}} xyp_{X,Y}(x, y) = \sum_{x \in \mathbb{R}} xp_X(x) \sum_{y \in \mathbb{R}} yp_Y(y) = \mathbf{E}(X)\mathbf{E}(Y).$$

□

Corollary 2.35. *Let $X_1, \ldots, X_n$ be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that $X_1, \ldots, X_n$ are pairwise independent. That is, X_i and X_j are independent whenever $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Assume that $X_1, \ldots, X_n$ take a finite number of values. Then*

$$\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \operatorname{var}(X_i).$$

Proof. Let $i, j \in \{1, \dots, n\}$ with $i \neq j$. Then by Theorem 2.34,

$$\operatorname{cov}(X_i, X_j) = \mathbf{E}((X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))) = \mathbf{E}(X_i X_j) - 2\mathbf{E}(X_i)\mathbf{E}(X_j) + \mathbf{E}(X_i)\mathbf{E}(X_j) = 0.$$

So, Lemma 2.28 concludes the proof. □

Exercise 2.36. Let X, Y, Z be discrete random variables. Let $f(y) = \mathbf{E}(X|Y = y)$ for any $y \in \mathbb{R}$. Then $f\colon \mathbb{R} \to \mathbb{R}$ is a function. In more advanced probability classes, we consider the random variable $f(Y)$, which is denoted by $\mathbf{E}(X|Y)$. Show that $\mathbf{E}(X + Z|Y) = \mathbf{E}(X|Y) + \mathbf{E}(Z|Y)$. Then, show that $\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}(X)$. That is, understanding $\mathbf{E}(X|Y)$ can help us to compute $\mathbf{E}(X)$.

Exercise 2.37. Give an example of two random variables X, Y that are independent. Prove that these random variables are independent.

Give an example of two random variables X, Y that are not independent. Prove that these random variables are not independent.

Finally, find two random variables X, Y such that $\mathbf{E}(XY) \neq \mathbf{E}(X)\mathbf{E}(Y)$.

Exercise 2.38. Is it possible to have a random variable X such that X is independent of X? Either find such a random variable X, or prove that it is impossible to find such a random variable X.

Exercise 2.39. Let $0 < p < 1$. Let n be a positive integer. Let $X_1, \dots, X_n$ be pairwise independent Bernoulli random variables. Compute the expected value of

$$S_n = \frac{X_1 + \cdots + X_n}{n}.$$

Then, compute the variance of $S_n - \mathbf{E}(S_n)$. Describe in words what this variance computation tells you as $n \to \infty$. Particularly, what does S_n "look like" as $n \to \infty$? (Consider the following statistical interpretation. Suppose each X_i is the result of some poll of person i, where $i \in \{1, \dots, n\}$. Suppose that each person's response is a Bernoulli random variable with parameter p, and each person's response is independent of each other person's response. Then S_n is the average of the results of the poll. If $S_n - \mathbf{E}(S_n)$ has small variance, then our poll is very accurate. So, how accurate is the poll as $n \to \infty$? Note that the accuracy of the poll does *not* depend on the size of the population you are sampling from!)

Exercise 2.40. Let X and Y be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. Assume that X and Y are independent. Assume that X and Y take a finite number of values. Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be functions. Then

$$\mathbf{E}(f(X)g(Y)) = \mathbf{E}(f(X))\mathbf{E}(g(Y)).$$

2.3.1. *Independence of Multiple Random Variables.*

Definition 2.41 (**Independence of Random Variables**). Let $X_1, \dots, X_n$ be discrete random variables on a sample space Ω, and let $\mathbf{P}$ be a probability law on Ω. We say that $X_1, \dots, X_n$ **are independent** if

$$\mathbf{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i), \qquad \forall\, x_1, \dots, x_n \in \mathbb{R}.$$

Remark 2.42. Suppose $X_1, \ldots, X_n$ are discrete, independent random variables taking a finite number of values. Let $f_1, \ldots, f_n$ be functions from $\mathbb{R}$ to $\mathbb{R}$. Similar to Exercise 2.40 we have

$$\mathbf{E}(\prod_{i=1}^{n} f_i(X_i)) = \prod_{i=1}^{n} \mathbf{E}(f_i(X_i)).$$

In particular,

$$\mathbf{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbf{E}(X_i).$$

Proposition 2.43. *Let $X_1, \ldots, X_n$ be discrete random variables on a sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $X_1, \ldots, X_n$ are independent. Then, for any subset S of $\{1, \ldots, n\}$, the random variables $\{X_i\}_{i \in S}$ are independent. In particular, $X_1, \ldots, X_n$ are pairwise independent.*

Proof. By reordering indices and iterating, it suffices to show that $X_1, \ldots, X_{n-1}$ are independent. That is, it suffices to show that

$$\mathbf{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i), \qquad \forall\, x_1, \ldots, x_{n-1} \in \mathbb{R}.$$

For any $x_n \in \mathbb{R}$, let $B_{x_n} = \{X_n = x_n\}$. Then $B_{x_n} \cap B_{y_n} = \emptyset$ if $x_n \neq y_n$, $x_n, y_n \in \mathbb{R}$, and $\cup_{x_n \in \mathbb{R}} B_{x_n} = \Omega$. So, using Axiom (ii) for probability laws in Definition 0.2,

$$\begin{aligned} \mathbf{P}(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) &= \mathbf{P}(\{X_1 = x_1\} \cap \cdots \cap \{X_{n-1} = x_{n-1}\} \cap (\cup_{x_n \in \mathbb{R}} B_{x_n})) \\ &= \sum_{x_n \in \mathbb{R}} \mathbf{P}(X_1 = x_1, \ldots, X_n = x_n). \qquad (*) \end{aligned}$$

Similarly,

$$\begin{aligned} \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) &= \mathbf{P}(\cup_{x_n \in \mathbb{R}} B_x) \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) \\ &= \sum_{x_n \in \mathbb{R}} \mathbf{P}(X_n = x_n) \prod_{i=1}^{n-1} \mathbf{P}(X_i = x_i) = \sum_{x_n \in \mathbb{R}} \prod_{i=1}^{n} \mathbf{P}(X_i = x_i). \qquad (**) \end{aligned}$$

So, the quantities $(*)$ and $(**)$ are equal, by assumption. □

Exercise 2.44. Find three random variables X_1, X_2, X_3 such that: X_1 and X_2 are independent; X_1 and X_3 are independent; X_2 and X_3 are independent; but such that X_1, X_2, X_3 are not independent.

Exercise 2.45. Let $0 < p < 1$. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Let $S_n = \sum_{i=1}^{n} X_i$. A moment generating function can help use to compute moments in various ways. Let $t \in \mathbb{R}$ and compute the moment generating function of X_i for each $i \in \{1, \ldots, n\}$. That is, show that

$$\mathbf{E}e^{tX_i} = (1 - p) + pe^t.$$

Then, using the product formula for independent random variables, show that

$$\mathbf{E}e^{tS_n} = [(1 - p) + pe^t]^n.$$

By differentiating the last equality at $t = 0$, and using the power series expansion of the exponential function, compute $\mathbf{E}S_n$ and $\mathbf{E}S_n^2$.

Exercise 2.46. $X_1, \ldots, X_n$ be independent discrete random variables. Show that

$$\mathbf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \leq x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

3. Continuous Random Variables

Up until this point, we have mostly focused on discrete random variables. These random variables take either a finite or countable number of values. However, we are often confronted with a continuous range of possible values. For example, if I throw a dart at a board, then there is a continuous range of places that the dart could land. Or, the price of a stock is (for many purposes) any possible positive real number. We now develop the theory of random variables which take a continuous range of values.

3.1. Continuous Random Variables.

Definition 3.1 (**Probability Density Function, PDF**)**.** A **probability density function** or PDF, is a function $f\colon \mathbb{R} \to [0, \infty)$ such that $\int_{-\infty}^{\infty} f(x)dx = 1$, and such that, for any $-\infty \leq a \leq b \leq \infty$, the integral $\int_a^b f(x)dx$ exists.

Definition 3.2 (**Continuous Random Variable**)**.** A random variable X on a sample space Ω is called **continuous** if there exists a probability density function f_X such that, for any $-\infty \leq a \leq b \leq \infty$, we have

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

We call f_X the **probability density function of** X.

Remark 3.3. Let X be a continuous random variable with density function f_X. Then for any $a \in \mathbb{R}$, $\mathbf{P}(X = a) = \int_a^a f_X(x)dx = 0$. Consequently, for any $-\infty < a \leq b < \infty$, we have

$$\mathbf{P}(a \leq X \leq b) = \mathbf{P}(a \leq X < b) = \mathbf{P}(a < X \leq b) = \mathbf{P}(a < X < b).$$

Remark 3.4. Let $I_1, I_2, \ldots$ be disjoint intervals in the real line $\mathbb{R}$. Let $B = \cup_{i=1}^{\infty} I_i$. Then from Axiom (ii) of Definition 0.2,

$$\mathbf{P}(X \in B) = \mathbf{P}(X \in \cup_{i=1}^{\infty} I_i) = \sum_{i=1}^{\infty} \mathbf{P}(X \in I_i) = \sum_{i=1}^{\infty} \int_{I_i} f_X(x)dx = \int_B f_X(x)dx.$$

The following Theorem is typically proven in advanced analysis classes.

Theorem 3.5 (**Fundamental Theorem of Calculus**)**.** *Let f_X be a probability density function. Then the function $g(t) = \int_{-\infty}^{t} f_X(x)dx$ is continuous at any $t \in \mathbb{R}$. Also, if f_X is continuous at a point x, then g is differentiable at $t = x$, and $g'(x) = f_X(x)$.*

Example 3.6. Let $\Omega = [0,1]$, and define $f_X : \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. Then $\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 dx = 1$, and $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $a \leq b$, we have

$$\mathbf{P}(a \leq X \leq b) = \int_{\max(0,\min(a,1))}^{\max(0,\min(b,1))} dx = \max(0, \min(b,1)) - \max(0, \min(a,1)).$$

In particular, if $0 \leq a < b \leq 1$, we have $\mathbf{P}(a \leq X \leq b) = b - a$. When X has this density function f_X, we say X is **uniformly distributed in** $[0,1]$.

Note that f_X is not a continuous function, but we still say that X is continuous since the function $g(t) = \int_{-\infty}^{t} f_X(x)dx$ is continuous, by the Fundamental Theorem of Calculus. Also, note that f_X only takes two values, but X can take any value in $[0,1]$. Finally, note that g is not differentiable when $t = 0$ or $t = 1$, but g is differentiable for any other $t \in \mathbb{R}$.

Example 3.7. Let $\Omega = [c,d]$, with $-\infty < c < d < \infty$ and define $f_X : \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{d-c}$ when $x \in [c,d]$, and $f_X(x) = 0$ otherwise. Then $\int_{-\infty}^{\infty} f_X(x)dx = \int_c^d \frac{1}{d-c}dx = 1$, and $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $-\infty < a \leq b < \infty$, we have

$$\mathbf{P}(a \leq X \leq b) = \frac{1}{d-c} \int_{\max(c,\min(a,d))}^{\max(c,\min(b,d))} dx = \frac{1}{d-c} \left(\max(c, \min(b,d)) - \max(c, \min(a,d))\right).$$

In particular, if $c \leq a < b \leq d$, we have $\mathbf{P}(a \leq X \leq b) = \frac{b-a}{d-c}$. When X has the density function f_X, we say that X is **uniformly distributed in** $[c,d]$.

Example 3.8. Let $\Omega = \mathbb{R}$, and define $f_X : \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for all $x \in \mathbb{R}$. Then $\int_{-\infty}^{\infty} f_X(x)dx = 1$ by Exercise 3.10 below and $f_X(x) \geq 0$ for all $x \in \mathbb{R}$, so f_X is a probability density function. So, if f_X is the density function of X, and if $-\infty \leq a \leq b \leq \infty$,

$$\mathbf{P}(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

We call X the **standard Gaussian** random variable or the **standard normal** random variable. The distribution f_X resembles a "bell curve."

The Gaussian comes up in many applications, and it has a certain "universality" property which is studied in more advanced probability classes. For example, if we make a histogram of test scores for a class with a large number of people, then the scores will look something like the distribution $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. And we can replace "test scores" with many other things, and the histogram will remain essentially the same. This is what is meant by "universality."

In general, we can intuitively think of a distribution function f_X as a histogram for the (random) values that X takes.

Example 3.9. Let $\lambda > 0$. Define $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and $f_X(x) = 0$ otherwise. Let's check that f_X satisfies Definition 3.1.

$$\int_{-\infty}^{\infty} f_X(x)dx = \lambda \int_0^{\infty} e^{-\lambda x} dx = \lambda \lim_{N \to \infty} [-\lambda^{-1}(e^{-\lambda N} - 1)] = 1.$$

A random variable X with this density f_X is called an **exponential random variable with parameter** λ. Exponential random variables can be used to model the expiration time of lightbulbs, or other electronic equipment.

Exercise 3.10. Verify that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$.

3.1.1. *Expected Value.* How should we define the expected value of a continuous random variable? Let's return to Example 3.6. Let $\Omega = [0,1]$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. Then X is uniformly distributed in $[0,1]$. Let n be a positive integer. We will try to approximate the expected value of X. Consider the intervals $[0,1/n)$, $[1/n,2/n)$, ..., $[(n-1)/n,1)$. Then, for each $i \in \{1,\dots,n\}$,

$$\mathbf{P}(X \in [(i-1)/n, i/n)) = \int_{(i-1)/n}^{i/n} dx = 1/n.$$

So, to estimate the expected value of X, let's just make the approximation that X takes the value i/n with probability $1/n$, for each $i \in \{1,\dots,n\}$. This is not quite true, but it is also not so far from the truth. Then we estimate the expected value of X by summing up the (approximate) values of X, multiplied by their probabilities of occurring:

$$\sum_{i=1}^{n} \frac{i}{n} \cdot \mathbf{P}(X \in [(i-1)/n, i/n)) = \sum_{i=1}^{n} \frac{i}{n} \frac{1}{n}.$$

We could compute this sum exactly, but it is perhaps better to see that this sum is a Riemann sum for the function $g(x) = x$ on the interval $[0,1]$. That is,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{i}{n} \cdot \mathbf{P}(X \in [(i-1)/n, i/n)) = \int_0^1 x dx = \int_{-\infty}^{\infty} x f_X(x) dx.$$

The last expression is exactly our definition of expected value for continuous random variables.

Definition 3.11 (**Expected Value**)**.** Let X be a continuous random variable with density function f_X. Assume that $\int_{-\infty}^{\infty} |x| \, f_X(x) dx < \infty$. We then define the **expected value of** X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

Let $g \colon \mathbb{R} \to \mathbb{R}$ be a function. We define

$$\mathbf{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

In particular, if n is a positive integer, we have

$$\mathbf{E}(X^n) = \int_{-\infty}^{\infty} x^n f_X(x) dx.$$

Comparing Definition 2.1 to Definition 3.11, we see that we have essentially replaced the sums with integrals. Also, we can use the same definition of variance as before.

Definition 3.12 (**Variance**)**.** Let X be a continuous random variable with density function f_X. We define the **variance of** X, denoted $\mathrm{var}(X)$, by

$$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2.$$

Many facts for discrete random variables also apply to continuous random variables. For example, the following restatements of Propositions 2.5 and 2.7 hold, with the same proof as before, where we replace the sums by integrals.

