All of Statistics: A Concise Course in Statistical Inference

Solutions by David A. Lee

All errors, typographical and substantive, and other offenses, are entirely my own.

Contents

Chapter 1: Probability

Import Packages

Please see the associated GitHub repo for all code and comments to simulations. [LINK HERE]

```
import numpy as np
from numpy.random import choice
import random
import matplotlib.pyplot as plt
import scienceplots
```

Question: 1.10.1

Fill in the details of the proof of Theorem 1.8. Also, prove the monotone decreasing case.

Theorem (Wasserman 1.8) (Continuity of Probabilities)

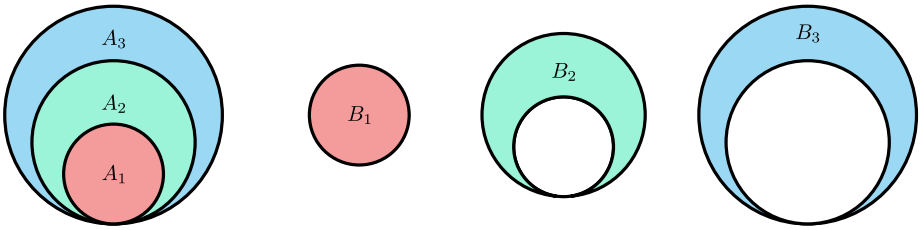
If $A_n \to A$ then

$$\mathbb{P}(A_n) \to \mathbb{P}(A)$$

as $n \to \infty$.

Proof. **Monotone Increasing Case.** The intuition we exploit here is that monotonicity alone does not allow us to employ Kolmogorov Axiom 3. Therefore, we need to rewrite the monotone A's in such a manner that the union of rewritten sets is disjoint.

Let A_n be monotone increasing. By definition, $A_1 \subset A_2 \subset \cdots$. Let $A = \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i$ and $B_1 = A_1, B_2 = \{\, \omega \in \Omega \colon \omega \in A_2, \omega \notin A_1 \,\}$, and so on. Take any B_i, B_j, with $i < j$. Our intuition here is that each B_i consists of the "new" ω that is "added" to the previous A_{i-1} to create the new A_i.



Now, let $\omega \in B_i$. We must have $\omega \notin B_j$, since $\omega \notin B_i$ is a condition for membership in B_j. Thus $B_i \cap B_j = \varnothing$, and each of the B's are pairwise disjoint. Since $A_1 \subset \cdots \subset A_n$, we have

$$A_n = \bigcup_{i=1}^{n} A_i$$

Now consider $\bigcup_{i=1}^{n} A_i$. Let $\omega \in A_n$. Then $\omega \in B_n$ implies $\omega \in \bigcup_{i=1}^{n} B_n$. Conversely, let $\omega \in \bigcup_{i=1}^{n} B_n$. Then $\omega \in B_i$ for some i, meaning we must have $\omega \in A_i$; with $A_i \subset A_n$, we must have $\omega \in A_n$. Ergo, $A_n = \bigcup_{i=1}^{n} B_n$.

We now have established

$$A_n = \bigcup_{i=1}^{n} A_n = \bigcup_{i=1}^{n} B_n$$

and can conclude

$$\lim_{n\to\infty} A_n = \lim_{n\to\infty} \bigcup_{i=1}^{n} A_n = \lim_{n\to\infty} \bigcup_{i=1}^{n} B_n = A$$

By Kolmogorov Axiom 3 and the disjointness of the B's, it follows that

$$P(A_n) = P\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} P(B_i)$$

leaving the final step of taking the limits

$$\lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \sum_{i=1}^{\infty} P(B_i) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = P(A)$$

Monotone Decreasing Case. With A_n monotone decreasing so that $A_1 \supset A_2 \supset \cdots$, let $A = \lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i$. By de Morgan's laws, we have

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

and when applied to the monotone increasing case,

$$\lim_{n\to\infty} P(A_n^c) = 1 - \lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i^c\right) = P(A^c) = 1 - P(A)$$

we can conclude

$$1 - \lim_{n\to\infty} P(A_n) = 1 - P(A) \implies \lim_{n\to\infty} P(A_n) = P(A)$$

□

Question: 1.10.2

Prove the statements in equation 1.1.

Property (Wasserman Eq. 1.1)

$$\begin{array}{rcl} \mathbb{P}(\varnothing) & = & 0 \\ A \subset B & \implies & \mathbb{P}(A) \leq \mathbb{P}(B) \\ 0 \leq & \mathbb{P}(A) & \leq 1 \\ \mathbb{P}(A^c) & = & 1 - \mathbb{P}(A) \\ A \bigcap B = \varnothing & \implies & P\big(A \bigcup B\big) = \mathbb{P}(A) + \mathbb{P}(B) \end{array}$$

Proof. Let A, B be events.

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

We know that $\Omega^c = \varnothing$. Then $\Omega^c \cap \Omega = \varnothing$ and $\Omega^c \cup \Omega = \Omega$. Combining Kolmogorov Axioms 2 and 3, conclude $\mathbb{P}(\Omega) = \mathbb{P}(\Omega^c) + \mathbb{P}(\Omega) = \mathbb{P}(\varnothing) + \mathbb{P}(\Omega) = 1$. Another application of Axiom 2 yields $\mathbb{P}(\varnothing) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0$.

$A \subset B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$

Consider A and $B - A$. We have $A \cap B - A = \varnothing$, as $\omega \in A$ implies $\omega \notin B - A$ and $\omega \in B - A$ implies $\omega \notin A$.

Now we prove $B = A \cup B - A$. If $\omega \in A \cup B - A$ either $\omega \in A$ or $\omega \in B - A$. If $\omega \in A$, then since $A \subset B$, we have $\omega \in B$. If $\omega \in B - A$, then $\omega \in B$. And if $\omega \in B$, then $\omega \in B - A$ implies $\omega \in A \cup B - A$. Thus $B = A \cup B - A$.

By Axiom 3,

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

since $\mathbb{P}(B - A) \geq 0$, this implies $\mathbb{P}(A) \leq \mathbb{P}(B)$.

$0 \leq \mathbb{P}(A) \leq 1$

By Axiom 1, $\mathbb{P}(A) \geq 0$. By Axioms 2 and 3, $\mathbb{P}(\Omega) = 1$. Then $\Omega = A \cup A^c$ and $A \cap A^c = \varnothing$, which in turn implies $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1$. Since $\mathbb{P}(A^c) \geq 0$, we have $\mathbb{P}(A) \leq 1$. Thus $0 \leq \mathbb{P}(A) \leq 1$.

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

Since $A^c \cap A = \varnothing$ and $A^c \cup A = \Omega$, by Axioms 2 and 3, we have $\mathbb{P}(\Omega) = 1 = \mathbb{P}(A) + \mathbb{P}(A^c)$. We conclude that $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

$A \cap B = \varnothing \implies \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

By disjointness of A, B, by Axiom 3, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$. □

Question: 1.10.3

Let Ω be a sample space and let $A_1, A_2, \ldots,$ be events. Define $B_n = \bigcup_{i=n}^{\infty} A_i$ and $C_n = \bigcap_{i=n}^{\infty} A_i$.