Proposition 3.13 (**Properties of Expected Value**). *Let X be a continuous random variable. Let a, b be constants. Then*

$$\mathbf{E}(aX + b) = a\mathbf{E}(X) + b.$$

Proof. Using Definition 3.11 and Definition 3.1

$$\mathbf{E}(aX + b) = \int_{-\infty}^{\infty} (ax + b) f_X(x) dx = a \int_{-\infty}^{\infty} x f_X(x) + b \int_{-\infty}^{\infty} f_X(x) dx = a\mathbf{E}(X) + b \cdot 1.$$

□

Proposition 3.14 (**Properties of Variance**). *Let X be a continuous random variable. Let a, b be constants. Then*

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.$$

Moreover,

$$\mathrm{var}(aX + b) = a^2 \mathrm{var}(X).$$

Proof. Using Definition 3.11 and Definition 3.1,

$$\begin{aligned}
\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2 &= \int_{-\infty}^{\infty} (x - \mathbf{E}(X))^2 f_X(x) dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2\mathbf{E}(X) \int_{-\infty}^{\infty} x f_X(x) dx + (\mathbf{E}(X))^2 \int_{-\infty}^{\infty} f_X(x) dx \\
&= \mathbf{E}(X^2) - 2\mathbf{E}(X)\mathbf{E}(X) + (\mathbf{E}(X))^2 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2.
\end{aligned}$$

From Proposition 3.13, $\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$. So, using Definition 3.11,

$$\begin{aligned}
\mathrm{var}(aX + b) &= \mathbf{E}(aX + b - (a\mathbf{E}(X) + b))^2 = \mathbf{E}(aX - a\mathbf{E}(X))^2 = \mathbf{E}(a^2(X - \mathbf{E}(X))^2) \\
&= a^2 \mathbf{E}(X - \mathbf{E}(X))^2 = a^2 \mathrm{var}(X).
\end{aligned}$$

□

Example 3.15. We revisit Example 3.6. Let $\Omega = [0, 1]$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = 1$ when $x \in [0, 1]$, and $f_X(x) = 0$ otherwise. Then X is uniformly distributed in $[0, 1]$. We compute

$$\mathbf{E}(X) = \int_0^1 x dx = \frac{1}{2}, \qquad \mathbf{E}(X^2) = \int_0^1 x^2 dx = \frac{1}{3}.$$

$$\mathrm{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

In particular, if X is uniformly distributed in $[0, 1]$, then the average value of X is $1/2$.

Example 3.16. We revisit Example 3.8. Let $\Omega = \mathbb{R}$, and define $f_X \colon \mathbb{R} \to \mathbb{R}$ so that $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for all $x \in \mathbb{R}$. Then X is a standard Gaussian random variable. We compute

$$\mathbf{E}(X) = \int_{-\infty}^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \int_0^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} - \int_0^{\infty} x e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = 0.$$

Exercise 3.17. Let X be a continuous random variable with distribution function $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, $\forall\ x \in \mathbb{R}$. Show that $\mathrm{var}(X) = 1$.

Example 3.18. We reconsider Example 3.9. Let $\lambda > 0$. Define $f_X(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, and $f_X(x) = 0$ otherwise. Then X is an **exponential random variable with parameter** λ. Using integration by parts, we compute

$$\mathbf{E}(X) = \lambda \int_0^\infty x e^{-\lambda x} dx = -\int_0^\infty x \frac{d}{dx} e^{-\lambda x} dx = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

$$\mathbf{E}(X^2) = \lambda \int_0^\infty x^2 e^{-\lambda x} dx = -\int_0^\infty x^2 \frac{d}{dx} e^{-\lambda x} dx = \int_0^\infty 2x \frac{d}{dx} e^{-\lambda x} dx = \frac{2}{\lambda}\mathbf{E}(X) = \frac{2}{\lambda^2}.$$

$$\operatorname{var}(X) = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Exercise 3.19. Let X be a random variable such that $f_X(x) = x$ when $0 \leq x \leq \sqrt{2}$ and $f_X(x) = 0$ otherwise. Compute $\mathbf{E}X^2$ and $\mathbf{E}X^3$.

3.2. **Cumulative Distribution Function (CDF).** Our treatments of discrete and continuous random variables have been similar but different. We had to repeat ourselves several times, and some concepts seem similar but not identical. Thankfully, a unified treatment of both discrete and continuous random variables can be done. This unified treatment comes from examining the probability that a random variable X satisfies $\mathbf{P}(X \leq x)$, for any $x \in \mathbb{R}$.

Definition 3.20 (**Cumulative Distribution Function**)**.** Let X be a random variable. The **cumulative distribution function of** X, denoted F_X, is a function $F_X \colon \mathbb{R} \to [0,1]$ defined by

$$F_X(x) = \mathbf{P}(X \leq x), \qquad \forall\, x \in \mathbb{R}.$$

Remark 3.21. If X is a discrete random variable, then

$$F_X(x) = \mathbf{P}(X \leq x) = \sum_{y \in \mathbb{R} \colon\, y \leq x} p_X(y).$$

If X is a continuous random variable with density function f_X, then

$$F_X(x) = \mathbf{P}(X \leq x) = \int_{-\infty}^x f_X(t) dt.$$

Proposition 3.22 (**Properties of the Distribution Function**)**.** *Let X be a random variable. The cumulative distribution function F_X satisfies the following properties:*

- *If $x \leq y$, then $F_X(x) \leq F_X(y)$.*
- $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$.
- *If X is discrete, then F_X is piecewise constant.*
- *If X is continuous, then F_X is continuous.*

Remark 3.23. If X is a continuous random variable with probability density function f_X, and if f_X is continuous at a point $x \in \mathbb{R}$, then Theorem 3.5 implies that $\frac{d}{dx} F_X(x) = f_X(x)$.

Example 3.24. Let X be a uniformly distributed random variable in $[0,1]$. In Example 3.6, we showed that X has the distribution function f_X where $f_X(x) = 1$ when $x \in [0,1]$, and $f_X(x) = 0$ otherwise. So,

$$F_X(x) = \int_{-\infty}^x f_X(t) dt = \int_0^{\max(0,\min(x,1))} dt = \begin{cases} x, & x \in [0,1] \\ 0, & x < 0 \\ 1, & x > 1. \end{cases}$$

Note also that

$$\frac{d}{dx}F_X(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & x < 0 \text{ or } x > 1 \\ \text{undefined}, & x = 0 \text{ or } x = 1 \end{cases}$$

So, the derivative of F_X may not exist at some points, but $\frac{d}{dx}F_X(x) = f_X(x)$ for any $x \in (-\infty, 0) \cup (0,1) \cup (1, \infty)$.

Example 3.25 (**Maximum of Independent Variables**)**.** Let X_1, X_2 be two independent discrete random variable with identical CDFs. That is, $\mathbf{P}(X_1 \leq x) = \mathbf{P}(X_2 \leq x)$ for all $x \in \mathbb{R}$. Define the random variable Y by

$$Y = \max(X_1, X_2).$$

Using Exercise 2.46, for any $x \in \mathbb{R}$, we have

$$\mathbf{P}(Y \leq x) = \mathbf{P}(X_1 \leq x, X_2 \leq x) = \mathbf{P}(X_1 \leq x)\mathbf{P}(X_2 \leq x) = [\mathbf{P}(X_1 \leq x)]^2.$$

That is, the CDF of Y is the square of the CDF of X_1.

More generally, if $X_1, X_2, \ldots, X_n$ are independent, discrete random variable with identical CDFs, and if

$$Y = \max(X_1, \ldots, X_n),$$

then for any $x \in \mathbb{R}$,

$$\mathbf{P}(Y \leq x) = [\mathbf{P}(X_1 \leq x)]^n.$$

We can think of Y as the maximum score on a test with n test takers, or the longest throw of a shot put, etc.

Example 3.26. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $p = 1/2$, so that $\mathbf{P}(X_i = 0) = \mathbf{P}(X_i = 1) = 1/2$ for all $1 \leq i \leq n$. Also,

$$\mathbf{P}(X_1 \leq x) = \begin{cases} 0 & , \text{ if } x < 0 \\ 1/2 & , \text{ if } 0 \leq x < 1 \\ 1 & , \text{ if } x \geq 1 \end{cases}.$$

Let $Y = \max(X_1, \ldots, X_n)$. Then

$$\mathbf{P}(Y \leq x) = [\mathbf{P}(X_1 \leq x)]^n = \begin{cases} 0 & , \text{ if } x < 0 \\ (1/2)^n & , \text{ if } 0 \leq x < 1 \\ 1 & , \text{ if } x \geq 1 \end{cases}.$$

That is, $p_Y(0) = (1/2)^n$ and $p_Y(1) = 1 - (1/2)^n$. That is, Y is a Bernoulli random variable with parameter $1 - (1/2)^n$.

Exercise 3.27. Let X be a random variable such that $X = 1$ with probability 1. Show that X is not a continuous random variable. That is, there does not exist a probability density function f such that $\mathbf{P}(X \leq a) = \int_{-\infty}^{a} f(x)dx$ for all $x \in \mathbb{R}$. (Hint: if X were continuous, then the function $g(t) = \int_{-\infty}^{a} f(x)dx$ would be continuous, by the Fundamental Theorem of Calculus.)

3.3. **Normal Random Variables.**

Definition 3.28 (**Normal Random Variable**)**.** Let $\mu \in \mathbb{R}$, $\sigma > 0$. A continuous random variable X is said to be **normal** or **Gaussian** with mean μ and variance σ^2 if X has the following PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall\, x \in \mathbb{R}.$$

In particular, a **standard normal** or **standard Gaussian** random variable is defined to be a normal with $\mu = 0$ and $\sigma = 1$.

Exercise 3.29. Verify that a Gaussian random variable X with mean μ and variance σ^2 actually has mean μ and variance σ^2.

Let $a, b \in \mathbb{R}$ with $a \neq 0$. Show that $aX + b$ is a normal random variable with mean $a\mu + b$ and variance $a^2\sigma^2$.

In particular, conclude that $(X - \mu)/\sigma$ is a standard normal.

The Gaussian is probably one of the single most useful random variables within mathematics, and within applications of mathematics. Here is a sample result that shows the usefulness of the Gaussian.

Theorem 3.30 (**De Moivre-Laplace Theorem**)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $1/2$. Recall that X_1 has mean $1/2$ and variance $1/4$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - (1/2)n}{\sqrt{n}\sqrt{1/4}} \leq a \right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

That is, when n is large, the CDF of $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ is roughly the same as that of a standard normal. In particular, if you flip n fair coins, then the number of heads you get should typically be in the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, when n is large.

Remark 3.31. The random variable $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ has mean zero and variance 1, just like the standard Gaussian. So, the normalizations of $X_1 + \cdots + X_n$ we have chosen are sensible. Also, to explain the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, note that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2} \right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-\frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n - \frac{n}{2} \leq \frac{\sqrt{n}}{2} \right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-1 \leq \frac{X_1 + \cdots + X_n - \frac{n}{2}}{\sqrt{n}/2} \leq 1 \right) = \int_{-1}^{1} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} \approx .6827.
\end{aligned}$$

In fact, there is nothing special about the parameter $1/2$ in the above theorem.

Theorem 3.32 (**De Moivre-Laplace Theorem, Second Version**)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Recall that X_1 has mean p and variance $p(1-p)$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - pn}{\sqrt{n}\sqrt{p(1-p)}} \leq a \right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

In fact, there is nothing special about Bernoulli random variables in the above theorem. (See the Central Limit Theorem in Theorem 5.20 below.)

Exercise 3.33. Using the De Moivre-Laplace Theorem, estimate the probability that 10^6 coin flips of fair coins will result in more than $501,000$ heads. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

3.4. **Joint PDFs.**

Definition 3.34 (**Joint Probability Density Function, Two Variables**)**.** A **joint probability density function (PDF)** for two random variables is a function $f\colon \mathbb{R}^2 \to [0,\infty)$ such that $\iint_{\mathbb{R}^2} f(x,y)dxdy = 1$, and such that, for any $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, the integral $\int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy$ exists.

Definition 3.35. Let X, Y be two continuous random variables on a sample space Ω. We say that X and Y are **jointly continuous** with **joint PDF** $f_{X,Y}\colon \mathbb{R}^2 \to [0,\infty)$ if, for any subset $A \subseteq \mathbb{R}^2$, we have

$$\mathbf{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)dxdy.$$

In particular, choosing $A = [a,b] \times [c,d]$ with $-\infty \le a < b \le \infty$ and $-\infty \le c < d \le \infty$, we have

$$\mathbf{P}(a \le X \le b, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=a}^{x=b} f_{X,Y}(x,y)dxdy.$$

We define the **marginal PDF** f_X of X by

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

We define the **marginal PDF** f_Y of Y by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx, \qquad \forall\, y \in \mathbb{R}.$$

Note that

$$\mathbf{P}(c \le Y \le d) = \mathbf{P}(-\infty \le X \le \infty, c \le Y \le d) = \int_{y=c}^{y=d}\int_{x=-\infty}^{x=\infty} f_{X,Y}(x,y)dxdy.$$

Comparing this formula with Definition 3.2, we see that the marginal PDF of Y is exactly the PDF of Y. Similarly, the marginal PDF of X is the PDF of X.

Example 3.36. In this example, we take it as given that

$$\frac{1}{2\pi}\iint_{\mathbb{R}^2} e^{-(x^2+y^2)/2}dxdy = 1.$$

Suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y)\in A)=\frac{1}{2\pi}\iint_A e^{-(x^2+y^2)/2}dxdy.$$

That is, we can think of X as the x-coordinate of a randomly thrown dart, and we can think of Y as the y-coordinate of a randomly thrown dart on the infinite dartboard $\mathbb{R}^2$.

In this case, the marginals are both standard Gaussians:

$$f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\int_{-\infty}^{\infty}e^{-y^2/2}\frac{dy}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}e^{-x^2/2},\qquad\forall x\in\mathbb{R}.$$
$$f_Y(y)=\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\int_{-\infty}^{\infty}e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}=\frac{1}{\sqrt{2\pi}}e^{-y^2/2},\qquad\forall y\in\mathbb{R}.$$

That is, if we only keep track of the x-coordinate of the random dart, then this x-coordinate is a standard Gaussian itself. And if we only keep track of the y-coordinate of the random dart, then this y-coordinate is also a standard Gaussian.

Example 3.37 (**Buffon's Needle**)**.** Suppose a needle of length $\ell>0$ is kept parallel to the ground. The needle is dropped onto the ground with a random position and orientation. The ground has a grid of equally spaced horizontal lines, where the distance between two adjacent lines is $d>0$. Suppose $\ell<d$. What is the probability that the needle touches one of the lines? (Since $\ell<d$, the needle can touch at most one line.)

Let x be the distance of the midpoint of the needle from the closest line. Let θ be the acute angle formed by the needle and any horizontal line. The tip of the needle exactly touches the line when $\sin\theta=x/(\ell/2)=2x/\ell$. So, any part of the needle touches some line if and only if $x\le(\ell/2)\sin\theta$. Since the needle has a uniformly random position and orientation, we model X,Θ as random variables with joint distribution uniform on $[0,d/2]\times[0,\pi/2]$. So,

$$f_{X,\Theta}(x,\theta)=\begin{cases}\frac{4}{\pi d}, & x\in[0,d/2]\text{ and }\theta\in[0,\pi/2]\\ 0, & \text{otherwise.}\end{cases}$$

(Note that $\iint_{\mathbb{R}^2}f_{X,\Theta}(x,\theta)dxd\theta=1$.) And the probability that the needle touches one of the lines is

$$\begin{aligned}\iint_{0\le x\le(\ell/2)\sin\theta}f_{X,\Theta}(x,\theta)dxd\theta&=\int_{\theta=0}^{\theta=\pi/2}\int_{x=0}^{x=(\ell/2)\sin\theta}\frac{4}{\pi d}dxd\theta\\&=\frac{2\ell}{\pi d}\int_{\theta=0}^{\theta=\pi/2}\sin\theta d\theta=\frac{2\ell}{\pi d}[-\cos\theta]_{\theta=0}^{\theta=\pi/2}=\frac{2\ell}{\pi d}.\end{aligned}$$

Note that $x\le\ell/2<d/2$ always, so the set $0\le x\le(\ell/2)\sin\theta$ is still contained in the set $x\in[0,d/2]$.

In particular, when $\ell=d$, the probability is $2/\pi$.

Definition 3.38. Let X,Y be random variables with joint PDF $f_{X,Y}$. Let $g\colon\mathbb{R}^2\to\mathbb{R}$. Then

$$\mathbf{E}g(X,Y)=\iint_{\mathbb{R}^2}g(x,y)f_{X,Y}(x,y)dxdy.$$

In particular,

$$\mathbf{E}(XY)=\iint_{\mathbb{R}^2}xyf_{X,Y}(x,y)dxdy.$$

Exercise 3.39. Let X, Y be random variables with joint PDF $f_{X,Y}$. Let $a, b \in \mathbb{R}$. Using Definition 3.38, show that $\mathbf{E}(aX + bY) = a\mathbf{E}X + b\mathbf{E}Y$.

Theorem 3.40 (Fubini Theorem). *Let $h\colon \mathbb{R}^2 \to \mathbb{R}$ be a continuous function such that $\iint_{\mathbb{R}^2} |h(x,y)|\, dxdy < \infty$. Then*

$$\iint_{\mathbb{R}^2} h(x,y)dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y)dx \right) dy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y)dy \right) dx.$$

Exercise 3.41. Let X, Y be random variables. For any $y \in \mathbb{R}$, assume that $\mathbf{E}(X|Y = y) = e^{-|y|}$. Also, assume that Y has an exponential distribution with parameter $\lambda = 2$. Compute $\mathbf{E}X$.

3.5. Independence.

Definition 3.42. Let X, Y be random variables with joint PDF $f_{X,Y}$. We say that X and Y are **independent** if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \qquad \forall\, x, y \in \mathbb{R}.$$

More generally, random variables $X_1, \ldots, X_n$ with joint PDF $f_{X_1,\ldots,X_n}$ are **independent** if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \qquad \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Example 3.43. We continue Example 3.36. We suppose X and Y have a joint PDF so that

$$\mathbf{P}((X,Y) \in A) = \frac{1}{2\pi} \iint_A e^{-(x^2+y^2)/2} dxdy, \qquad \forall\, A \subseteq \mathbb{R}^2.$$

We showed in Example 3.36 that X and Y are both standard normals. We verified in Example 3.36 that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ forall $x, y \in \mathbb{R}$. So, X and Y are independent.

Proposition 3.44. *Let X, Y be two independent random variables with joint PDF $f_{X,Y}$. Let $A, B \subseteq \mathbb{R}$. Then the events $\{X \in A\}$ and $\{Y \in B\}$ are independent.*

Proof. Using Definition 3.42 and Theorem 3.40,

$$\begin{aligned}\mathbf{P}(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y)dydx = \int_A \int_B f_Y(y)dy f_X(x)dx \\ &= (\int_A f_X(x)dx)(\int_B f_Y(y)dy) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B).\end{aligned}$$

□

Theorem 3.45. *Let X, Y be two independent random variables with joint PDF $f_{X,Y}$. Then*

$$\mathbf{E}(XY) = (\mathbf{E}X)(\mathbf{E}Y).$$

More generally, if $g, h\colon \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}(g(X)h(Y)) = (\mathbf{E}g(X))(\mathbf{E}h(Y)).$$

More generally, if $X_1, \ldots, X_n$ are independent random variables with joint PDF $f_{X_1,\ldots,X_n}$, and if $g_1, \ldots, g_n\colon \mathbb{R} \to \mathbb{R}$, then

$$\mathbf{E}(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} \mathbf{E}(g_i(X_i)).$$

Proof. We prove the second statement since it implies the first. Using Definitions 3.38 and 3.42, and Theorem 3.40

$$E(g(X)h(Y)) = \iint_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)dxdy = \iint_{\mathbb{R}^2} g(x)h(y)f_X(x)f_Y(y)dxdy$$
$$= (\int_{\mathbb{R}} g(x)f_X(x)dx)(\int_{\mathbb{R}} h(y)f_Y(y)dy) = (\mathbf{E}g(X))(\mathbf{E}h(Y)).$$

□

Exercise 3.46. Let X, Y be independent random variables with joint PDF $f_{X,Y}$. Show that

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y).$$

Exercise 3.47. Let X and Y be uniformly distributed random variables on $[0,1]$. Assume that X and Y are independent. Compute the following probabilities:

- $\mathbf{P}(X > 3/4)$
- $\mathbf{P}(Y < X)$
- $\mathbf{P}(X + Y < 1/2)$
- $\mathbf{P}(\max(X,Y) > 1/2)$
- $\mathbf{P}(XY < 1/3)$.

Exercise 3.48. Let X_1, Y_1 be random variables with joint PDF f_{X_1,Y_1}. Let X_2, Y_2 be random variables with joint PDF f_{X_2,Y_2}. Let $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ and let $S\colon \mathbb{R}^2 \to \mathbb{R}^2$ so that $ST(x,y) = (x,y)$ and $TS(x,y) = (x,y)$ for every $(x,y) \in \mathbb{R}^2$. Let $J(x,y)$ denote the determinant of the Jacobian of S at (x,y). Using the change of variables formula from multivariable calculus, show that

$$f_{X_2,Y_2}(x,y) = f_{X_1,Y_1}(S(x,y))\,|J(x,y)|\,.$$

Exercise 3.49 (Numerical Integration). In computer graphics in video games, etc., various integrations are performed in order to simulate lighting effects. Here is a way to use random sampling to integrate a function in order to quickly and accurately render lighting effects. Let $\Omega = [0,1]$, and let $\mathbf{P}$ be the uniform probably law on Ω, so that if $0 \le a < b \le 1$, we have $\mathbf{P}([a,b]) = b - a$. Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbf{P}(X_i \in [a,b]) = b - a$ for all $0 \le a < b \le 1$, for all $i \in \{1, \ldots, n\}$. Let $f\colon [0,1] \to \mathbb{R}$ be a continuous function we would like to integrate. Instead of integrating f directly, we instead compute the quantity

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Show that

$$\lim_{n\to\infty} \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \int_0^1 f(t)dt.$$

$$\lim_{n\to\infty} \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = 0.$$

That is, as n becomes large, $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ is a good estimate for $\int_0^1 f(t)dt$.

3.6. **Joint CDF.**

Definition 3.50 (**Joint CDF**)**.** Let X, Y be random variables. We define the **joint CDF** of X, Y to be the function

$$F_{X,Y}(x,y) = \mathbf{P}(X \le x, Y \le y), \quad \forall\, x, y \in \mathbb{R}.$$

More generally, if $X_1, \dots, X_n$ are random variables, we define the **joint CDF** of $X_1, \dots, X_n$ to be the function

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbf{P}(X_1 \le x_1, \dots, X_n \le x_n), \quad \forall\, x_1, \dots, x_n \in \mathbb{R}.$$

Remark 3.51. If X, Y are independent random variables with joint PDF $f_{X,Y}$, then Proposition 3.44 says that

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

More generally, if $X_1, \dots, X_n$ are independent random variables with joint PDF $f_{X_1,\dots,X_n}$, then

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \quad \forall\, x_1, \dots, x_n \in \mathbb{R}.$$

Remark 3.52. In fact, we can use the last equality as a *definition* in order to define independence of general random variables. That is, we say random variables $X_1, \dots, X_n$ are independent if

$$F_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i), \quad \forall\, x_1, \dots, x_n \in \mathbb{R}.$$

4. Limit Theorem Preliminaries: Covariance, Transforms

4.1. **Introduction to Limit Theorems.** Suppose I flip a fair coin 10^9 times. Then I should expect to get roughly $\frac{1}{2}10^9$ heads and $\frac{1}{2}10^9$ tails. This is formalized in the Law of Large Numbers. Or, suppose I have a casino game where the casino wins 51% of the time. Then over a long period of time, the casino will make money; the Law of Large Numbers guarantees that! However, if I do flip 10^9 fair coins, it is unlikely that I will get *exactly* $\frac{1}{2}10^9$ heads. (What is the exact probability?) There will typically be some small fluctuations around $\frac{1}{2}10^9$. But about how close to $\frac{1}{2}10^9$ will the number of heads be? This question is answered precisely by the Central Limit Theorem. In your previous probability class, you may have mentioned the Central Limit Theorem applied to coin flips, which is known as the De Moivre-Laplace Theorem:

Theorem 4.1 (**De Moivre-Laplace Theorem**)**.** *Let $X_1, \dots, X_n$ be independent Bernoulli random variables with parameter $1/2$, so that $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = 0) = 1/2$. Recall that X_1 has mean $1/2$ and variance $1/4$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - (1/2)n}{\sqrt{n}\sqrt{1/4}} \le a\right) = \int_{-\infty}^{a} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}}.$$

That is, when n is large, the CDF of $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ is roughly the same as that of a standard normal. In particular, if you flip n fair coins, then the number of heads you get should typically be in the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, when n is large.

Remark 4.2. The random variable $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ has mean zero and variance 1, just like the standard Gaussian. So, the normalizations of $X_1+\cdots+X_n$ we have chosen are sensible. Also, to explain the interval $(n/2-\sqrt{n}/2, n/2+\sqrt{n}/2)$, note that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left(\frac{n}{2}-\frac{\sqrt{n}}{2} \le X_1+\cdots+X_n \le \frac{n}{2}+\frac{\sqrt{n}}{2}\right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-\frac{\sqrt{n}}{2} \le X_1+\cdots+X_n-\frac{n}{2} \le \frac{\sqrt{n}}{2}\right) \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left(-1 \le \frac{X_1+\cdots+X_n-\frac{n}{2}}{\sqrt{n}/2} \le 1\right) = \int_{-1}^{1} e^{-t^2/2}\frac{dt}{\sqrt{2\pi}} \approx .6827.
\end{aligned}$$

Exercise 4.3. Let X and Y be nonnegative random variables. Recall that we can define

$$\mathbf{E}X := \int_0^\infty \mathbf{P}(X>t)dt.$$

Assume that $X \le Y$. Conclude that $\mathbf{E}X \le \mathbf{E}Y$.

More generally, if X satisfies $\mathbf{E}\,|X| < \infty$, we define $\mathbf{E}X := \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0)$. If X, Y are any random variables with $X \le Y$, $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|Y| < \infty$, show that $\mathbf{E}X \le \mathbf{E}Y$.

4.2. **Covariance.** Recall that the covariance of two random variables X and Y, denoted $\mathrm{cov}(X,Y)$, is

$$\mathrm{cov}(X,Y) = \mathbf{E}((X-\mathbf{E}(X))(Y-\mathbf{E}(Y))).$$

In particular, $\mathrm{cov}(X,X) = \mathbf{E}(X-\mathbf{E}(X))^2 = \mathrm{var}(X)$.

Definition 4.4. Let X, Y be random variables. We say that X, Y are **uncorrelated** if $\mathrm{cov}(X,Y) = 0$.

Exercise 4.5. Let X, Y be random variables with $\mathbf{E}X^2 < \infty$ and $\mathbf{E}Y^2 < \infty$. Prove the **Cauchy-Schwarz inequality**:

$$\mathbf{E}(XY) \le (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Then, deduce the following when X, Y both have finite variance:

$$|\mathrm{cov}(X,Y)| \le (\mathrm{var}(X))^{1/2}(\mathrm{var}(Y))^{1/2}.$$

(Hint: in the case that $\mathbf{E}Y^2 > 0$, expand out the product $\mathbf{E}(X - Y\mathbf{E}(XY)/\mathbf{E}Y^2)^2$.)

Recall in Lemma 2.28, we proved the following for discrete random variables, though the proof applies for any random variables.

Lemma 4.6. *Let $X_1, \ldots, X_n$ be random variables with $\mathrm{var}(X_i) < \infty$ for all $1 \le i \le n$. Then*

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i) + 2\sum_{1\le i<j\le n} \mathrm{cov}(X_i, X_j).$$

Proof.

$$\begin{aligned}\operatorname{var}(\sum_{i=1}^n X_i) &= \mathbf{E}(\sum_{i=1}^n X_i - \mathbf{E}(\sum_{i=1}^n X_i))^2 = \mathbf{E}(\sum_{i=1}^n (X_i - \mathbf{E}(X_i)))^2 \\ &= \mathbf{E}\left(\sum_{i=1}^n (X_i - \mathbf{E}(X_i))^2\right) + 2\mathbf{E}\left(\sum_{1\le i<j\le n} (X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))\right) \\ &= \sum_{i=1}^n \operatorname{var}(X_i) + 2\sum_{1\le i<j\le n} \operatorname{cov}(X_i, X_j).\end{aligned}$$

The assumption $\operatorname{var}(X_i) < \infty$ for all $1 \le i \le n$ and Exercise 4.5 ensure that all of the above quantities are finite. □

As in Corollary 2.35, Lemma 4.6 immediately implies:

Corollary 4.7. *Let $X_1, \dots, X_n$ be random variables that are pairwise uncorrelated. That is, $\operatorname{cov}(X_i, X_j) = 0$ for any $i, j \in \{1, \dots, n\}$ with $i \ne j$. Then*

$$\operatorname{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \operatorname{var}(X_i).$$

Corollary 4.8. *Let $X_1, \dots, X_n$ be independent random variables. Then*

$$\operatorname{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \operatorname{var}(X_i).$$

Proof. Let $i, j \in \{1, \dots, n\}$ with $i \ne j$. Then, using independence,

$$\operatorname{cov}(X_i, X_j) = \mathbf{E}((X_i - \mathbf{E}(X_i))(X_j - \mathbf{E}(X_j))) = \mathbf{E}(X_i X_j) - 2\mathbf{E}(X_i)\mathbf{E}(X_j) + \mathbf{E}(X_i)\mathbf{E}(X_j) = 0.$$

So, Corollary 4.7 concludes the proof. □

Exercise 4.9. Let X be a binomial random variable with parameters $n = 2$ and $p = 1/2$. So, $\mathbf{P}(X = 0) = 1/4$, $\mathbf{P}(X = 1) = 1/2$ and $\mathbf{P}(X = 2) = 1/4$. And X satisfies $\mathbf{E}X = 1$ and $\mathbf{E}X^2 = 3/2$.

Let Y be a geometric random variable with parameter $1/2$. So, for any positive integer k, $\mathbf{P}(Y = k) = 2^{-k}$. And Y satisfies $\mathbf{E}Y = 2$ and $\mathbf{E}Y^2 = 6$.

Let Z be a Poisson random variable with parameter 1. So, for any nonnegative integer k, $\mathbf{P}(Z = k) = \frac{1}{e}\frac{1}{k!}$. And Z satisfies $\mathbf{E}Z = 1$ and $\mathbf{E}Z^2 = 2$.

Let W be a discrete random variable such that $\mathbf{P}(W = 0) = 1/2$ and $\mathbf{P}(W = 4) = 1/2$, so that $\mathbf{E}W = 2$ and $\mathbf{E}W^2 = 8$.

Assume that X, Y, Z and W are all independent. Compute

$$\operatorname{var}(X + Y + Z + W).$$

Exercise 4.10. Let $X_1, \dots, X_n$ be random variables with finite variance. Define an $n \times n$ matrix A such that $A_{ij} = \operatorname{cov}(X_i, X_j)$ for any $1 \le i, j \le n$. Show that the matrix A is positive semidefinite. That is, show that for any $y = (y_1, \dots, y_n) \in \mathbb{R}^n$, we have

$$y^T A y = \sum_{i,j=1}^n y_i y_j A_{ij} \ge 0.$$

4.3. **Transforms.** Generally speaking, a transform is a way of creating one function from another function. For example, the moment generating function associates a real-valued function to a random variable. And the characteristic function (or Fourier transform) associates a complex-valued function to a random variable.

Definition 4.11 (**Moment Generating Function**)**.** Let X be a random variable. The **moment generating function** of X is a function $M_X\colon \mathbb{R} \to \mathbb{R}$ defined by

$$M_X(t) := \mathbf{E}(e^{tX}), \quad \forall\, t \in \mathbb{R}.$$

Remark 4.12. For certain random variables X, the moment generating function may not exist. For example, if X is a continuous random variable with density function $f_X(x) = x^{-2}$ for any $x > 1$, and $f_X(x) = 0$ otherwise. Then $M_X(t) = \int_1^\infty e^{tx} f_X(x)dx$ does not exist when $t > 0$.

Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value. Then

$$\frac{d}{dt}|_{t=0} M_X(t) = \mathbf{E}\left(\frac{d}{dt}_{t=0} e^{tX}\right) = \mathbf{E}(X).$$

That is, the first derivative of the moment generating function at $t = 0$ is equal to the first moment of X. More generally, the n^{th} derivative of the moment generating function at $t = 0$ is equal to the n^{th} moment of X:

Exercise 4.13. Let X be a random variable. Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value any number of times. For any positive integer n, show that

$$\frac{d^n}{dt^n}|_{t=0} M_X(t) = \mathbf{E}(X^n).$$

So, in principle, all moments of X can be computed just by taking derivatives of the moment generating function.