(a) Show that $B_1 \supset B_2 \supset \cdots$ and that $C_1 \subset C_2 \subset \cdots$.

(b) Show that $\omega \in \bigcap_{n=1}^{\infty} B_n$ if and only if ω belongs to an infinite number of the events $A_1, A_2, \ldots$.

(c) Show that $\omega \in \bigcup_{n=1}^{\infty} C_n$ if and only if ω belongs to all the events $A_1, A_2, \ldots$ except possibly a finite number of those events.

PROOF. (a) For the base case, we prove $B_1 \supset B_2$. If $\omega \in B_2 = \bigcup_{i=2}^{\infty} A_i$, then $\omega \in A_1 \cup \bigcup_{i=2}^{\infty} A_i = B_1$. Thus $B_1 \supset B_2$.

Now assume $B_1 \supset \cdots \supset B_n$ is true. We want to prove that $B_1 \supset \cdots \supset B_n \supset B_{n+1}$. It will suffice to show $B_n \supset B_{n+1}$. If $\omega \in B_{n+1} = \bigcup_{i=n+1}^{\infty} A_i$, then we have $\omega \in A_n \cup \bigcup_{i=n+1}^{\infty} A_i = B_n$, so $B_n \supset B_{n+1}$.

The proof for the C_n's is analogous. For the base case, $C_1 \subset C_2$, if $\omega \in C_1 = \bigcap_{i=1}^{\infty} A_i$ then $\omega \in A_i$ for all i. Then we must have $\omega \in A_2, A_3, \ldots$, and consequently, $\omega \in C_2 = \bigcap_{i=2}^{\infty} A_2$.

For the induction hypothesis, suppose that $C_1 \subset \cdots \subset C_n$. We must show $C_n \subset C_{n+1}$. If $\omega \in C_n = \bigcap_{i=n}^{\infty} A_i$, then we have $\omega \in A_{n+1}, \ldots$ implying $\omega \in C_{n+1} = \bigcap_{i=n+1}^{\infty} A_n$.

(b) ($\Longrightarrow$) Let $\omega \in \bigcap_{n=1}^{\infty} B_n$. Suppose ω belongs only to a finite number of events A_i. Pick the A_i with the highest subscript i. Now consider $B_{i+1} = \bigcup_{j=i+1}^{\infty} A_j$. Our assumption of finitude implies $\omega \notin B_{i+1}$, a contradiction. We will come to this contradiciton for any choice of A_i. Thus it must be that ω belogns to an infinite number of events A_i.

($\Longleftarrow$) Now suppose ω belongs to an infinite number of events A_i. We must show that $\omega \in \bigcap_{n=1}^{\infty} B_n$.

For $n = 1$, we must have $\omega \in B_1$, as ω is in infinitely many A_i's, and $B_1 = \bigcup_{i=1}^{\infty} A_i$.

Now suppose $\omega \in B_n = \bigcup_{i=n}^{\infty} A_i$. To demonstrate $\omega \in B_{n+1}$, we realize that we cannot have only $\omega \in A_n, \omega \notin A_{n+1}, \ldots$ as this would contradict the infinitude of membership. Thus we must have membership in A_i's with $i > n$. Ergo, $\omega \in B_{n+1}$.

By induction, $\omega \in B_i$ for all i. Thus $\omega \in \bigcap_{n=1}^{\infty} B_n$.

(c) ($\Longrightarrow$) Let $\omega \in \bigcup_{n=1}^{\infty} C_n$. Then $\omega \in C_n = \bigcap_{i=n}^{\infty} A_i$ for at least one n. If $\omega \in C_1 = \bigcap_{i=1}^{\infty} A_i$, then ω belongs to all events $A_1, A_2, \ldots$. Suppose otherwise that $\omega \in C_n$ for $n \neq 1$. Then ω lacks membership in at least one of the $A_1, \ldots, A_{n+1}$; in other words, it lacks membership in a finite number of those events.

($\Longleftarrow$) Let ω belong to all $A_1, A_2, \ldots$ or otherwise lack membership in a finite number of A_i. In the first case, $\omega \in \bigcap_{i=1}^{\infty} A_i = C_1$ implies $\omega \in \bigcup_{n=1}^{\infty} C_n$. Otherwise, pick the A_i with $\omega \notin A_i$ and the highest i. Then we have $\omega \in \bigcap_{j=i+1}^{\infty} A_j = C_{i+1}$, and still $\omega \in \bigcup_{n=i+1}^{\infty} C_n$. □

Question: 1.10.4

Let $\{ A_i : i \in I \}$ be a collection of events where I is an arbitrary index set. Show that

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i \in I} A_i^c \quad \text{and} \quad \left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c$$

Hint: First prove this for $I = \{1, \ldots, n\}$.

PROOF. These are the famous De Morgan's Laws.

(1) Let $\omega \in \left(\bigcup_{i \in I} A_i \right)^c$. Suppose $I = \{1, \ldots, n\}$. Then we have $\omega \notin \bigcup_{i \in I} A_i$, with the following chain of implications: $\omega \notin A_1, \ldots, A_n$ to $\omega \in A_1^c, \ldots, A_n^c$ to $\omega \in \bigcap_{i \in I} A_i^c$. Conversely, if $\omega \in \bigcap_{i \in I} A_i^c$, we must have $\omega \notin A_1, \ldots, A_n$, implying $\omega \notin \bigcup_{i \in I} A_i$, allowing us to conclude $\omega \in \left(\bigcup_{i \in I} A_i \right)^c$.

That is the finite case. Suppose now that we prove the infinite case. In the base case, trivially $\omega \in A_1^c$ implies itself. If we assume it holds for $n - 1$ A_i's, apply the proof for the n A_i's to prove the infinite case by induction.

(2) Let $\omega \in \left(\bigcap_{i \in I} A_i \right)^c$ for $i \in \{1, \ldots, n\}$. Then $\omega \notin \bigcap_{i \in I} A_i$, meaning ω is bereft of membership for at least one A_i. It follows that $\omega \in A_i^c$, implying $\omega \in \bigcup_{i \in I} A_i^c$. Now suppose $\omega \in \bigcup_{i \in I} A_i^c$. Then $\omega \in A_i^c$ for at least one i. If so, we cannot have $\omega \in \bigcap_{i \in I} A_i$, so it must follow that $\omega \in \left(\bigcap_{i \in I} A_i \right)^c$. Thus $\left(\bigcap_{i \in I} A_i \right)^c = \bigcup_{i \in I} A_i^c$. □

Question: 1.10.5

Suppose we toss a fair coin until we get exactly two heads. Describe the sample space S. What is the probability that exactly k tosses are required?