Example 4.14. Let X be an exponential random variable with parameter $\lambda > 0$. That is, $f_X(x) = \lambda e^{-\lambda x}$ for any $x \geq 0$, and $f_X(x) = 0$ otherwise. Then for any $t < \lambda$,

$$\begin{aligned} M_X(t) &= \lambda \int_0^\infty e^{tx} e^{-\lambda x} dx = \lambda \int_0^\infty e^{(t-\lambda)x} dx \\ &= \lambda \lim_{N\to\infty} \frac{1}{t-\lambda} [e^{(t-\lambda)x}]_{x=0}^{x=N} = \frac{\lambda}{\lambda - t}. \end{aligned}$$

From Exercise 4.13, $\mathbf{E}X = \frac{d}{dt}|_{t=0} M_X(t) = \frac{\lambda}{\lambda^2} = \lambda^{-1}$. More generally, it follows by induction that for any integer $n > 0$,

$$\mathbf{E}X^n = \frac{d^n}{dt^n}|_{t=0} M_X(t) = n!\lambda^{-n}.$$

Instead of proving this equality by induction, we use power series. Let $t \in \mathbb{R}$ with $|t| < 1$. From the summation formula for geometric series,

$$\frac{1}{1-t} = \sum_{k=0}^{\infty} t^n.$$

That is, for any $t \in \mathbb{R}$ with $|t| < \lambda$,

$$M_X(t) = \frac{\lambda}{\lambda - t} = \frac{1}{1 - (t/\lambda)} = \sum_{k=0}^{\infty} (t/\lambda)^k.$$

So, from Exercise 4.13, if n is a positive integer, then

$$\mathbf{E}X^n = \frac{d^n}{dt^n}|_{t=0} M_X(t) = \sum_{k=0}^{\infty} \frac{d^n}{dt^n}|_{t=0} (t/\lambda)^k = \frac{d^n}{dt^n}|_{t=0} (t/\lambda)^n = n!\lambda^{-n}.$$

The Gaussian density has a fairly simple moment generating function.

Proposition 4.15. *Let X be a standard Gaussian random variable. Then*

$$M_X(t) = e^{t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

Proof.

$$\begin{aligned} M_X(t) = \mathbf{E}e^{tX} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} e^{t^2/2} dx \\ &= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx = e^{t^2/2}. \end{aligned}$$

□

Exercise 4.16. Using the explicit formula for the moment generating function, compute an explicit formula for all moments of the Gaussian random variable. (The $2n^{th}$ moment of X should be something resembling a factorial.)

Proposition 4.17. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$M_{X_1 + \cdots + X_n}(t) = \prod_{j=1}^{n} M_{X_j}(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Since $X_1, \ldots, X_n$ are independent, $e^{tX_1}, \ldots, e^{tX_n}$ are independent, for any $t \in \mathbb{R}$. So,

$$M_{X_1 + \cdots + X_n}(t) = \mathbf{E}e^{t(X_1 + \cdots + X_n)} = \mathbf{E} \prod_{j=1}^{n} e^{tX_j} = \prod_{j=1}^{n} \mathbf{E}e^{tX_j} = \prod_{j=1}^{n} M_{X_j}(t)$$

□

Example 4.18. Let X be a binomial distributed random variable with parameters n and $0 < p < 1$. That is, X can be written as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameter p. Then by Proposition 4.17, for any $t \in \mathbb{R}$,

$$M_X(t) = \prod_{j=1}^{n} M_{X_j}(t) = (M_{X_1}(t))^n = ((1-p)e^{0 \cdot t} + pe^t)^n = (1 - p + pe^t)^n.$$

In some cases, the moment generating function uniquely determines the random variable.

Theorem 4.19 (Lévy Continuity Theorem, Weak Form). *Let X, Y be random variables. Assume that $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Then X and Y have the same CDF.*

Exercise 4.20. Construct two random variables $X, Y\colon \Omega \to \mathbb{R}$ such that $X \neq Y$ but $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and such that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$.

Exercise 4.21. Unfortunately, there exist random variables X, Y such that $\mathbf{E}X^n = \mathbf{E}Y^n$ for all $n = 1, 2, 3, \ldots$, but such that X, Y do not have the same CDF. First, explain why this does not contradict the Lévy Continuity Theorem, Weak Form. Now, let $-1 < a < 1$, and define a density

$$f_a(x) := \begin{cases} \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2}} (1 + a\sin(2\pi \log x)) & \text{, if } x > 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Suppose X_a has density f_a. If $-1 < a, b < 1$, show that $\mathbf{E}X_a^n = \mathbf{E}X_b^n$ for all $n = 1, 2, 3, \ldots$. (Hint: write out the integrals, and make a change of variables $s = \log(x) - n$.)

From Exercise 4.13, the moment generating function of a random variable X contains all information about the moments of X. However, as mentioned in Remark 4.12, $M_X(t)$ may not exist for many values of t. So, studying the moment generating function may not be so helpful for certain random variables. Fortunately, the closely related characteristic function will always exist, and it also contains all information about the moments of X

5. Limit Theorems

We now start to build up the machinery that is used to prove the two "big theorems" of probability: the Law of Large Numbers, and the Central Limit Theorem. We begin with some useful inequalities.

5.1. Markov and Chebyshev Inequalities.

Markov's inequality says that a random variable with finite expected value cannot be too large very often.

Proposition 5.1 (The Markov Inequality). *Let X be a nonnegative random variable. Then*

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}X}{t}, \qquad \forall\, t > 0.$$

Proof. Let $t > 0$. Let Y be a random variable such that

$$Y = \begin{cases} t & \text{, if } X \geq t \\ 0 & \text{, if } X < t. \end{cases}$$

By definition of Y, we have $Y \leq X$. Therefore, $\mathbf{E}Y \leq \mathbf{E}X$ by Exercise 4.3. By the definition of Y, $\mathbf{E}Y = t\mathbf{P}(X \geq t)$. That is,

$$t\mathbf{P}(X \geq t) \leq \mathbf{E}(X).$$

□

Remark 5.2. A nearly identical proof shows that $\mathbf{P}(X > t) \leq \frac{\mathbf{E}X}{t}$, for all $t > 0$.

Markov's inequality is commonly applied in the following ways.

Corollary 5.3. *Let X be a random variable. Then*

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

More generally, if n is a positive integer, then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

Proof. The first assertion follows immediately by applying Proposition 5.1 to $|X|$. For the second assertion, we use the first assertion to get

$$\mathbf{P}(|X| \geq t) = \mathbf{P}(|X|^n \geq t^n) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

□

The second inequality of Corollary 5.3 is fairly useful, since if many moments of $|X|$ are bounded, then $\mathbf{P}(|X| \geq t)$ decays very rapidly.

Replacing X by $X - \mu$ and taking $n = 2$ in Corollary 5.3 gives:

Corollary 5.4 (**Chebyshev Inequality**). *Let X be a random variable with mean μ. Then*

$$\mathbf{P}(|X - \mu| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}, \qquad \forall\, t > 0.$$

Or, replacing t by $t\sqrt{\mathrm{var}(X)}$,

$$\mathbf{P}(|X - \mu| \geq t\sqrt{\mathrm{var}(X)}) \leq \frac{1}{t^2}, \qquad \forall\, t > 0.$$

Exercise 5.5. Let X be a standard Gaussian random variable. Let $t > 0$ and let n be a positive even integer. Show that

$$\mathbf{P}(X > t) \leq \frac{(n-1)(n-3)\cdots(3)(1)}{t^n}.$$

That is, the function $t \mapsto \mathbf{P}(X > t)$ decays faster than any monomial.

Exercise 5.6. Let X be a random variable. Let $t > 0$. Show that

$$\mathbf{P}(|X| > t) \leq \frac{\mathbf{E}X^4}{t^4}.$$

Proposition 5.7 (**Borel-Cantelli Lemma**). *Let $A_1, A_2, \ldots$ be events with $\sum_{n=1}^{\infty} \mathbf{P}(A_n) < \infty$. Let $B := \{\sum_{n=1}^{\infty} 1_{A_n} = \infty\}$, so that B is the event that infinitely many of the events $A_1, A_2, \ldots$ occur. Then $\mathbf{P}(B) = 0$.*

5.2. **Weak Law of Large Numbers.**

Definition 5.8. Let $X_1, X_2, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ are **identically distributed** if $X_1, X_2, \ldots$ all have the same CDF. That is, $\mathbf{P}(X_i \leq t) = \mathbf{P}(X_j \leq t)$ for all $t \in \mathbb{R}$ and for all positive integers i, j.

Remark 5.9. If $X_1, X_2, \ldots$ are identically distributed random variables, then $\mathbf{E}X_i = \mathbf{E}X_j$ for all positive integers i, j.

We know intuitively that, if the results of independent experiments are averaged, then the average will become close to the expected value of a single experiment. Indeed, one way to intuitively think about expected value is as the average of many repeated experiments. The Law of Large Numbers makes the previous statement rigorous. For now, we only prove a weak version of this statement, though a stronger version will be proven later.

Theorem 5.10 (**Weak Law of Large Numbers**). *Let $X_1, X_2, \ldots$ be independent identically distributed random variables. Assume that $\mu \in \mathbb{R}$ and $\mathbf{E}X_1 = \mu$. Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) = 0.$$

Proof. We make the additional assumption that $\mathrm{var}(X_1) < \infty$. Removing this assumption relies on things outside of this class. From Corollary 4.7,

$$\mathrm{var}\left(\frac{X_1 + \cdots + X_n}{n} \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{1}{n} \mathrm{var}(X_1).$$

So, Chebyshev's inequality implies that

$$\mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{1}{n} \varepsilon^{-2} \mathrm{var}(X_1).$$

Letting $n \to \infty$ concludes the proof. □

Example 5.11. Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with parameter $1/2$. Let $n := 10^4$, $\varepsilon := 10^{-2}$. Then

$$\mathbf{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \frac{1}{2} \right| \geq \frac{1}{100} \right) \leq 10^{-4} 10^4 (1/4) = \frac{1}{4}.$$

5.3. **Convergence in Probability.**

Definition 5.12. We say that a sequence of random variables $Y_1, Y_2, \ldots$ **converges in probability** to a random variable Y if: for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \varepsilon) = 0.$$

More formally, if Ω is the sample space, then $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(\omega \in \Omega\colon\ |Y_n(\omega) - Y(\omega)| > \varepsilon) = 0$.

Remark 5.13. So, the Weak Law of Large numbers says: if X_1, X_2 are independent identically distributed random variables with $\mu := \mathbf{E}X_1 \in \mathbb{R}$, then the random variables $\frac{X_1 + \cdots + X_n}{n}$ converge in probability to the constant μ.

Example 5.14. For any $n \geq 1$, let Y_n be a random variable such that $\mathbf{P}(Y_n = n^2) = 1/n$, and $\mathbf{P}(Y_n = 0) = 1 - 1/n$. Then $Y_1, Y_2, \ldots$ converges in probability to 0. For any $\varepsilon > 0$,

$$\mathbf{P}(|Y_n - 0| > \varepsilon) = \mathbf{P}(|Y_n| > \varepsilon) = \mathbf{P}(Y_n = n^2) = 1/n.$$

Therefore, $\lim_{n\to\infty} \mathbf{P}(|Y_n - 0| > \varepsilon) = 0$.

However, note that convergence in probability does not imply convergence in expected value, since $\lim_{n\to\infty} \mathbf{E}Y_n = \lim_{n\to\infty} n = \infty$, whereas the expected value of 0 is just 0.

Proposition 5.15 (**Uniqueness of the Limit**). *Suppose $Y_1, Y_2, \ldots$ converges in probability to Y. Also, suppose $Y_1, Y_2, \ldots$ converges in probability to Z. Then $\mathbf{P}(Z \neq Y) = 0$.*

Proof. From the triangle inequality, for any $n \geq 1$,

$$|Z - Y| = |Z - Y_n + Y_n - Y| \leq |Z - Y_n| + |Y - Y_n|.$$

So, for any $\varepsilon > 0$, if $|Z - Y| \geq \varepsilon$, then either $|Z - Y_n| \geq \varepsilon/2$ or $|Y - Y_n| \geq \varepsilon/2$. That is, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\{\omega \in \Omega\colon\ |Z(\omega) - Y(\omega)| \geq \varepsilon\}$$
$$\subseteq \{\omega \in \Omega\colon\ |Z(\omega) - Y_n(\omega)| \geq \varepsilon/2\} \cup \{\omega \in \Omega\colon\ |Y(\omega) - Y_n(\omega)| \geq \varepsilon/2\}.$$

Therefore, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\mathbf{P}(|Z - Y| \geq \varepsilon) \leq \mathbf{P}(|Z - Y_n| \geq \varepsilon/2) + \mathbf{P}(|Y - Y_n| \geq \varepsilon/2).$$

The left side does not depend on n. So, letting $n \to \infty$, we get $\mathbf{P}(|Z - Y| \geq \varepsilon) = 0$, for all $\varepsilon > 0$. Now,

$$\{Z \neq Y\} \subseteq \cup_{t=1}^{\infty} \{|Z - Y| \geq 1/t\}.$$

Therefore, $\mathbf{P}(Z \neq Y) \leq \sum_{t=1}^{\infty} \mathbf{P}(|Z - Y| \geq 1/t) = 0$. So, $\mathbf{P}(Z \neq Y) = 0$. □

Exercise 5.16. Let $X_1, X_2, \ldots$ be independent random variables, each with exponential distribution with parameter $\lambda = 1$. For any $n \geq 1$, let $Y_n := \max(X_1, \ldots, X_n)$. Let $0 < a < 1 < b$. Show that $\mathbf{P}(Y_n \leq a \log n) \to 0$ as $n \to \infty$, and $\mathbf{P}(Y_n \leq b \log n) \to 1$ as $n \to \infty$. Conclude that $Y_n / \log n$ converges to 1 in probability as $n \to \infty$.

Exercise 5.17. We say that random variables $X_1, X_2, \ldots$ converge to a random variable X in L_2 if

$$\lim_{n \to \infty} \mathbf{E}\,|X_n - X|^2 = 0.$$

Show that, if $X_1, X_2, \ldots$ converge to X in L_2, then $X_1, X_2, \ldots$ converges to X in probability.

Is the converse true? Prove your assertion.

Exercise 5.18. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables such that $\mathbf{E}\,|X_1| < \infty$ and $\operatorname{var}(X_1) < \infty$. For any $n \geq 1$, define

$$Y_n := \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

Show that $Y_1, Y_2, \ldots$ converges in probability. Express the limit in terms of $\mathbf{E}X_1$ and $\operatorname{var}(X_1)$.

5.4. **Central Limit Theorem.** The following is a stronger version of Theorem 4.19.

Theorem 5.19 (**Lévy Continuity Theorem**). *Let $X_1, X_2, \ldots$ be random variables and let X be a random variable. For any fixed $t \in \mathbb{R}$, assume that $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$. Then for any fixed $t \in \mathbb{R}$ such that $\mathbf{P}(X \leq t)$ is continuous, we have $\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \mathbf{P}(X \leq t)$.*

In particular, if X, Y are random variables with $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$, then X, Y are identically distributed.

We are finally able to prove the generalization of the De Moivre Laplace Theorem, Theorem 4.1, to arbitrary random variables.

Theorem 5.20 (**Central Limit Theorem**). *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables. Let Z be a standard Gaussian random variable. Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. Assume that $\mathbf{E}X_1 = \mu$ and $\operatorname{var}(X_1) = \sigma^2$. Then for any $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq t\right) = \int_{-\infty}^{t} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbf{P}(Z \leq t).$$

Remark 5.21. The random variable $\frac{X_1+\cdots+X_n-\mu n}{\sigma\sqrt{n}}$ has mean zero and variance 1, just like the standard Gaussian Z.