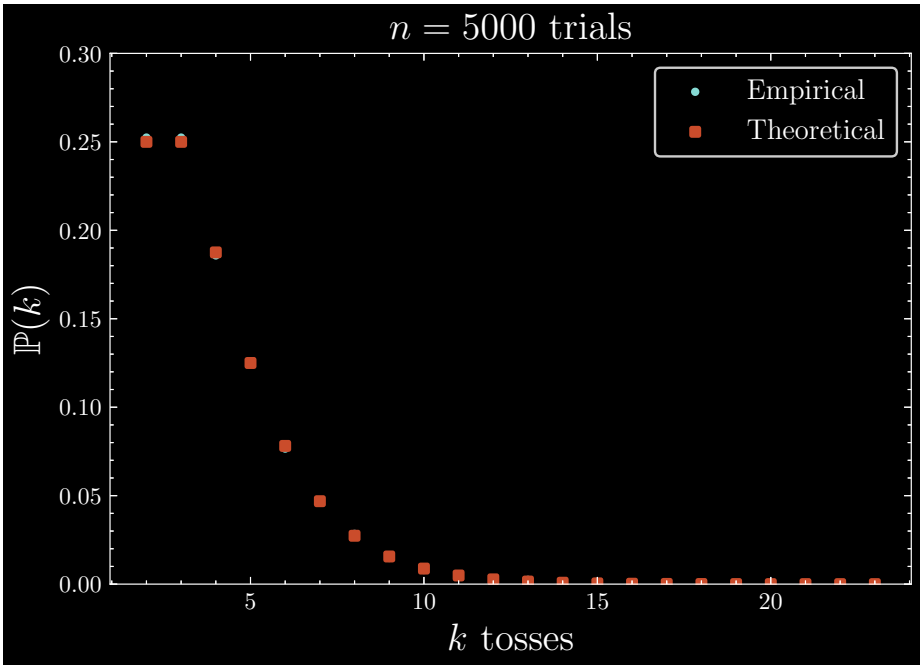
Descriptively, the sample space S is the set of events where the first head appears on any one of the $1, \ldots, k-1$-th toss. In general, the probability that exactly k tosses are needed is

$$\mathbb{P}(k \text{ tosses}) = (k-1)(1/2)^k$$

The reasoning is that since the second head is fixed on the k-th toss, there are $k-1$ tosses for the first head to be flipped. We simulate the coin toss experiment below:

```
def cointoss():
    tosses, heads = [], 0
    while (heads < 2):
        flip = random.randint(0,1)
        if (flip == 0):
            tosses.append('Heads')
            heads = tosses.count('Heads')
```

```
        else:
            tosses.append('Tails')
    return len(tosses)

def empirical(n):
    numtosses, probs = [], []
    for i in range(0,n):
        numtosses.append(cointoss())
    for x in np.unique(numtosses):
        probs.append((x, numtosses.count(x)/n))
    return probs

def theoretical():
    probs = []
    for i in range(2,24):
        probs.append((i, (i-1)*(0.5)**i))
    return probs

n = 5000
```



When we simulate the experiment for a large number of trials, we can see that the empirical results are quite close to the expected theoretical outcomes.

Question: 1.10.6

Let $\Omega = \{0, 1, \dots, \}$. Prove that there does not exist a uniform distribution on Ω (i.e., if $\mathbb{P}(A) = \mathbb{P}(B)$ whenever $|A| = |B|$, then $\mathbb{P}$ cannot satisfy the axioms of probability).

Proof. Suppose that Ω has a uniform distribution, and $\mathbb{P}(A) = \mathbb{P}(B)$ whenever $|A| = |B|$. Consider the disjoint singletons $\{0\}, \{1\}, \dots$. By Axiom 3, we have

$$\mathbb{P}\left(\bigcup_{k=0}^{\infty} \{k\}\right) = \sum_{k=0}^{\infty} \mathbb{P}(\{k\})$$

However,

$$\bigcup_{k=0}^{\infty} \{k\} = \Omega$$

which from Axiom 2 implies

$$\mathbb{P}\left(\bigcup_{k=0}^{\infty} \{k\}\right) = \mathbb{P}(\Omega) = 1$$

But by our uniform distribution assumption, each constituent of the sum of singleton probabilities is zero (picking a particular integer out of infinite possibilities is effectively zero probability). Thus we must have

$$\sum_{k=0}^{\infty} \mathbb{P}(\{k\}) = 0$$

But $\mathbb{P}\big(\bigcup_{k=0}^{\infty}\{k\}\big) = 1$, a contradiction. □

Question: 1.10.7

Let $A_1, A_2, \dots$ be events. Show that

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

Hint: Define $B_n = A_n - \bigcup_{i=1}^{n-1} A_i$. Then show that the B_n are disjoint and that $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$.

Proof. Define $B_n = A_n - \bigcup_{i=1}^{n-1} A_i$. First we prove the B_n are disjoint. Take B_k, B_j for $k \neq j$. Without loss of generality, suppose $k < j$. Let $\omega \in B_k$. Then $\omega \in A_k - \bigcup_{i=1}^{k-1} A_i$; importantly, $\omega \in A_k$. Then we must have $\omega \in \bigcup_{i=1}^{j-1} A_i$, implying $\omega \notin A_j - \bigcup_{i=1}^{j-1} A_i = B_j$.

Conversely, if $\omega \in B_j = A_j - \bigcup_{i=1}^{j-1} A_i$, we must have $\omega \notin \bigcup_{i=1}^{j-1} A_i$. Then we have $\omega \notin A_k$, implying $\omega \notin A_k - \bigcup_{i=1}^{k-1} A_i$. Thus $B_i \cap B_j = \varnothing$ and B_i, B_j are disjoint.

Now, we establish $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$. Let $\omega \in \bigcup_{n=1}^{\infty}$. Then $\omega \in A_k$ for some k('s). Choose the A_k with the lowest k. Then $\omega \in A_k - \bigcup_{i=1}^{k-1} A_i = B_k$ implies $\omega \in \bigcup_{n=1}^{\infty} B_n$.

Conversely, let $\omega \in \bigcup_{n=1}^{\infty} B_n$. Then $\omega \in B_k = A_k - \bigcup_{i=1}^{k=1} A_i$ for some k('s). Then $\omega \in A_k$ implies $\omega \in \bigcup_{n=1}^{\infty} A_n$. Thus $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$.

By Axiom 3 and the disjointness of B_n, we have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(B_n)$$

Since $B_n = A_n - \bigcup_{i=1}^{n-1} A_i \subset A_n$, we have $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ for all n. Then $\sum_{n=1}^{\infty} \mathbb{P}(B_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$. At last we may conclude

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

□

Question: 1.10.8

Suppose that $\mathbb{P}(A_i) = 1$ for each i. Prove that

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1$$

Proof. By premise, $\mathbb{P}(A_i = 1)$ implies $\mathbb{P}(A_i^c) = 1 - \mathbb{P}(A_i) = 0$. Now, since $A_1^c, A_2^c, \ldots$ all have probability zero, it must be the case that $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i^c) = 0$. By De Morgan's laws,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i^c\right) = \mathbb{P}\left(\left(\bigcap_{i=1}^{\infty} A_i\right)^c\right) = 0$$

Allowing us to conclude

$$\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = 1 - \mathbb{P}\left(\left(\bigcap_{i=1}^{\infty} A_i\right)^c\right) = 1$$

□

Question: 1.10.9

For fixed B such that $\mathbb{P}(B) > 0$, show that $\mathbb{P}(\cdot \mid B)$ satifies the axioms of probability.