Proof. We make the additional assumption that all moment generating functions of the random variables exist for all $t \in \mathbb{R}$, and they are differentiable

For any $j \geq 1$, let $Y_j := (X_j - \mu)/\sigma$. Then $Y_1, Y_2, \ldots$ are independent and identically distributed, $\mathbf{E}Y_j = 0$ and $\mathbf{E}Y_j^2 = 1$, $\forall\ j \geq 1$. We will show that $\lim_{n\to\infty} \mathbf{P}(\frac{Y_1+\cdots+Y_n}{\sqrt{n}} \leq t) = \mathbf{P}(Z \leq t)$, $\forall\ t \in \mathbb{R}$. From Theorem 5.19 and Proposition 4.15, it suffices to show:

$$\lim_{n\to\infty} \mathbf{E}e^{t\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \mathbf{E}e^{tZ} = e^{t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

From Proposition 4.17,

$$\mathbf{E}e^{t\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \prod_{j=1}^{n} \mathbf{E}e^{tY_j/\sqrt{n}} = (\mathbf{E}e^{tY_1/\sqrt{n}})^n.$$

So, it suffices to show:

$$\lim_{n\to\infty} \log \mathbf{E}e^{t\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = t^2/2, \qquad \forall\, t \in \mathbb{R}.$$

Denote $c(t) := \log \mathbf{E}e^{tY_1} = \log M_{Y_1}(t)$ for all $t \in \mathbb{R}$. Recall that $M_{Y_1}(0) = 1$, $M'_{Y_1}(0) = \mathbf{E}Y_1 = 0$, and $M''_{Y_1}(0) = \mathbf{E}Y_1^2 = 1$. Therefore, $c(0) = 1$,

$$c'(0) = \frac{M'_{Y_1}(0)}{M_{Y_1}(0)} = 0,$$

$$c''(0) = \frac{M''_{Y_1}(0)M_{Y_1}(0) - [M'_{Y_1}(0)]^2}{[M_{Y_1}(0)]^2} = 1.$$

So, using LHôpital's rule, twice,

$$\lim_{n\to\infty} \log \mathbf{E}e^{t\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \lim_{n\to\infty} \log(\mathbf{E}e^{tY_1/\sqrt{n}})^n = \lim_{n\to\infty} n \log \mathbf{E}e^{tY_1/\sqrt{n}} = \lim_{n\to\infty} \frac{\log \mathbf{E}e^{tY_1/\sqrt{n}}}{1/n}$$
$$= \lim_{s\to 0} \frac{\log \mathbf{E}e^{tY_1 s}}{s^2} = \lim_{s\to 0} \frac{c(ts)}{s^2} = \lim_{s\to 0} \frac{tc'(ts)}{2s} = \lim_{s\to 0} \frac{t^2 c''(ts)}{2} = \frac{t^2}{2}.$$

□

Definition 5.22 (Convergence in Distribution). Let $X, X_1, X_2, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ **converge in distribution** to X if, for any $t \in \mathbb{R}$ such that the CDF of X is continuous at t,

$$\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \mathbf{P}(X \leq t).$$

So, the Central Limit Theorem, Theorem 5.20, says: if $X_1, X_2, \ldots$ are independent, identically distributed random variables with $\mu := \mathbf{E}X_1$ and $\sigma^2 := \mathrm{Var}(X_1)$ with $\sigma > 0$, then the random variables $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ converge in distribution to the standard Gaussian random variable. This fact is rather remarkable, since it holds no matter what distribution X_1 has! In this sense, the Gaussian random variable is "universal."

Exercise 5.23. This exercise demonstrates that geometry in high dimensions is different than geometry in low dimensions.

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Let $\|x\| := \sqrt{x_1^2 + \cdots + x_n^2}$. Let $\varepsilon > 0$. Show that for all sufficiently large n, "most" of the cube $[-1, 1]^n$ is contained in the annulus

$$A := \{x \in \mathbb{R}^n \colon (1-\varepsilon)\sqrt{n/3} \leq \|x\| \leq (1+\varepsilon)\sqrt{n/3}\}.$$

That is, if $X_1, \ldots, X_n$ are each independent and identically distributed in $[-1, 1]$, then for n sufficiently large

$$\mathbf{P}((X_1, \ldots, X_n) \in A) \geq 1 - \varepsilon.$$

(Hint: apply the weak law of large numbers to $X_1^2, \ldots, X_n^2$.)

Exercise 5.24 (**Confidence Intervals**)**.** Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a, b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, if we sample 625 bank members independently and uniformly at random (from a very large bank membership), then the number of these people with savings accounts lies in the interval $[a, b]$. (Hint: if Y is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

Exercise 5.25 (**Hypothesis Testing**)**.** Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let p be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the i^{th} spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\operatorname{var}(X_1)}$. If the null hypothesis is true, and if Y is a standard Gaussian random variable

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right| \geq 2\right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel n times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is called a type I error. In this test, we set the type I error percentage to be 5%. (The type I error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

Theorem 5.26 (**Strong Law of Large Numbers**)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume that $\mu = \mathbf{E}X_1$. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Proof. We prove the Theorem under the stronger assumption that $\mathbf{E}X_1^4 < \infty$. For any $j \geq 1$, let $Y_j := X_j - \mu$. We are required to show $\mathbf{P}\left(\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = 0\right) = 1$. Note that $Y_1, Y_2, \ldots$ are independent identically distributed random variables with $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^4 < \infty$. We compute

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = \sum_{1 \leq i,j,k,\ell \leq n} \mathbf{E}Y_i Y_j Y_k Y_\ell.$$

By independence, terms with $i \neq j = k = \ell$ vanish, since they become $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j^3 = 0$. Terms with i, j, k, ℓ distinct also vanish, since $\mathbf{E}Y_iY_jY_kY_\ell = \mathbf{E}Y_i\mathbf{E}Y_j\mathbf{E}Y_k\mathbf{E}Y_\ell = 0$. The remaining nonvanishing terms are $i = j = k = \ell$ and the six permutations of $i = j \neq k = \ell$. That is,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 = n\mathbf{E}Y_1^4 + 6[n(n-1)/2](\mathbf{E}Y_1^2)^2.$$

By Jensen's Inequality, Exercise 2.23,

$$\mathbf{E}(Y_1 + \cdots + Y_n)^4 \leq n\mathbf{E}Y_1^4 + 3n(n-1)\mathbf{E}Y_1^4 \leq 4n^2\mathbf{E}Y_1^4. \qquad (*)$$

By Markov's Inequality, Proposition 5.1, for any $t > 0$,

$$\mathbf{P}\Big(\Big|\frac{Y_1 + \cdots + Y_n}{n}\Big| > t\Big) \leq \frac{\mathbf{E}(Y_1 + \cdots + Y_n)^4}{t^4n^4} \overset{(*)}{\leq} \frac{4\mathbf{E}Y_1^4}{t^4n^2}.$$

So $\sum_{n=1}^{\infty} \mathbf{P}(\big|\frac{Y_1+\cdots+Y_n}{n}\big| > t) < \infty$ and by Borel-Cantelli, Proposition 5.7, $\forall\ t > 0$,

$$\mathbf{P}\Big(\Big|\frac{Y_1 + \cdots + Y_n}{n}\Big| > t \text{ for infinitely many } n \geq 1\Big) = 0.$$

Since this holds for any $t > 0$, we conclude that $\frac{Y_1+\cdots+Y_n}{n}$ converges almost surely to 0. □

6. Estimation of Parameters

Exercise 6.1. Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell the mean or standard deviation.

5.1715, 3.2925, 5.2172, 6.1302, 4.9889, 5.5347, 5.2269, 4.1966, 4.7939, 3.7127

5.3884, 3.3529, 3.4311, 3.6905, 1.5557, 5.9384, 4.8252, 3.7451, 5.8703, 2.7885

To the best of your ability, determine what the mean and standard deviation are of this random variable. (This question is a bit open-ended, so there could be more than one correct way of justifying your answer.)

A basic problem in statistics is to fit data to an unknown probability distribution. As in Exercise 6.1, we might have a list of numbers, and we known these numbers follow some Gaussian distribution, but we might not know the mean and variance of this Gaussian. We then want to infer the mean and variance from the data. In this example, there are two unknown parameters. In general, we might want to estimate any number of unknown parameters.

Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. We can regard $\{f_\theta \colon \theta \in \Theta\}$ as either a family of probability density functions, or a family of probability mass functions. If Y is a statistic that is used to estimate the parameter θ that fits the data at hand, we then refer to Y as a **point estimator** or **estimator**.

Example 6.2. In Exercise 6.1 we have a random sample $X_1, \ldots, X_{20}$ from a Gaussian distribution with unknown mean and variance. We denote the unknown Gaussians as

$$\{f_\theta \colon \theta \in \Theta\} = \{f_{\mu,\sigma}(x) \colon (\mu, \sigma) \in \mathbb{R}^2,\ \mu \in \mathbb{R},\ \sigma > 0\} = \Big\{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R},\ \sigma > 0\Big\}.$$

One estimator for the unknown mean μ is the sample mean

$$\frac{X_1 + \cdots + X_{20}}{20}.$$

A "less good" estimator for the unknown mean μ could be $X_1 + X_2$ or $(X_1 + X_3)/2$.

As previously discussed, an estimator for the unknown variance σ^2

$$\frac{1}{19}\sum_{i=1}^{20}(X_i - \overline{X})^2.$$

And an estimator for the unknown parameter σ itself is

$$S := \sqrt{\frac{1}{19}\sum_{i=1}^{20}(X_i - \overline{X})^2}.$$

As we see from this example, there are many ways of defining estimators for various unknown parameters. One focus of this course will be criteria for determining if an estimator is "good" or not.

There are many different ways to create estimators. A priori, it might not be clear which estimator is the best. One desirable property of an estimator is that it is unbiased.

Definition 6.3. Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}^k$ and let $Y := t(X_1, \ldots, X_n)$ be an estimator for $g(\theta)$. Let $g \colon \Theta \to \mathbb{R}^k$. We say that Y is **unbiased** for $g(\theta)$ if

$$\mathbf{E}_\theta Y = g(\theta), \qquad \forall\, \theta \in \Theta.$$

For example, we saw in Exercise 6.4 that the sample mean and sample variance are unbiased estimates of the mean and variance, respectively.

Exercise 6.4. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n. Assume that $\mu := \mathbf{E}X_1 \in \mathbb{R}$ and $\sigma := \sqrt{\operatorname{var}(X_1)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

- $\operatorname{Var}(\overline{X}) = \sigma^2/n$.
- $\mathbf{E}S^2 = \sigma^2$.

6.1. Method of Moments.

Definition 6.5 (Consistency). Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of distributions. Let $Y_1, Y_2, \ldots$ be a sequence of estimators of $g(\theta)$ where $g \colon \Theta \to \mathbb{R}^k$. We say that $Y_1, Y_2, \ldots$ is **consistent** for $g(\theta)$ if, for any $\theta \in \Theta$, $Y_1, Y_2, \ldots$ converges in probability to the constant value $g(\theta)$, with respect to the probability distribution f_θ.

Typically, we will take Y_n to be a function of a random sample of size n, for all $n \geq 1$.

Example 6.6. Let $X_1, \ldots, X_n$ be a random sample of size n with distribution f_θ. The Weak Law of Large Numbers, Theorem 5.10, says that the sample mean is consistent when $\mathbf{E}_\theta |X_1| < \infty$ for all $\theta \in \Theta$. More generally, if $j \geq 1$ is a positive integer such that $\mathbf{E}_\theta |X_1|^j < \infty$ for all $\theta \in \Theta$, then the j^{th} sample moment

$$M_j = M_j(\theta) := \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

is also consistent (as $n \to \infty$), i.e. as $n \to \infty$, M_j converges in probability to the j^{th} moment

$$\mu_j(\theta) := \mathbf{E}X_1^j.$$

Note also that if $h\colon \mathbb{R}^k \to \mathbb{R}^k$ is continuous, and if $Y_1, Y_2, \ldots$ is consistent for $g(\theta)$, then $h(Y_1), h(Y_2), \ldots$ is consistent for $h(g(\theta))$ by Exercise 6.7.

Exercise 6.7. Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}$ be random variables that converge in probability to $X\colon \Omega \to \mathbb{R}$. Let $f\colon \mathbb{R} \to \mathbb{R}$ be continuous. Then $f(X_1), f(X_2), \ldots$ converges in probability to $f(X)$.

Definition 6.8 (Method of Moments). Let $g\colon \Theta \to \mathbb{R}^k$. Suppose we want to estimate $g(\theta)$ for any $\theta \in \Theta$. Suppose there exists $h\colon \mathbb{R}^j \to \mathbb{R}^k$ such that

$$g(\theta) = h(\mu_1, \ldots, \mu_j).$$

Then the estimator

$$h(M_1, \ldots, M_j)$$

is a **method of moments** estimator for $g(\theta)$.

Example 6.9. To estimate the mean μ, we can use $\Theta = \mathbb{R} = \{\mu_1 \in \mathbb{R}\}$, $j = 1$ and $h(\mu_1) = \mu_1$, so that a method of moments estimator of μ_1 is the sample mean M_1.

To estimate the standard deviation, we can use $\Theta = \mathbb{R} \times (0, \infty) = \{(\mu_1, \mu_2)\colon \mu_1 \in \mathbb{R}, \mu_2 > 0\}$, $j = 2$ and $h(\mu_1, \mu_2) = \sqrt{\mu_2 - \mu_1^2}$, so that a method of moments estimator of the standard deviation is $\sqrt{M_2 - M_1^2}$.

This estimation approach is good in that it uses essentially no assumptions about model parameters. Perhaps for this reason, the method of moments is one of the oldest methods of point estimation, originating in the late 1800s. However, when information about model parameters is available, often the method of moments does not work well (despite being consistent). In the following example, we demonstrate an estimator with much smaller variance than the method of moments estimator.

Example 6.10. Suppose $X_1, \ldots, X_n$ is a random sample of size n from the uniform distribution on the interval $[0, \theta]$ and $\theta > 0$ is unknown. Since $\mathbf{E}_\theta X_1 = \theta/2$, a method of moment estimator for θ is $2M_1 = \frac{2}{n}\sum_{i=1}^n X_i$. This estimator is unbiased and consistent (by Example 6.6), but its variance is $\frac{1}{3n}\theta^2$. It turns out the estimator $(1 + 1/n)X_{(n)}$ is unbiased and consistent for θ with a smaller variance. From Definition 6.11 we have

$$\begin{aligned}
\mathbf{E}(1 + 1/n)X_{(n)} &= (1 + 1/n)\int_0^\theta \mathbf{P}(X_{(n)} > t)dt = (1 + 1/n)\int_0^\theta [1 - \mathbf{P}(X_{(n)} < t)]dt \\
&= (1 + 1/n)\int_0^\theta [1 - \mathbf{P}(X_{(n)} < t)]dt = (1 + 1/n)(\theta - \int_0^\theta \mathbf{P}(X_1 < t)^n dt) \\
&= (1 + 1/n)(\theta - \int_0^\theta (t/\theta)^n dt) = (1 + 1/n)(\theta - \theta^{-n}\theta^{n+1}/(n+1)) \\
&= (1 + 1/n)(\theta - \theta/(n+1)) = \theta\frac{n+1}{n}\frac{n}{n+1} = \theta.
\end{aligned}$$

From Definition 6.11, $\operatorname{var}((1+1/n)X_{(n)})$ is equal to

$$\frac{(n+1)^2}{n^2}\mathbf{E}X_{(n)}^2 - \theta^2 = \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} > t)dt - \theta^2$$
$$= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\int_0^\theta 2t\mathbf{P}(X_{(n)} < t)dt$$
$$= \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^{-n}\int_0^\theta 2tt^n dt = \theta^2\Big(\frac{(n+1)^2}{n^2} - 1\Big) - \frac{(n+1)^2}{n^2}\theta^2\frac{2}{n+2}$$
$$= \frac{\theta^2}{n^2(n+2)}\Big((n+1)^2(n+2) - n^2(n+2) - 2(n+1)^2\Big)$$
$$= \frac{\theta^2}{n^2(n+2)}\Big([(n+1)^2 - n^2](n+2) - 2(n+1)^2\Big)$$
$$= \frac{\theta^2}{n^2(n+2)}\Big([2n+1](n+2) - 2(n+1)^2\Big) = \frac{\theta^2}{n^2(n+2)}(5n - 4n + 2 - 2) = \frac{\theta^2}{n(n+2)}.$$

In fact, $(1+1/n)X_{(n)}$ is the uniform minimum variance unbiased estimator for θ (and we call this estimator UMVU), though we will not prove it.

Definition 6.11 (**Expected Value**)**.** Let Ω be a sample space, let $\mathbf{P}$ be a probability law on Ω. Let X be a random variable on Ω. Assume that $X\colon \Omega \to [0,\infty)$. We define the **expected value** of X, denoted $\mathbf{E}(X)$, by

$$\mathbf{E}(X) = \int_0^\infty \mathbf{P}(X > t)dt.$$

More generally, if $g\colon [0,\infty) \to [0,\infty)$ is a differentiable function such that g' is continuous and $g(0) = 0$, we define

$$\mathbf{E}g(X) = \int_0^\infty g'(t)\mathbf{P}(X > t)dt.$$

In particular, taking $g(t) = t^n$ for any positive integer n, for any $t \geq 0$, we have

$$\mathbf{E}X^n = \int_0^\infty nt^{n-1}\mathbf{P}(X > t)dt.$$

For a general random variable X, if $\mathbf{E}\max(X,0) < \infty$ and if $\mathbf{E}\max(-X,0) < \infty$, we then define $\mathbf{E}(X) = \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0)$. Otherwise, we say that $\mathbf{E}(X)$ is undefined.

Example 6.12. Suppose we have a binomial random variable with unknown parameters n, p. We want to find method of moments estimators for n and for p. It is known that $\mathbf{E}X_1 = np$ and $\mathbf{E}X_1^2 = np(1-p) + n^2p^2$. So, we solve for n, p in the system of equations

$$\mu_1 = np, \qquad \mu_2 = np(1-p) + n^2p^2,$$

to get an estimator for n:

$$N := \frac{M_1^2}{M_1 - (M_2 - M_1^2)}, \qquad \text{since} \qquad n = \frac{\mu_1^2}{\mu_1 - (\mu_2 - \mu_1^2)},$$

and an estimator for p:

$$P := \frac{M_1}{N}, \qquad \text{since} \qquad p = \frac{\mu_1}{n}.$$

(To solve the system, note that the second equation says $\mu_2 = (1-p)\mu_1 + \mu_1^2 = (1-\mu_1/n)\mu_1 + \mu_1^2$, and then solve for n.)

The Central Limit Theorem implies that the combination of a large number of independent identically distributed random actions results in a Gaussian distribution. For this reason, one can often (but not always) assume that sampling from a large population is sampling from the normal distribution with unknown mean and variance. Since this Gaussian assumption is so common, we discuss properties of sampling from the normal in this section.

Proposition 6.13. *Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample from the Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation.*

- *$\overline{X}$ and S are independent random variables.*
- *$\overline{X}$ is a Gaussian random variable with mean μ and variance σ^2/n.*
- *$(n-1)S^2/\sigma^2$ is a chi-squared distributed random variable with $n-1$ degrees of freedom.*

6.2. **Evaluating Estimators.** Even if an estimator is unbiased, its distribution of values might be quite far from $g(\theta)$. Recall that we made a similar observation that the Law of Large Numbers does not give any information about the Central Limit Theorem. It is desirable to examine the distribution of values of the estimator. The most common way to check the quality of an estimator in this sense is to examine the mean-squared error, or squared L_2 norm, of the estimator minus $g(\theta)$:

$$\mathbf{E}_\theta(Y - g(\theta))^2.$$

If the estimator is unbiased, this quantity is equal to the variance of Y.

Definition 6.14 (**UMVU**)**.** Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Let $g \colon \Theta \to \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X_1, \ldots, X_n)$ be an unbiased estimator for $g(\theta)$. We say that Y is **uniformly minimum variance unbiased (UMVU)** if, for any other unbiased estimator Z for $g(\theta)$, we have

$$\mathrm{Var}_\theta(Y) \leq \mathrm{Var}_\theta(Z), \qquad \forall\, \theta \in \Theta.$$

Remark 6.15. Unfortunately the UMVU might not exist. Suppose we want a UMVU for a binomial random variable X with known parameter n and unknown parameter $0 < \theta < 1$, and we want an estimator for $\theta/(1-\theta)$. In fact, no unbiased estimate exists for this function, since $\mathbf{E}_\theta t(X) = \sum_{j=0}^n \binom{n}{j} t(j)\theta^j(1-\theta)^{n-j}$ and this is a polynomial in θ, i.e. only polynomials in θ of degree at most n can possible have unbiased estimates. And $\theta/(1-\theta)$ is not a polynomial in θ.

6.3. **Efficiency of an Estimator.** Another desirable property of an estimator is high efficiency. That is, the estimator is good with a small number of samples. One way to quantify "good" in the previous sentence is to define a notion of information and to try to maximize the information content of the estimator.

Definition 6.16 (**Fisher Information**)**.** Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of multivariable probability densities or probability mass functions. Assume $\Theta \subseteq \mathbb{R}$. Let X be a random vector with distribution f_θ. Define the **Fisher information** of the family to be

$$I(\theta) = I_X(\theta) := \mathbf{E}_\theta\left(\frac{d}{d\theta} \log f_\theta(X)\right)^2, \qquad \forall\, \theta \in \Theta,$$

if this quantity exists and is finite.

In order for the Fisher information to be well defined, the set $\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ should not depend on θ, otherwise the derivative $\frac{d}{d\theta} \log f_\theta(X)$ might not be well-defined.

If $\{f_\theta \colon \theta \in \Theta\}$ are n-dimensional probability densities, note that

$$\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) = \int_{\mathbb{R}^n} \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{d}{d\theta} f_\theta(x) dx = \frac{d}{d\theta} \int_{\mathbb{R}^n} f_\theta(x) dx = \frac{d}{d\theta}(1) = 0.$$

Similarly, if $\{f_\theta \colon \theta \in \Theta\}$ are multivariable probability mass functions, $\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) = 0$. So, we could equivalently define

$$I(\theta) = \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X)\Big), \qquad \forall\, \theta \in \Theta.$$

(Here we assume we can differentiate under the integral sign.) We also have another equivalent definition:

$$\begin{aligned} \mathbf{E}_\theta \frac{d^2}{d\theta^2} \log f_\theta(X) &= \int_{\mathbb{R}^n} \frac{d}{d\theta} \frac{\frac{d}{d\theta} f_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int_{\mathbb{R}^n} \frac{f_\theta(x) \frac{d^2}{d\theta^2} f_\theta(x) - \Big(\frac{d}{d\theta} f_\theta(x)\Big)^2}{[f_\theta(x)]^2} f_\theta(x) dx \\ &= \int_{\mathbb{R}^n} \frac{d^2}{d\theta^2} f_\theta(x) - \Big(\frac{d}{d\theta} \log f_\theta(x)\Big)^2 f_\theta(x) dx = 0 - I_X(\theta) = -I_X(\theta). \end{aligned}$$

The Fisher information expresses the amount of "information" a random variable has.

Example 6.17. Let $\sigma > 0$ and let $f_\theta(x) := \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\theta)^2/[2\sigma^2]}$ for all $\theta \in \Theta$, $x \in \mathbb{R}$. We have

$$I(\theta) = \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \frac{-(X-\theta)^2}{2\sigma^2}\Big) = \frac{1}{\sigma^4} \mathrm{Var}_\theta(X - \theta) = \frac{1}{\sigma^2}.$$

For the Gaussian case, we interpret "more information" as σ small, since then the variance is small, so more "information" is conveyed by a single sample than when σ is large. The Fisher information also agrees with our intuitive notion of information since the information of a joint distribution of independent random variables is equal to the sum of the separate informations.

Proposition 6.18. *Let X be a random variable with distribution from $\{f_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Assume that X and Y are independent. Then*

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta), \qquad \forall\, \theta \in \Theta.$$

Proof. Since X and Y are independent, (X, Y) has distribution from the multivariate density $f_\theta(X) g_\theta(Y)$. Also, $\frac{d}{d\theta} \log f_\theta(X)$ and $\frac{d}{d\theta} \log g_\theta(Y)$ are independent for any $\theta \in \Theta$, so

$$\begin{aligned} I_{(X,Y)}(\theta) &= \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log[f_\theta(X) g_\theta(Y)]\Big) = \mathrm{Var}_\theta\Big(\frac{d}{d\theta}[\log f_\theta(X) + \log g_\theta(Y)]\Big) \\ &= \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log f_\theta(X)\Big) + \mathrm{Var}_\theta\Big(\frac{d}{d\theta} \log g_\theta(X)\Big) = I_X(\theta) + I_Y(\theta). \end{aligned}$$

□

Exercise 6.19. Let X be a random variable with distribution from $\{f_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Let Y be a random variable with distribution from $\{g_\theta \colon \theta \in \Theta\}$ (densities or mass functions). Show that

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_{Y|X=x}(\theta), \qquad \forall\, \theta \in \Theta,\, x \in \mathbb{R}.$$

(If X, Y are continuous random variables, recall that $Y|X$ has density $f_{X,Y}(x,y)/f_X(x)$ for any fixed x. And if X, Y are discrete random variables, recall that $Y|X$ has mass function $\mathbf{P}(X = x, Y = y)/\mathbf{P}(Y = y)$.)

Our primary interest in information is the following inequality. Theorem 6.20 gives a lower bound on the variance of unbiased estimators of θ.

Theorem 6.20 (Cramér-Rao/ Information Inequality). *Let $X \colon \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta \colon \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be statistic. For any $\theta \in \Theta$ let $g(\theta) := \mathbf{E}_\theta Y$. Then*

$$\mathrm{Var}_\theta(Y) \geq \frac{|g'(\theta)|^2}{I_X(\theta)}, \qquad \forall\, \theta \in \Theta.$$

In particular, if Y is unbiased for θ,

$$\mathrm{Var}_\theta(Y) \geq \frac{1}{I_X(\theta)}, \qquad \forall\, \theta \in \Theta.$$

Equality occurs for some $\theta \in \Theta$ only when $\frac{d}{d\theta} \log f_\theta(X)$ and $Y - \mathbf{E}_\theta Y$ are multiples of each other.

(Here we assume we can differentiate under the integral sign. Also, we assume that $\{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ does not depend on θ, and for a.e. $x \in \mathbb{R}^n$, $(d/d\theta) f_\theta(x))$ exists and is finite.)

Remark 6.21. In the case that $X_1, \ldots, X_n$ are i.i.d. real-valued random variables and $X = (X_1, \ldots, X_n)$, Proposition 6.18 says that $I_X(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = n I_{X_1}(\theta)$. And if Y is unbiased for θ, Theorem 6.20 says

$$\mathrm{Var}_\theta(Y) \geq \frac{1}{n I_{X_1}(\theta)}, \qquad \forall\, \theta \in \Theta.$$

Proof. For any $\theta \in \Theta$ let $g(\theta) := \mathbf{E}_\theta Y$. We assume that X is continuous, the discrete case being similar. Using $\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) = 0$ and Remark 2.26,

$$|g'(\theta)| = \left| \frac{d}{d\theta} \int_{\mathbb{R}^n} f_\theta(x) t(x) dx \right| = \left| \int_{\mathbb{R}^n} \frac{d}{d\theta} \log f_\theta(x) t(x) f_\theta(x) dx \right| = \left| \mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X) t(X) \right|$$
$$= \left| \mathrm{Cov}_\theta\left(\frac{d}{d\theta} \log f_\theta(X), t(X)\right) \right| \leq \sqrt{\mathrm{Var}_\theta\left(\frac{d}{d\theta} \log f_\theta(X)\right) \mathrm{Var}_\theta(t(X))} = \sqrt{I_X(\theta) \mathrm{Var}_\theta(t(X))}.$$

The equality case follows from Remark 2.26 and the known equality case of the Cauchy-Schwarz Inequality (see Theorem 2.27). □

For a one-parameter family of distributions, the equality case of Theorem 6.20 allows us to find a UMVU for θ. To find such an estimator, we look for affine functions of $\frac{d}{d\theta} \log f_\theta(X)$.

Example 6.22. Suppose $f_\theta(x) := \theta x^{\theta-1} 1_{0<x<1}$ for all $x \in \mathbb{R}, \theta > 0$. (This is a beta distribution with $\beta = 1$.) We have

$$\frac{d}{d\theta} \log f_\theta(x) = \frac{1}{\theta} + \log x, \qquad \forall\, 0 < x < 1.$$

A vector $X = (X_1, \ldots, X_n)$ of n independent samples from f_θ is distributed according to the product $\prod_{i=1}^n f_\theta(x_i)$, so that

$$\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(x_i) = \sum_{i=1}^n \Big(\frac{1}{\theta} + \log x_i\Big) = n\Big(\frac{1}{\theta} + \frac{1}{n} \log \prod_{i=1}^n x_i\Big), \qquad \forall\, 0 < x_i < 1,\ 1 \le i \le n.$$

By Theorem 6.20, any function of $\frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i)$ (plus a constant) is UMVU for its expectation. So, for example,

$$Y := -\frac{1}{n} \log \prod_{i=1}^n X_i$$

is UMVU of its expectation, which is $\frac{1}{\theta}$ since $\mathbf{E}_\theta \frac{d}{d\theta} \log \prod_{i=1}^n f_\theta(X_i) = 0$.

Theorem 6.20 suggests the following quantity represents the efficiency of an estimator.

Definition 6.23 (**Efficiency**)**.** Let $X \colon \Omega \to \mathbb{R}^n$ be a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta \colon \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$. Let $t \colon \mathbb{R}^n \to \mathbb{R}$ and let $Y := t(X)$ be statistic. Define the **efficiency** of Y to be

$$\frac{1}{I_X(\theta) \mathrm{Var}_\theta(Y)}, \qquad \forall\, \theta \in \Theta,$$

if this quantity exists and is finite. If Z is another statistic, we define the **relative efficiency** of Y to Z to be

$$\frac{I_X(\theta) \mathrm{Var}_\theta(Z)}{I_X(\theta) \mathrm{Var}_\theta(Y)} = \frac{\mathrm{Var}_\theta(Z)}{\mathrm{Var}_\theta(Y)}, \qquad \forall\, \theta \in \Theta.$$

6.4. **Maximum Likelihood Estimator.** Let $X_1, \ldots, X_n$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. So, we denote the joint distribution of $X_1, \ldots, X_n$ as

$$\prod_{i=1}^n f_\theta(x_i), \qquad \forall\, 1 \le i \le n.$$

If we have data $x \in \mathbb{R}^n$, recall that we defined the function $\ell \colon \Theta \to [0, \infty)$

$$\ell(\theta) := \prod_{i=1}^n f_\theta(x_i)$$

and called it the **likelihood function**.

Definition 6.24 (**Maximum Likelihood Estimator**)**.** The **maximum likelihood estimator** (MLE) Y is the estimator maximizing the likelihood function. That is, $Y := t(X)$, $t \colon \mathbb{R}^n \to \Theta$ and $t(x_1, \ldots, x_n)$ is defined to be any value of $\theta \in \Theta$ that maximizes the function

$$\prod_{i=1}^n f_\theta(x_i),$$

if this value of θ exists. A priori, the θ maximizing $\ell(\theta)$ might not exist, and it might not be unique

Remark 6.25. Maximizing the likelihood $\ell(\theta)$ is equivalent to maximizing $\log \ell(\theta)$, since log is monotone increasing.

It is relatively easy to construct examples where the MLE is not unique.

Example 6.26. Let $f_\theta(x_1) := 1_{[\theta,\theta+1]}(x_1)$ for all $x_1, \theta \in \mathbb{R}$. Then, for all $x_1, \ldots, x_n, \theta \in \mathbb{R}$, we have

$$\prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n 1_{[\theta,\theta+1]}(x_i) = \prod_{i=1}^n 1_{x_i \in [\theta,\theta+1]}.$$

So, if $x_1 = \cdots = x_n = 0$, we have

$$\prod_{i=1}^n f_\theta(x_i) = 1_{0 \in [\theta,\theta+1]} = 1_{\theta \in [-1,0]}.$$

That is, any value of $\theta \in [-1, 0]$ is a maximum of the likelihood function, i.e. there are infinitely many maxima of the likelihood function. This is certainly not desirable.