Proof. (i) For independent $\cdot, B$, we have $\mathbb{P}(\cdot B) = \mathbb{P}(\cdot)\mathbb{P}(B)$, and if $\mathbb{P}(\cdot) \geq 0$ then $\mathbb{P}(\cdot \mid B) = \mathbb{P}(\cdot) \geq 0$. Otherwise, $\mathbb{P}(\cdot B), \mathbb{P}(B) \geq 0$ implies $\mathbb{P}(\cdot \mid B) \geq 0$.

(ii) If $\cdot = \Omega$, then

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

(iii) If $A_i \cap A_j = \varnothing$, then $(A_i \cap B) \cap (A_j \cap B) = \varnothing$. Then we have

$$\begin{aligned}
\mathbb{P}(A_i \cup A_j \mid B) &= \frac{\mathbb{P}((A_i \cup A_j) \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}((A_i \cap B) \cup (A_j \cap B))}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(A_i \cap B) + \mathbb{P}(A_j \cap B)}{\mathbb{P}(B)} \\
&= \mathbb{P}(A_i \mid B) + \mathbb{P}(A_j \mid B)
\end{aligned}$$

To generalize for infinitely many disjoint events A_i, observe that

$$\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B = \bigcup_{i=1}^{\infty} (A_i \cap B)$$

Applying the logic for the case with two disjoint events, we can immediately conclude

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B)$$

□

Question: 1.10.10

You have probably heard it before. Now you can solve it rigorously. It is called the "Monty Hall Problem." A prize is placed at random behind one of three doors. You pick a door. To be concrete, let's suppose you always pick door 1. Now Monty Hall chooses one of the other two doors, opens it and shows you that it is empty. He then gives you the opportunity to keep your door or switch to the other unopened door. Should you stay or switch? Intuition suggests it doesn't matter. The correct answer is you should switch. Prove it. It will help to specify the sample space and the relevant events carefully. Thus write $\Omega = \{ (\omega_1, \omega_2) : \omega_i \in \{1, 2, 3\} \}$ where ω_1 is where the prize is and ω_2 is the door Monty opens.

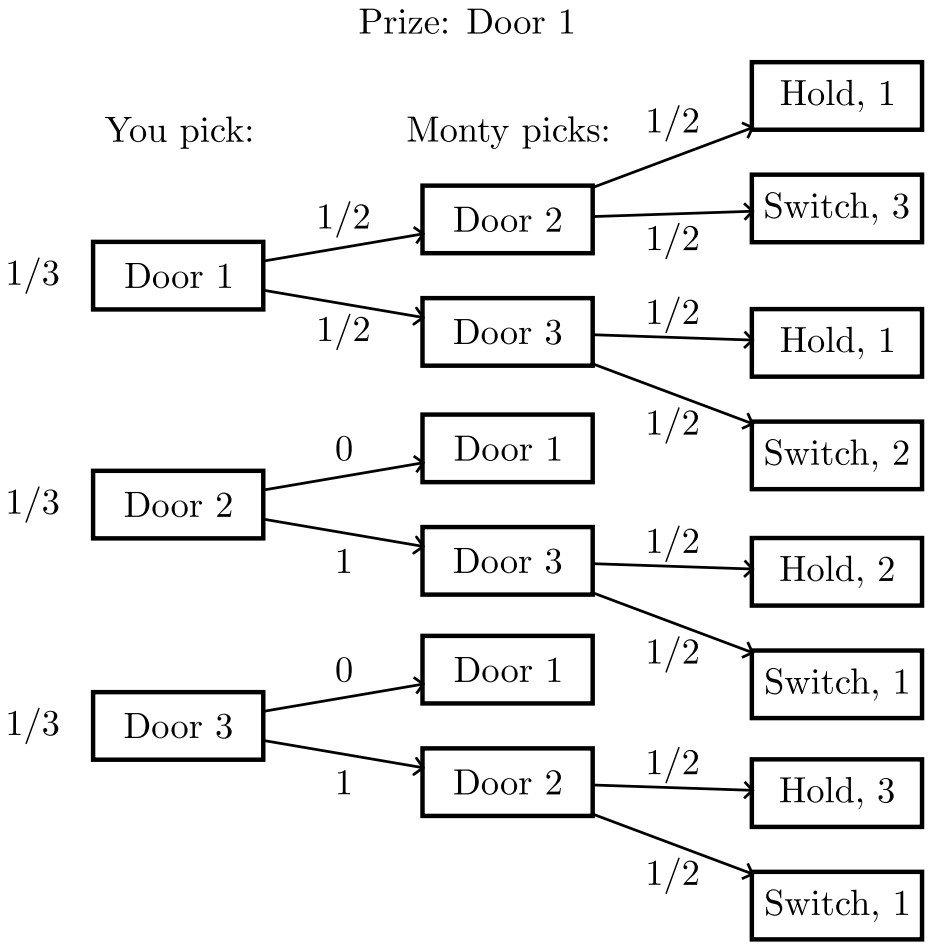
There are some insights we should be aware of. Firstly, there is equal chance of the prize appearing behind any one of the three doors, i.e., $\mathbb{P}(\text{Prize, Door } i) = 1/3$. Secondly, the choice of door for Monty Hall *depends on our choice of door*. Our pick of door also follows equal probability for each door, or $1/3$.

Using Bayes' Theorem, without loss of generality, suppose you pick door 1, the prize lies behind door 2, and Monty Hall picks door 3. We wish to find the probability that the prize is behind door 2, *given that Monty Hall opens door 3*.

$$\begin{aligned}
\mathbb{P}(\text{Prize, Door 2} \mid \text{Monty, Door 3}) &= \frac{\mathbb{P}(\text{M, D 3} \mid \text{P, D 2})\mathbb{P}(\text{P, D 2})}{\sum \mathbb{P}(\text{M, D 3} \mid \text{P, D } i)\mathbb{P}(\text{P, D } i)} \\
&= \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 1/3 + 0 \cdot 1/3} \\
&= \boxed{2/3}
\end{aligned}$$

Here, it is advantageous to switch doors, as the prior probability $(1/3)$ changes to the posterior probability $(2/3)$ once Monty Hall diverges additional information that his door is empty.

It is useful to illustrate the sample space by way of a probability tree:



Assuming without loss of generality that the prize lies behind door 1, we can calculate the conditional probabilities of winning given that we either hold or switch.

Let Y be your choice of door and M be Monty Hall's. Moreover, let α be

$$\begin{aligned}
\alpha &= \mathbb{P}(Y=1)\mathbb{P}(M=2 \mid Y=1)\mathbb{P}(\text{Hold } 1 \mid M=2) \\
&\quad + \mathbb{P}(Y=1)\mathbb{P}(M=3 \mid Y=1)\mathbb{P}(\text{Hold } 1 \mid M=3) \\
&\quad + \mathbb{P}(Y=2)\mathbb{P}(M=3 \mid Y=2)\mathbb{P}(\text{Hold } 2 \mid M=3) \\
&\quad + \mathbb{P}(Y=3)\mathbb{P}(M=2 \mid Y=3)\mathbb{P}(\text{Hold } 3 \mid M=2) \\
&= 1/3 \cdot 1/2 \cdot 1/2 + 1/3 \cdot 1/2 \cdot 1/2 + 1/3 \cdot 1/2 + 1/3 \cdot 1/2 \\
&= 1/2
\end{aligned}$$

We calculate

$$\begin{aligned}\mathbb{P}(\text{Win} \mid \text{Hold}) &= \underbrace{\frac{\mathbb{P}(Y=1)\mathbb{P}(M=2 \mid Y=1)\mathbb{P}(\text{Hold } 1 \mid M=2)}{\alpha}} \\ &\quad + \underbrace{\frac{\mathbb{P}(Y=1)\mathbb{P}(M=3 \mid Y=1)\mathbb{P}(\text{Hold } 1 \mid M=3)}{\alpha}} \\ &= \frac{1/3 \cdot 1/2 \cdot 1/2 + 1/3 \cdot 1/2 \cdot 1/2}{1/2} \\ &= \boxed{1/3}\end{aligned}$$

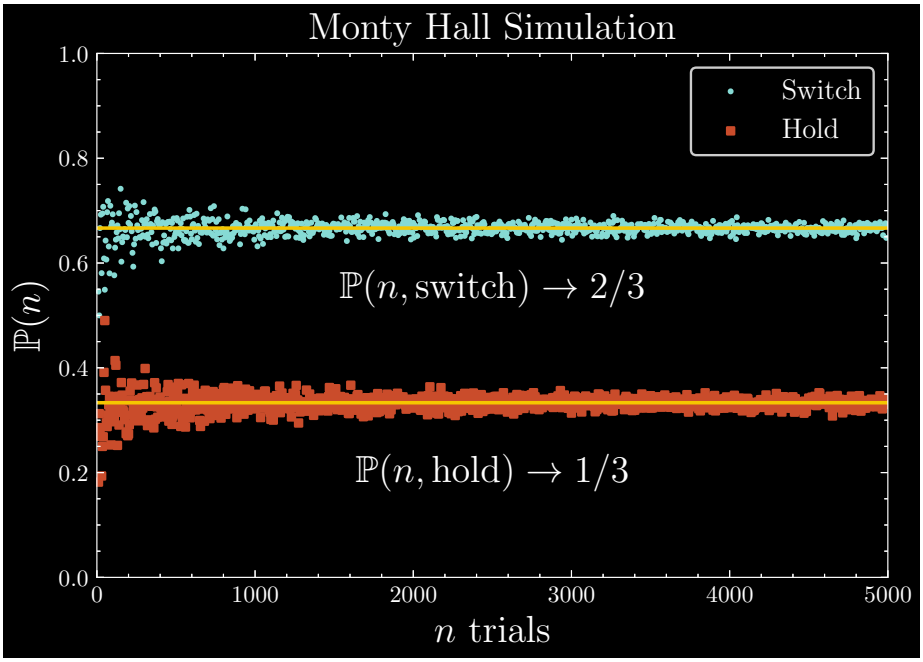
An analogous calculation for the switching conditional probability gives us

$$\mathbb{P}(\text{Win} \mid \text{Switch}) = \boxed{2/3}$$

Lastly, there is a famous anecdote of Erdös refusing to accept the result of the Monty Hall problem until examining a simulation of the game. You may find yourself burdened by a similar sentiment. Indeed, the experimental result aligns with our derivations above:

```
def montyhallswitch():
    doors = list(range(2, 4))
    prizedoor = random.randint(1,3)
    if (prizedoor == 2 or prizedoor == 3):
        montydoors = [x for x in doors if x != prizedoor]
        montypicks = montydoors[0]
    else:
        montypicks = random.randint(2,3)
    doors.remove(montypicks)
    yourswitch = doors[0]
    return(prizedoor == yourswitch)

def montyhallhold():
    prizedoor = random.randint(1,3)
    return(prizedoor == 1)

def iterationswitch(n):
    runs = []
    for i in range(1,n+1):
        runs.append(montyhallswitch())
    probs = runs.count(True)/(n+1)
    return probs

def iterationhold(n):
    runs = []
    for i in range(1,n+1):
        runs.append(montyhallhold())
    probs = runs.count(True)/(n+1)
    return probs

n = 5000
```

```
trials = list(range(10, n+1, 5))
switchprob, holdprob = [], []
for i in trials:
    switchprob.append(iterationswitch(i))
    holdprob.append(iterationhold(i))
```



Question: 1.10.11

Suppose that A and B are independent events. Show that A^c and B^c are independent events.

Proof. By premise, $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$. We have $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(B^c) = 1 - \mathbb{P}(B)$. To find $\mathbb{P}(A^cB^c)$, we observe these disjoint events:

$$\begin{aligned} \mathbb{P}(A - B) &= \mathbb{P}(A) - \mathbb{P}(AB) \\ \mathbb{P}(B - A) &= \mathbb{P}(B) - \mathbb{P}(AB) \\ &\mathbb{P}(AB) \\ &\mathbb{P}(A^cB^c) \end{aligned}$$

Thus we have

$$1 = \mathbb{P}(A - B) + \mathbb{P}(B - A) + \mathbb{P}(AB) + \mathbb{P}(A^cB^c)$$

which implies

$$\begin{aligned} \mathbb{P}(A^cB^c) &= 1 - \mathbb{P}(A - B) - \mathbb{P}(B - A) - \mathbb{P}(AB) \\ &= 1 - (\mathbb{P}(A) - \mathbb{P}(AB)) - (\mathbb{P}(B) - \mathbb{P}(AB)) - \mathbb{P}(AB) \\ &= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(AB) \end{aligned}$$

Now calculate

$$\begin{aligned}\mathbb{P}(A^c)\mathbb{P}(B^c) &= (1-\mathbb{P}(A))(1-\mathbb{P}(B)) \\ &= 1-\mathbb{P}(A)-\mathbb{P}(B)+\mathbb{P}(A)\mathbb{P}(B) \\ &= 1-\mathbb{P}(A)-\mathbb{P}(B)+\mathbb{P}(AB)\end{aligned}$$

Thus $\mathbb{P}(A^cB^c) = \mathbb{P}(A^c)\mathbb{P}(B^c)$, and A^c and B^c are independent. □

Question: 1.10.12

There are three cards. The first is green on both sides, the second is red on both sides and the third is green on one side and red on the other. We choose a card at random and we see one side (also chosen at random). If the side we see is green, what is the probability that the other side is also green? Many people intuitively answer 1/2. Show that the correct answer is 2/3.