If the likelihood function is continuous and Θ is compact, then at least one maximum of the likelihood function must exist.

A common assumption of a probability density function is that it is logarithmically concave. We will describe how this condition guarantees the uniqueness of the MLE. For a proof of consistency of the MLE under certain assumptions, see the Keener book, Theorem 9.11.

Recall that $\phi\colon \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ with $x \neq y$ and for any $t \in (0, 1)$,

$$\phi(tx + (1-t)y) \leq t\phi(x) + (1-t)\phi(y).$$

And $\phi\colon \mathbb{R}^n \to \mathbb{R}$ is strictly convex if this inequality is always a strict inequality. We also say ϕ is concave if $-\log \phi$ is convex, and ϕ is strictly concave if $-\log \phi$ is strictly convex.

Definition 6.27 (**Log-Concave**)**.** We say that $\phi\colon \mathbb{R}^n \to [0, \infty)$ is **logarithmically concave** or **log concave** if $\log \phi$ is concave, i.e. $-\log \phi$ is convex.

For example, the function $\phi(x) = e^{-x^2}$, $x \in \mathbb{R}$, is log concave, since $\log \phi$ is concave. If we allow ϕ to take infinite values, then $1_{[-1,0]}$ is log-concave, so Example 6.26 shows that log-concavity still does not guarantee uniqueness of the maximum of the likelihood function. However, strict log-concavity does guarantee uniqueness.

Proposition 6.28. *Let $f_\theta\colon \mathbb{R} \to [0, \infty)$ be a family of probability density functions, where $\theta \in \Theta \subseteq \mathbb{R}^k$. Fix $x_1, \ldots, x_n \in \mathbb{R}$. Assume that the function*

$$\theta \mapsto f_\theta(x_i)$$

is strictly log-concave, for every $1 \leq i \leq n$. Fix $x_1, \ldots, x_n \in \mathbb{R}$. Then the likelihood function

$$\theta \mapsto \prod_{i=1}^n f_\theta(x_i)$$

has at most one maximum value.

Proof. The function $\theta \mapsto \log f_\theta(x_i)$ is strictly concave for all $1 \leq i \leq n$, so the function

$$\theta \mapsto \sum_{i=1}^n \log f_\theta(x_i) = \log \prod_{i=1}^n f_\theta(x_i)$$

is strictly concave by Exercise 6.31. From Exercise 6.29, this function has at most one global maximum. □

Exercise 6.29. Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. Let $x \in \mathbb{R}^n$ be a local minimum of f. Show that x is in fact a global minimum of f.

Show also that if f is strictly convex, then there is at most one global minimum of f.

Now suppose additionally that f is a C^1 function (all derivatives of f exist and are continuous), and $x \in \mathbb{R}^n$ satisfies $\nabla f(x) = 0$. Show that x is a global minimum of f.

Exercise 6.30. Let A be a real $m \times n$ matrix. Let $x \in \mathbb{R}^n$ and let $b \in \mathbb{R}^m$. Show that the function $f\colon \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \frac{1}{2}\|Ax - b\|^2$ is convex. Moreover, show that

$$\nabla f(x) = A^T(Ax - b), \qquad D^2 f(x) = A^T A.$$

(Here $D^2 f$ denotes the matrix of second derivatives of f.)

So, if $\nabla f(x) = 0$, i.e. if $A^T Ax = A^T b$, then x is the global minimum of f. And if A has full rank, then $A^T A$ is invertible, so that $x = (A^T A)^{-1} A^T b$ is the global minimum of f.

Exercise 6.31. Let $f_1, \dots, f_n\colon \mathbb{R} \to \mathbb{R}$ be n strictly convex functions on $\mathbb{R}$. Define $g\colon \mathbb{R}^n \to \mathbb{R}$ by

$$g(x_1, \dots, x_n) := \sum_{i=1}^n f(x_i), \qquad \forall\, (x_1, \dots, x_n) \in \mathbb{R}^n.$$

Show that $g\colon \mathbb{R}^n \to \mathbb{R}$ is strictly convex.

Exercise 6.32. Let $f\colon \mathbb{R}^2 \to \mathbb{R}$ be a C^1 function (all derivatives of f exist and are continuous). Suppose there exists $z \in \mathbb{R}$ such that, for any $x_1 \in \mathbb{R}$, we have

$$f(x_1, z) < f(x_1, x_2), \qquad \forall\, x_2 \neq z.$$

Assume also that the function

$$x_1 \mapsto f(x_1, z)$$

is strictly convex. Show that f has at most one global minimum.

Example 6.33. Consider a random sample from a Gaussian distribution with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$, so that $\theta = (\mu, \sigma)$. The value of θ maximizing

$$\log \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x_i - \mu)^2/[2\sigma^2]) = \sum_{i=1}^n -\log\sigma - \frac{1}{2}\log(2\pi) - \frac{(x_i - \mu)^2}{2\sigma^2}$$

can be found by differentiating in the two parameters. We have

$$\frac{\partial}{\partial \mu} \log \ell(\theta) = \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2}, \qquad \frac{\partial}{\partial \sigma} \log \ell(\theta) = \sum_{i=1}^n -\sigma^{-1} + \sigma^{-3}(x_i - \mu)^2,$$

Setting both terms equal to zero, we get

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2.$$

This is the unique critical point of the function $\ell(\theta)$. It remains to show that this critical point is the global maximum of $\ell(\theta)$. It follows from Exercise 2.8 that, if $z \neq \frac{1}{n}\sum_{i=1}^n x_i$, then

$$\sum_{i=1}^n \Big(x_i - \frac{1}{n}\sum_{i=1}^n x_i\Big)^2 < \frac{1}{n}\sum_{i=1}^n (x_i - z)^2.$$

Therefore, for any such $z \in \mathbb{R}$

$$\log \ell(\frac{1}{n}\sum_{i=1}^n x_i, \sigma) > \log \ell(z, \sigma).$$

So, we need only show that $\log \ell(\frac{1}{n}\sum_{i=1}^n x_i, \sigma)$ is maximized when $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2}$. Since

$$\frac{\partial}{\partial \sigma} \log \ell(\theta) = \sigma^{-3} \sum_{i=1}^n -\sigma^2 + (x_i - \mu)^2,$$

the function $\sigma \mapsto \log \ell(\mu, \sigma)$ is increasing, and then decreasing, so that the global maximum occurs at the unique critical point.

It is known that the sample mean M_1 is UMVU for the mean. Let

$$Y = Y_n = Y_n(X_1, \ldots, X_n) := \frac{1}{n}\sum_{j=1}^n \Big(X_j - \frac{1}{n}\sum_{i=1}^n X_i\Big)^2.$$

We also know from Proposition 6.13 that Y is asymptotically unbiased for σ^2, i.e.

$$\lim_{n\to\infty} \frac{\mathbf{E}Y}{\sigma^2} = \lim_{n\to\infty} \frac{n-1}{n} = 1.$$

We will show that Y has asymptotically optimal variance. If we fix $\mu \in \mathbb{R}$ and look at the information of the n-dimensional Gaussian X, we get by modifying Example 6.17 and using Proposition 6.18

$$\begin{aligned} I_X(\sigma) = nI_{X_1}(\sigma) &= n\mathrm{Var}_\sigma\Big(\frac{d}{d\sigma}\frac{-(X_1-\mu)^2}{2\sigma^2}\Big) = n\sigma^{-6}\mathrm{Var}_\sigma[(X_1-\mu)^2] \\ &= n\sigma^{-6}\mathbf{E}_\sigma((X_1-\mu)^4 - \sigma^4) = 2n\sigma^{-2}. \end{aligned}$$

By the Cramér-Rao Inequality, Theorem 6.20, with $g(\sigma) = \mathbf{E}_\sigma(Y) = \sigma^2(n-1)/n$ (using Proposition 6.13), the variance of any unbiased estimator Z of $\sigma^2(n-1)/n$ satisfies

$$\mathrm{Var}_\sigma(Z) \geq \frac{|g'(\sigma)|^2}{I_X(\sigma)} = \frac{4\sigma^2(n-1)^2}{n^2 2n\sigma^{-2}} = \frac{2\sigma^4(n-1)^2}{n^3}.$$

And by Proposition 6.13,

$$\mathrm{Var}_\sigma(Y) = \mathrm{Var}_\sigma\Big[\frac{\sigma^2}{n}\frac{1}{\sigma^2}\sum_{j=1}^n \Big(X_j - \frac{1}{n}\sum_{i=1}^n X_i\Big)^2\Big] = \frac{\sigma^4}{n^2}2(n-1) = \frac{2\sigma^4(n-1)}{n^2}.$$

In summary,

$$\lim_{n\to\infty} \frac{\mathbf{E}Y}{\sigma^2} = 1, \qquad \lim_{n\to\infty} \frac{\mathrm{Var}_\sigma(Y)}{|g'(\sigma)|^2 / I_X(\sigma)} = 1.$$

That is, the estimator Y is asymptotically unbiased (as $n \to \infty$) and it asymptotically achieves the optimal variance bound in the Cramér-Rao Inequality.

Example 6.34. Consider a random sample that is uniform on $[0,\theta]$ with $\theta > 0$ unknown. The value of θ maximizing

$$\prod_{i=1}^{n} \frac{1}{\theta} 1_{[0,\theta]}(x_i) = \theta^{-n} 1_{x_1,\dots,x_n \in [0,\theta]} = \theta^{-n} 1_{x_{(1)},x_{(n)} \in [0,\theta]}$$

occurs when θ is as small as possible such that the likelihood is nonzero, since θ^{-n} is a decreasing function in θ. Once $\theta < x_{(n)}$, this expression is zero, so the smallest value of θ giving a nonzero likelihood is $\theta = x_{(n)}$. So, the unique global maximum occurs at $\theta = x_{(n)}$, so that $X_{(n)}$ is the MLE for θ. In contrast, recall that the UMVU for θ is $(1 + 1/n)X_{(n)}$, so both are asymptotically equivalent, though the MLE is biased.

Example 6.35. Consider a random sample from the exponential density $1_{x>0}\theta e^{-\theta x}$ with $\theta > 0$ unknown. Then

$$\log \prod_{i=1}^{n} 1_{x_i>0} \theta e^{-\theta x_i} = 1_{x_1,\dots,x_n>0} \log \theta - \theta \sum_{i=1}^{n} x_i.$$

So,

$$\frac{d}{d\theta} \log \prod_{i=1}^{n} 1_{x_i>0} \theta e^{-\theta x_i} = 1_{x_1,\dots,x_n>0} \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

As a function of θ, the likelihood is increasing for small θ and decreasing for large θ, so there is a unique maximum of

$$Y := \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i},$$

which is the MLE for θ. It turns out that

$$\operatorname{Var}(Y) = \operatorname{Var}\Big[n^{-1/2}\sqrt{n}\Big(\frac{1}{\overline{X}_n} - \theta\Big)\Big] = \frac{1}{n}\theta^2(1 + o(1)).$$

On the other hand, the information inequality, Theorem 6.20, says the smallest possible variance of an unbiased estimator of θ is

$$1/\operatorname{Var}\Big(\frac{n}{\theta} - \sum_{i=1}^{n} X_i\Big) = 1/(n\theta^{-2}) = \theta^2/n.$$

So, the MLE asymptotically achieves the optimal variance for an estimator of θ.

Example 6.36. Consider a random sample from the exponential density $1_{x>0}\theta e^{-\theta x}$ with $\theta > 0$ unknown. That is, we continue the previous example. Instead of finding an MLE for θ, suppose we want an MLE for $e^{-\theta}$. From the previous example, we can immediately conclude that

$$\psi = e^{-1/\sum_{i=1}^{n} x_i}.$$

by with $g(\theta) := e^{-\theta}$. Proposition 6.37.

Proposition 6.37 (Functional Equivariance of MLE). *Let $g\colon \Theta \to \Theta'$ be a bijection. Suppose Y is the MLE of θ. Then $g(Y)$ is the MLE of $g(\theta)$.*

Proof. By definition of the MLE Y, $Y(X_1,\dots,X_n)$ achieves the maximum value of $\theta \mapsto \ell(\theta)$. Writing $\ell(\theta) = \ell(g^{-1}g(\theta))$, we have the equivalent statement: $g(Y)(X_1,\dots,X_n)$ achieves the maximum value of $\theta' \mapsto \ell(g^{-1}(\theta'))$. □

So, unlike the UMVU, once we know the MLE for θ, we can easily get the MLE for invertible functions of θ.

Lemma 6.38 (**Likelihood Inequality**). *Let $X\colon \Omega \to \mathbb{R}^n$ be a random variable with probability density $f_\theta\colon \mathbb{R}^n \to [0,\infty)$. Let $f_\omega\colon \mathbb{R}^n \to [0,\infty)$ be another probability density. Assume that the probability laws $\mathbf{P}_\theta$ and $\mathbf{P}_\omega$ corresponding to f_θ and f_ω are not equal. Then the* ***Kullback-Leibler information***

$$I(\theta,\omega) := \mathbf{E}_\theta \log \frac{f_\theta(X)}{f_\omega(X)}$$

satisfies $I(\theta,\omega) > 0$.

Remark 6.39. If $\mathbf{P}_\theta(f_\omega(X) = 0 \text{ and } f_\theta(X) > 0) > 0$, then define $I(\theta,\omega) := \infty$, so there is nothing to prove. Also, in the definition of $I(\theta,\omega)$, if both densities take value zero, we define the ratio of zero over zero to be 1.

Proof. We may assume that $\mathbf{P}_\theta(f_\omega(X) = 0 \text{ and } f_\theta(X) > 0) = 0$. Note that $f_\theta(X) > 0$ with probability one with respect to $\mathbf{P}_\theta$. By Jensen's Inequality, Exercise 2.23,

$$-I(\theta,\omega) = \mathbf{E}_\theta \log \frac{f_\omega(X)}{f_\theta(X)} \le \log \mathbf{E}_\theta \frac{f_\omega(X)}{f_\theta(X)} = \log \int_{x\in\mathbb{R}^n\colon f_\theta(x)>0} \frac{f_\omega(x)}{f_\theta(x)} f_\theta(x)dx \le \log(1) = 0.$$

If $I(\theta,\omega) = 0$, then both of the inequalities above are equalities. The last inequality being an equality implies that $\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ and $\{x \in \mathbb{R}\colon f_\omega(x) > 0\}$ are equal almost everywhere. Since log is strictly concave, equality in the application of Jensen's Inequality implies that $\frac{f_\omega(X)}{f_\theta(X)}$ is constant almost surely (with respect to the probability law $\mathbf{P}_\theta$), therefore the densities f_ω and f_θ must be proportional, hence equal almost surely with respect to $\mathbf{P}_\theta$, so their corresponding probability laws are equal. □

Theorem 6.40 (**Consistency of MLE**). *Let $X_1, X_2, \ldots : \Omega \to \mathbb{R}^n$ be i.i.d. random variables with common probability density $f_\theta\colon \mathbb{R}^n \to [0,\infty)$. Fix $\theta \in \Theta \subseteq \mathbb{R}^m$. Suppose Θ is compact and $f_\theta(x_1)$ is a continuous function of θ for a.e. $x_1 \in \mathbb{R}$. (Then the maximum of $\ell(\theta)$ exists, since it is a continuous function on a compact set.) Assume that $\mathbf{E}_\theta \sup_{\theta'\in\Theta} |\log f_{\theta'}(X_1)| < \infty$, and $\mathbf{P}_\theta \neq \mathbf{P}_{\theta'}$, for all $\theta' \neq \theta$. Then, as $n \to \infty$, the MLE Y_n of θ converges in probability to the constant function θ, with respect to $\mathbf{P}_\theta$.*

Proof. For simplicity we assume that Θ is finite. For a full proof, see the Keener book, Theorem 9.11. Fix $\theta \in \Theta$.