PROOF. Let GG, RR, GR correspond to the first, second, and third cards. Using Bayes' Theorem: we calculate the probability of having drawn a GG card given that we see G, or GR given that we see G. Calculate

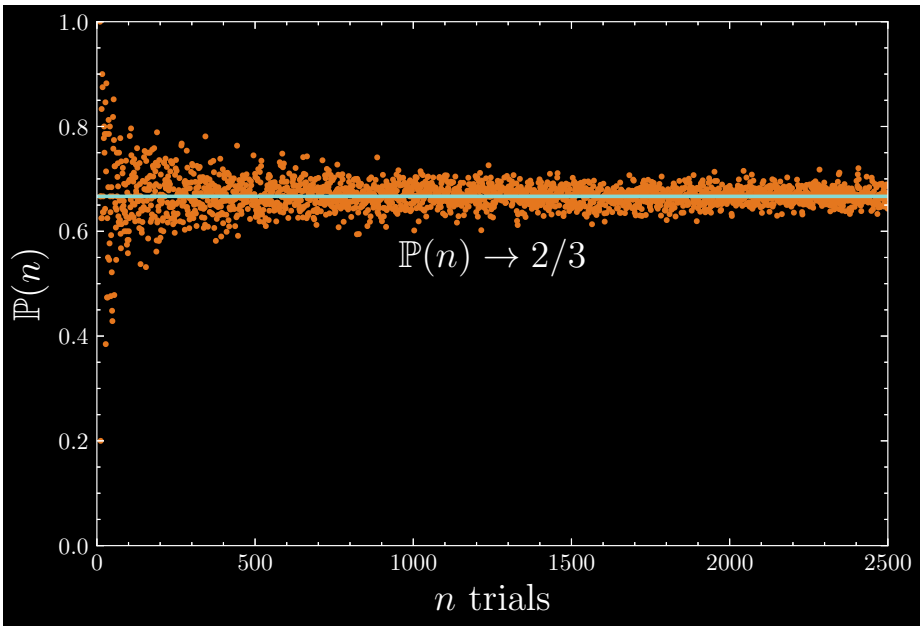
$$\begin{aligned}\mathbb{P}(GG \mid G) &= \frac{\mathbb{P}(G \mid GG)\mathbb{P}(GG)}{\mathbb{P}(G \mid GG)\mathbb{P}(GG)+\mathbb{P}(G \mid RR)\mathbb{P}(RR)+\mathbb{P}(G \mid GR)\mathbb{P}(GR)} \\ &= \frac{1 \cdot 1/3}{1 \cdot 1/3 + 0 \cdot 1/3 + 1/2 \cdot 1/3} \\ &= \boxed{2/3} \\ \mathbb{P}(GR \mid G) &= \frac{\mathbb{P}(G \mid GR)\mathbb{P}(GR)}{\mathbb{P}(G \mid GR)\mathbb{P}(GR)+\mathbb{P}(G \mid RR)\mathbb{P}(RR)+\mathbb{P}(G \mid GG)\mathbb{P}(GG)} \\ &= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} \\ &= \boxed{1/3}\end{aligned}$$

□

Simulating the card game, we find that the empirical outcome agrees with our theoretical result.

```
def cardgame():
    first, second, third = ('G','G'), ('R','R'), ('G','R')
    mychoice = random.choice([first, second, third])
    myside = random.choice(mychoice)
    if (myside == 'G' and 'R' not in mychoice):
        return(True, True)
    elif (myside == 'G' and 'R' in mychoice):
        return(True, False)
    else:
        return(False, False)
```

```
def iteration(n):
    runs = []
    for i in range(1,n+1):
        runs.append(cardgame())
    probs = runs.count((True, True))/
        (runs.count((True, True)) + runs.count((True, False)))
    return probs

n=2500
games = list(range(10, n+1))
cardprob = []
for i in games:
    cardprob.append(iteration(i))
```



Question: 1.10.13

Suppose that a fair coin is tossed repeatedly until both a head and tail have appeared at least once.

(a) Describe the sample space Ω.

(b) What is the probability that three tosses will be required?

(a) The sample space is given by

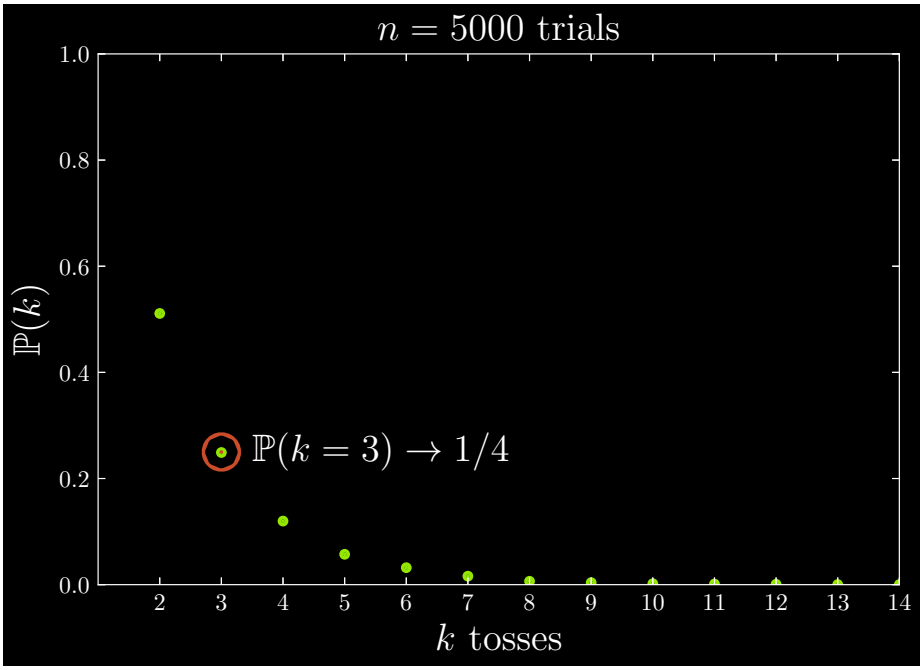
$$\Omega = \{HT, HHT, \ldots, \underbrace{H\cdots H}_{k}T, \ldots, TH, TTH, \ldots, \underbrace{T\cdots T}_{k}H, \ldots\}$$

(b) The theoretical derivation of the probability is

$$\begin{aligned}\mathbb{P}(3 \text{ tosses}) &= \mathbb{P}(HHT) + \mathbb{P}(TTH) \\ &= (1/2)^3 + (1/2)^3 \\ &= \boxed{1/4}\end{aligned}$$

By simulation, we can observe the empirical probability of three tosses aligns with the theoretical result, as well as the probabilities for other numbers of tosses:

```
def cointoss():
    tosses, heads, tails = [], 0, 0
    while (heads < 1 or tails < 1):
        flip = random.randint(0,1)
        if (flip == 0):
            tosses.append('Heads')
            heads = tosses.count('Heads')
        else:
            tosses.append('Tails')
            tails = tosses.count('Tails')
    return len(tosses)

def empirical(n):
    numtosses, probs = [], []
    for i in range(0,n):
        numtosses.append(cointoss())
    for x in np.unique(numtosses):
        probs.append((x, numtosses.count(x)/n))
    return probs

n = 5000
```



Question: 1.10.14

Show that if $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$ then A is independent of every other event. Show that if A is independent of itself then $\mathbb{P}(A)$ is either 0 or 1.

Proof. Let B be any event. Then if $\mathbb{P}(A) = 0$, we have $\mathbb{P}(AB) = 0$. Put differently, if A can never happen, then surely A *and* B could never happen. Now, it must also be true that $P(A)\mathbb{P}(B) = 0$, so it follows that $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) = 0$ and A is independent of all events.

Now suppose $\mathbb{P}(A) = 1$. Then $\mathbb{P}(AB) = \mathbb{P}(B)$. The reasoning is that since A will *always* happen, whether A and B happen together is contingent on B happening. Thus we can write $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)$, again proving independence of A relative to all other events.

Lastly, suppose A is independent of itself. Then $\mathbb{P}(AA) = \mathbb{P}(A) = \mathbb{P}(A)\mathbb{P}(A)$, implying $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$. □

Question: 1.10.15

The probability that a child has blue eyes is 1/4. Assume independence between children. Consider a family with 3 children.

(a) If it is known that at least one child has blue eyes, what is the probability that at least two children have blue eyes?

(b) If it is known that the youngest child has blue eyes, what is the probability that at least two children have blue eyes?

(a) The complement event to "at least one child has blue eyes" – call it A – is "no child has blue eyes." The probability that a child does not have blue eyes is $1 - p = 3/4$. Then we have

$$\mathbb{P}(A) = 1 - (3/4)^3 = 1 - 27/64 = 37/64$$

The probability that "at least two children have blue eyes" – call this one B – is the sum of the mutually exclusive probabilities of either two or three children having blue eyes. This is given by

$$\begin{aligned}\mathbb{P}(B) &= \mathbb{P}(2 \text{ children} + 3 \text{ children}) \\ &= \binom{3}{2}(1/4)^2(3/4) + (1/4)^3 \\ &= 9/64 + 1/64 \\ &= 10/64\end{aligned}$$

Now we can calculate the probability of at least two children having blue eyes, given that at least one child has blue eyes.

$$\begin{aligned}\mathbb{P}(B \mid A) &= \frac{\mathbb{P}(BA)}{\mathbb{P}(A)} \\ &= (10/64)(64/37) \\ &= \boxed{10/37}\end{aligned}$$

(b) The probability that the youngest child has blue eyes (call it C) is 1/4. Let D be the event that at least two children have blue eyes, with one of them being the youngest child. Then we have

$$\begin{aligned}\mathbb{P}(D) &= 2(1/4)^2(3/4) + (1/4)^3 \\ &= 6/64 + 1/64 \\ &= 7/64\end{aligned}$$

Lastly, the probability that at least two children have blue eyes given that the youngest child is known to have blue eyes is

$$\begin{aligned}\mathbb{P}(D \mid C) &= \frac{\mathbb{P}(DC)}{\mathbb{P}(C)} \\ &= \frac{7/64}{1/4} \\ &= \boxed{7/16}\end{aligned}$$

Question: 1.10.16

Prove Lemma 1.14.

Lemma (Wasserman 1.14)

If A and B are independent events then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Also, for any pair of events A and B,

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

Proof.

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

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

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

□

Question: 1.10.17

Show that

$$\mathbb{P}(ABC) = \mathbb{P}(A \mid BC)\mathbb{P}(B \mid C)\mathbb{P}(C)$$

Proof.

$$\mathbb{P}(ABC) = \mathbb{P}(A \mid BC)\mathbb{P}(BC) = \mathbb{P}(A \mid BC)\mathbb{P}(B \mid C)\mathbb{P}(C)$$

□

Question: 1.10.18

Suppose k events form a partition of the sample space Ω, i.e., they are disjoint and $\bigcup_{i=1}^{k} A_i = \Omega$. Assume that $\mathbb{P}(B) > 0$. Prove that if $\mathbb{P}(A_1 \mid B) < \mathbb{P}(A_1)$ then $\mathbb{P}(A_i \mid B) > \mathbb{P}(A_i)$ for some $i = 2, \ldots, k$.

Proof. By disjointness of events, we have

$$\sum_{i=1}^{k} \mathbb{P}(A_i) = 1 \quad \text{and} \quad \sum_{i=1}^{k} \mathbb{P}(A_i \mid B) = 1$$

with the second equality following from 1.10.9. Let $\mathbb{P}(A_1 \mid B) < \mathbb{P}(A_1)$. It cannot be the case that we have $\mathbb{P}(A_i \mid B) < \mathbb{P}(A_i)$ for the remaining $i = 2, \ldots, k$, for this would violate additivity of the conditional probabilities to unity:

$$\sum_{i=2}^{k} \mathbb{P}(A_i \mid B) < \sum_{i=2}^{k} \mathbb{P}(A_i) = 1$$

So would the condition that $\mathbb{P}(A_i \mid B) = \mathbb{P}(A_i)$ for the remaining i. Thus at least one of the $\mathbb{P}(A_i \mid B)$ must be greater than A_i for the unity sum condition to be met. □

Question: 1.10.19

Suppose that 30 percent of computer owners use a Macintosh, 50 percent use Windows, and 20 percent use Linux. Suppose that 65 percent of the Mac users have succumbed to a computer virus, 82 percent of the Windows users get the virus, and 50 percent of the Linux users get the virus. We select a person at random and learn that her system was infected with the virus. What is the probability that she is a Windows user?

By Bayes' Theorem:

$$\begin{aligned}\mathbb{P}(\text{Windows} \mid \text{Virus}) &= \frac{\mathbb{P}(\text{V} \mid \text{W})\mathbb{P}(\text{W})}{\mathbb{P}(\text{V} \mid \text{W})\mathbb{P}(\text{W}) + \mathbb{P}(\text{V} \mid \text{M})\mathbb{P}(\text{M}) + \mathbb{P}(\text{V} \mid \text{L})\mathbb{P}(\text{L})} \\ &= \frac{0.82 \cdot 0.5}{0.82 \cdot 0.5 + 0.65 \cdot 0.3 + 0.5 \cdot 0.2} \\ &= \boxed{0.582}\end{aligned}$$

Question: 1.10.20

A box contains 5 coins and each has a different probability of showing heads. Let $p_1, \ldots, p_5$ denote the probability of heads on each coin. Suppose that

$$p_1 = 0, p_2 = 1/4, p_3 = 1/2, p_4 = 3/4 \text{ and } p_5 = 1.$$

Let H denote "heads is obtained" and let C_i denote the event that coin i is selected.