For any $\theta' \in \Theta$ and $n \ge 1$, let $\ell_n(\theta') := \frac{1}{n}\sum_{i=1}^n \log f_{\theta'}(X_i)$. Denote $\Theta = \{\theta, \theta_1, \ldots, \theta_k\}$. By the Weak Law of Large Numbers, Theorem 5.10, for any $\theta' \in \Theta$, $\ell_n(\theta')$ converges in probability with respect to $\mathbf{P}_\theta$ to the constant $\mu(\theta') := \mathbf{E}_\theta \log f_{\theta'}(X_1)$ as $n \to \infty$. Since $\mathbf{P}_\theta \neq \mathbf{P}_{\theta'}$, for all $\theta' \neq \theta$, we have $\mu(\theta) > \mu(\theta')$ for all $\theta' \in \Theta$ with $\theta' \neq \theta$, by Lemma 6.38 (since $I(\theta,\theta') = \mu(\theta) - \mu(\theta') > 0$). For any $n \ge 1$, let

$$A_n := \{\ell_n(\theta) > \ell_n(\theta_j), \quad \forall\, 1 \le j \le k\}.$$

Then $\lim_{n\to\infty} \mathbf{P}_\theta(A_n) = 1$, and on the set A_n, the MLE Y_n is well-defined and unique with $Y_n = \theta$, so $\{Y_n = \theta\}^c \subseteq A_n^c$, and for any $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}_\theta(|Y_n - \theta| > \varepsilon) \le \lim_{n\to\infty} \mathbf{P}_\theta(A_n^c) = 0.$$

□

If $g\colon \Theta \to \Theta'$ is a bijection, it follows from Proposition 6.37 that the MLE for $g(\theta)$ is also consistent.

The above Theorem is analogous to a weak law of large numbers, since it gives convergence in probability of the MLE. Continuing this analogy, the following Theorem is analogous to the Central Limit Theorem, since it gives the limiting distribution of the MLE.

Theorem 6.41 (Limiting Distribution of MLE). *Let $\{f_\theta\colon \theta \in \Theta\}$ be a family of probability density functions, so that $f_\theta\colon \mathbb{R}^n \to [0,\infty)\ \forall\ \theta \in \Theta$. Let $X_1, X_2, \dots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \subseteq \mathbb{R}$. Assume the following*

(i) *The set $A := \{x \in \mathbb{R}\colon f_\theta(x) > 0\}$ does not depend on θ.*

(ii) *For every $x \in A$, $\partial^2 f_\theta(x)/\partial\theta^2$ exists and is continuous in θ.*

(iii) *The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbf{E}_\theta \frac{d}{d\theta} \log f_\theta(X_1) = 0$ and*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\Big(\frac{d}{d\theta}\log f_\theta(X_1)\Big)^2 = -\mathbf{E}_\theta \frac{d^2}{d\theta^2}\log f_\theta(X_1) > 0.$$

(iv) *For every θ in the interior of Θ, $\exists\ \varepsilon > 0$ such that*

$$\mathbf{E}_\theta \sup_{\theta' \in \Theta}\left|1_{\theta' \in [\theta-\varepsilon,\theta+\varepsilon]} \frac{d^2}{d[\theta']^2}\log f_{\theta'}(X_1)\right| < \infty.$$

(v) *The MLE Y_n of θ is consistent.*

Then, for any θ in the interior of Θ, as $n \to \infty$,

$$\sqrt{n}(Y_n - \theta)$$

converges in distribution to a mean zero Gaussian with variance $\frac{1}{I_{X_1}(\theta)}$, with respect to $\mathbf{P}_\theta$.

Remark 6.42. Combining this Theorem with Proposition 6.37, under the above assumptions (and also if the variance of the MLE converges), the MLE for θ achieves the asymptotically optimal variance in the Cramér-Rao Inequality, Theorem 6.20. The same holds for an invertible function of θ.

Proof. For simplicity we assume that Θ is finite. For a full proof, see the Keener book, Theorem 9.14. Fix $\theta \in \Theta$. (When Θ is finite, it has no interior, so the theorem is vacuous in this case, but the proof below is meant to illustrate the general case while avoiding a few technicalities.)

For any $\theta' \in \Theta$ and $n \geq 1$, let $\ell_n(\theta') := \frac{1}{n}\sum_{i=1}^n \log f_{\theta'}(X_i)$.

Choose $\varepsilon > 0$ sufficiently small such that $[\theta-\varepsilon, \theta+\varepsilon] \cap \Theta = \{\theta\}$. For any $n \geq 1$, let A_n be the event that $Y_n = \theta$. Since $Y_1, Y_2, \dots$ is consistent by Assumption (v), $\lim_{n\to\infty} \mathbf{P}_\theta(A_n) = 1$. Since Y_n maximizes ℓ_n, we have $\ell_n'(Y_n) = 0$ on A_n. (Since Θ is finite, this is not true, so take it as an additional assumption.) Taylor expanding ℓ_n' then gives

$$0 = \ell_n'(Y_n) = \ell_n'(\theta) + \ell_n''(Z_n)(Y_n - \theta), \qquad \text{if } A_n \text{ occurs},$$

where Z_n lies between θ and Y_n. Rewriting this equation gives

$$\sqrt{n}(Y_n - \theta) = \frac{\sqrt{n}\ell_n'(\theta)}{-\ell_n''(Z_n)}, \qquad \text{if } A_n \text{ occurs.} \qquad (*)$$

By Assumption (iii), the summed terms in $\ell_n'(\theta)$ i.i.d. random variables with mean zero and variance $I_{X_1}(\theta)$. So, the Central Limit Theorem 5.20 says that $\sqrt{n}\ell_n'(\theta)$ converges in distribution to a mean zero Gaussian with variance $I_{X_1}(\theta)$.

We now examine the denominator of $(*)$. By Assumption (iv) and the Weak Law of Large Numbers, $\ell_n''(\theta')$ converges in probability to $\mathbf{E}_\theta \ell_n''(\theta')$. Since $|Z_n - \theta| \leq |Y_n - \theta|$ when A_n occurs, we conclude that Z_n also converges in probability to θ as $n \to \infty$. Since Z_n only takes finitely many values, $\ell_n''(Z_n)$ converges in probability to $\mathbf{E}_\theta \ell_n''(\theta) \overset{(iii)}{=} -I_{X_1}(\theta)$. So, $(*)$ implies that $\sqrt{n}(Y_n - \theta)$ converges in distribution as $n \to \infty$ to a mean zero Gaussian with variance

$$\frac{I_{X_1}(\theta)}{[I_{X_1}(\theta)]^2} = \frac{1}{I_{X_1}(\theta)}.$$

So, we are done by Exercise 6.43. □

Exercise 6.43. Suppose $W_1, W_2, \ldots$ are random variables that converge in distribution to a random variable W, and $U_1, U_2, \ldots$ is any sequence of random variables. Let $A_1, A_2, \ldots \subseteq \Omega$ satisfy $\lim_{n\to\infty} \mathbf{P}(A_n) = 1$. Then, as $n \to \infty$

$$W_n 1_{A_n} + U_n 1_{A_n^c}$$

converges in distribution to W.

6.5. **Additional Comments.** The Cramér-Rao and Limiting Distribution for the MLE have analogous statements when Θ is a vector space.

Theorem 6.44 (**Multiparameter Cramér-Rao/ Information Inequality**). *Suppose $X\colon \Omega \to \mathbb{R}^n$ is a random variable with distribution from a family of multivariable probability densities or probability mass functions $\{f_\theta\colon \theta \in \Theta\}$. Assume that $\Theta \subseteq \mathbb{R}^m$ is an open set. We assume that $\{x \in \mathbb{R}^n\colon f_\theta(x) > 0\}$ does not depend on θ, and for a.e. $x \in \mathbb{R}^n$, and for all $1 \leq i \leq m$, $(\partial/\partial\theta_i) f_\theta(x)$ exists and is finite. Define the **Fisher information** of the family to be the $m \times m$ matrix $I(\theta) = I_X(\theta)$, so that if $1 \leq i, j \leq m$, the (i,j) entry of $I(\theta)$ is*

$$\mathrm{Cov}_\theta\Big(\frac{\partial}{\partial\theta_i}\log f_\theta(X), \frac{\partial}{\partial\theta_j}\log f_\theta(X)\Big) = \mathbf{E}_\theta\Big(\frac{\partial}{\partial\theta_i}\log f_\theta(X)\cdot\frac{\partial}{\partial\theta_j}\log f_\theta(X)\Big), \qquad \forall\,\theta\in\Theta,$$

and assume this quantity exists and is finite. Moreover, assume that $I(\theta)$ is an invertible matrix. (It is symmetric positive semidefinite by e.g. Exercise 6.45, but it might have a zero eigenvalue, a priori.)

Let $t\colon \mathbb{R}^n \to \mathbb{R}^m$ and let $Y := t(X)$ be statistic. For any $\theta \in \Theta$, let $g(\theta) := \mathbf{E}_\theta Y$ so that $g\colon \Theta \to \Theta$. Assume that all first order partial derivatives of g exist and are continuous. We assume we can differentiate under the integral sign. Let $Dg(\theta)$ denote the matrix of first order partial derivatives of g, and let $\mathrm{Var}_\theta(Y)$ denote the vector of variances of the components of Y. Then

$$\mathrm{Var}_\theta(Y) \geq (Dg(\theta))^T [I_X(\theta)]^{-1} Dg(\theta), \qquad \forall\,\theta\in\Theta.$$

In particular, if Y is unbiased for θ,

$$\mathrm{Var}_\theta(Y) \geq [I_X(\theta)]^{-1}, \qquad \forall\,\theta\in\Theta.$$

Equality occurs for some $\theta \in \Theta$ only when $\frac{d}{d\theta}\log f_\theta(X)$ and $Y - \mathbf{E}_\theta Y$ are multiples of each other.

Exercise 6.45. Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be a Gaussian random vector.

- Show that the covariance matrix $(a_{ij})_{1\le i,j\le d}$ of Z is symmetric, positive semidefinite. That is, for any $v \in \mathbb{R}^d$, we have

$$v^T a v = \sum_{i,j=1}^{d} v_i v_j a_{ij} \ge 0.$$

- Given any symmetric positive semidefinite matrix $(b_{ij})_{1\le i,j\le d}$, show that there exists a Gaussian random vector Z such that the covariance matrix of Z is $(b_{ij})_{1\le i,j\le d}$. (Hint: write the matrix b in its Cholesky decomposition $b = rr^*$, where r is a $d \times d$ real matrix. Let $e^{(1)}, \dots, e^{(d)}$ be the rows of r. Let $X_1, \dots, X_d$ be independent standard Gaussian random variables. Let $X := (X_1, \dots, X_d)$. Define $Z_i := \langle X, e^{(i)} \rangle$ for any $1 \le i \le d$.)

Theorem 6.46 (**Limiting Distribution of MLE**). *Let $\{f_\theta \colon \theta \in \Theta\}$ be a family of probability density functions, so that $f_\theta \colon \mathbb{R}^n \to [0,\infty)$ $\forall\ \theta \in \Theta$. Let $X_1, X_2, \dots$ be i.i.d. such that X_1 has density f_θ. Let $\Theta \subseteq \mathbb{R}^m$. Assume the following*

(i) *The set $A := \{x \in \mathbb{R}^n \colon f_\theta(x) > 0\}$ does not depend on θ.*

(ii) *For every $x \in A$, $\forall\ 1 \le i, j \le m$, $\frac{\partial^2 f_\theta(x)}{\partial\theta_i \partial\theta_j}$ exists and is continuous in θ.*

(iii) *The Fisher Information $I_{X_1}(\theta)$ exists and is finite, with $\mathbf{E}_\theta \nabla_\theta \log f_\theta(X_1) = 0$ and*

$$I_{X_1}(\theta) = \mathbf{E}_\theta\Big(\frac{\partial}{\partial\theta_i} \log f_\theta(X) \cdot \frac{\partial}{\partial\theta_j} \log f_\theta(X)\Big) = -\mathbf{E}_\theta D_\theta^2 \log f_\theta(X_1).$$

(D_θ^2 denotes the matrix of iterated second order derivatives in θ.) Moreover, assume that $I_{X_1}(\theta)$ is an invertible matrix.

(iv) *For every θ in the interior of Θ, $\forall\ 1 \le i, j \le m$, $\exists\ \varepsilon > 0$ such that*

$$\mathbf{E}_\theta \sup_{\theta' \in \Theta} \left| 1_{\theta' \in [\theta-\varepsilon, \theta+\varepsilon]} \frac{\partial^2}{\partial\theta_i' \partial\theta_j'} \log f_{\theta'}(X_1) \right| < \infty.$$

(v) *The MLE Y_n of θ is consistent.*

Then, for any θ in the interior of Θ, as $n \to \infty$,

$$\sqrt{n}(Y_n - \theta)$$

converges in distribution to a mean zero Gaussian random vector with covariance matrix $[I_{X_1}(\theta)]^{-1}$, with respect to $\mathbf{P}_\theta$.

7. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{R}$ denotes the set of real numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \colon x_i \in \mathbb{R} \, \forall \, 1 \leq i \leq n\}$

$f \colon A \to B$ means f is a function with domain A and range B. For example,

$f \colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall \, a \in A$, we have $a \in B$, so A is contained in B

$A \setminus B := \{a \in A \colon a \notin B\}$

$A^c := \Omega \setminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$\mathbf{P}$ denotes a probability law on Ω

$\mathbf{P}(A|B)$ denotes the conditional probability of A, given B.

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

Let X be a discrete random variable on a sample space Ω, so that $X \colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$. Let $A \subseteq \Omega$. Let Y be another discrete random variable

$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega \colon X(\omega) = x\}), \, \forall \, x \in \mathbb{R}$

the Probability Mass Function (PMF) of X

$\mathbf{E}(X)$ denotes the expected value of X

$\operatorname{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2$, the variance of X

$\sigma_X = \sqrt{\operatorname{var}(X)}$, the standard deviation of X

$X|A$ denotes the random variable X conditioned on the event A.

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

$1_A \colon \Omega \to \{0,1\}$, denotes the indicator function of A, so that

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

Let X, Y be a continuous random variables on a sample space Ω, so that $X, Y \colon \Omega \to \mathbb{R}$. Let $-\infty \leq a \leq b \leq \infty$, $-\infty \leq c \leq d \leq \infty$. Let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$.

$f_X \colon \mathbb{R} \to [0, \infty)$ denotes the Probability Density Function (PDF) of X, so

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

$f_{X,Y} \colon \mathbb{R} \to [0, \infty)$ denotes the joint PDF of X and Y, so

$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y) dx dy$$

$f_{X|A}$ denotes the Conditional PDF of X given A

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

Let X be a random variable on a sample space Ω, so that $X \colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x, t \in \mathbb{R}$. Let $i := \sqrt{-1}$.

$F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(\{\omega \in \Omega \colon X(\omega) \leq x\})$
the Cumulative Distibution Function (CDF) of X.

$M_X(t) = \mathbf{E}e^{tX}$ denotes the Moment Generating Function of X at $t \in \mathbb{R}$

$\phi_X(t) = \mathbf{E}e^{itX}$ denotes the Characteristic Function (or Fourier Transform) of X at $t \in \mathbb{R}$

Let $g, h \colon \mathbb{Z} \to \mathbb{R}$. Let $t \in \mathbb{Z}$.

$$(g * h)(t) = \sum_{j \in \mathbb{Z}} g(j) h(t-j) \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{Z}$$

Let $g, h \colon \mathbb{R} \to \mathbb{R}$. Let $t \in \mathbb{R}$.

$$(g * h)(t) = \int_{-\infty}^{\infty} g(x) h(t-x) dx \text{ denotes the convolution of } g \text{ and } h \text{ at } t \in \mathbb{R}$$

Let $f, g \colon \mathbb{R} \to \mathbb{R}$. We use the notation $f(t) = o(g(t))$, $\forall\, t \in \mathbb{R}$ to denote $\lim_{t \to 0} \left| \frac{f(t)}{g(t)} \right| = 0$.

USC Mathematics, Los Angeles, CA
E-mail address: stevenmheilman@gmail.com