(a) Select a coin at random and toss it. Suppose a head is obtained. What is the posterior probability that coin i was selected ($i = 1, \ldots, 5$)? In other words, find $\mathbb{P}(C_i \mid H)$ for $i = 1, \ldots, 5$.

(b) Toss the coin again. What is the probability of another head? In other words find $\mathbb{P}(H_2 \mid H_1)$ where $H_j =$ "heads on toss j."

Now suppose that the experiment was carried out as follows: We select a coin at random and toss it until a head is obtained.

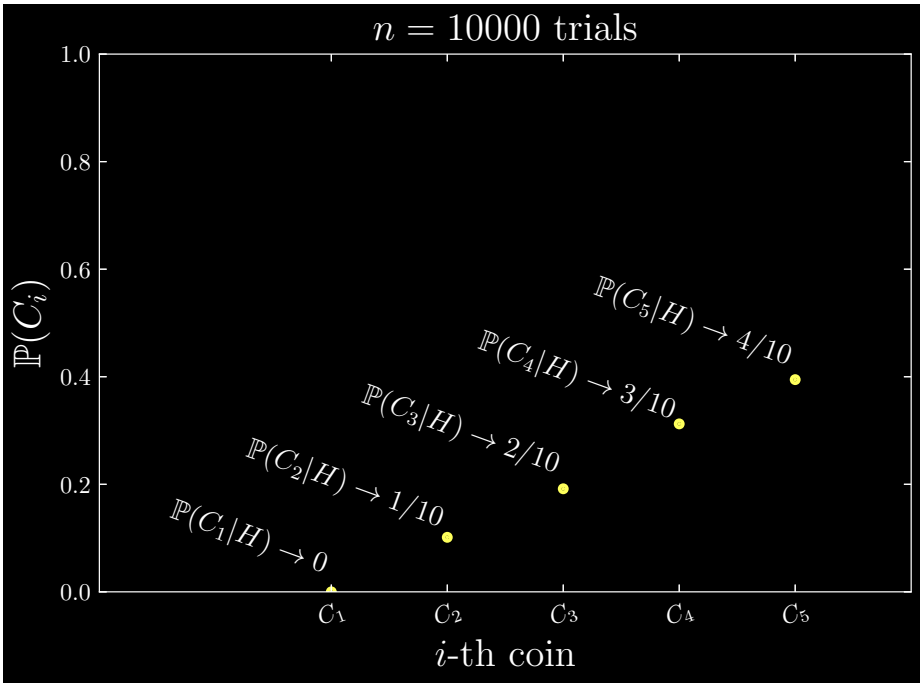
(c) Find $\mathbb{P}(C_i \mid B_4)$ where $B_4 =$ "first head is obtained in toss 4."

Apply Bayes' Theorem for all parts.
(a) Here, $\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i) = 1/5 \cdot \left(\sum p_i\right) = 1/5 \cdot 5/2 = 1/2$.

$$\mathbb{P}(C_1 \mid H) = \frac{\mathbb{P}(H \mid C_1)\mathbb{P}(C_1)}{\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i)} = \boxed{0}$$
$$\mathbb{P}(C_2 \mid H) = \frac{\mathbb{P}(H \mid C_2)\mathbb{P}(C_2)}{\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i)} = \frac{1/4 \cdot 1/5}{1/2} = \boxed{1/10}$$
$$\mathbb{P}(C_3 \mid H) = \frac{\mathbb{P}(H \mid C_3)\mathbb{P}(C_3)}{\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i)} = \frac{1/2 \cdot 1/5}{1/2} = \boxed{2/10}$$
$$\mathbb{P}(C_4 \mid H) = \frac{\mathbb{P}(H \mid C_4)\mathbb{P}(C_4)}{\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i)} = \frac{3/4 \cdot 1/5}{1/2} = \boxed{3/10}$$
$$\mathbb{P}(C_5 \mid H) = \frac{\mathbb{P}(H \mid C_5)\mathbb{P}(C_5)}{\sum \mathbb{P}(H \mid C_i)\mathbb{P}(C_i)} = \frac{1 \cdot 1/5}{1/2} = \boxed{4/10}$$

Empirical simulation:

```
def coinselect():
  coin = random.randint(1,5)
  return coin

def coinprob(i):
  coinface = ['H', 'T']
  headsprob=[0, 0.25, 0.5, 0.75, 1]
  randomFace = choice(
    coinface, 1, p=[headsprob[i—1], 1—headsprob[i—1]])
  return randomFace[0], i

def iteration(n):
  sample, probs = [], []
  for i in range(1,n):
    sample.append(coinprob(coinselect()))
  heads = len([i for i in sample if 'H' in i])
  for i in list(range(1,6)):
    probs.append(sample.count(('H',i)) / heads)
  return probs

n = 10000
```



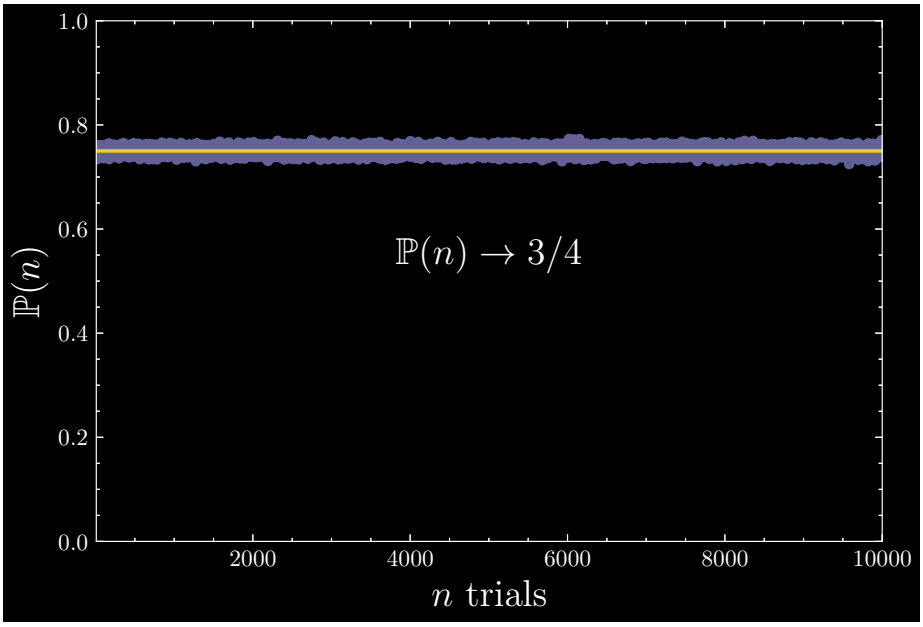
(b) By the Law of Total Probability, $\mathbb{P}(H_1) = 1/5 \cdot \sum p_i = 1/2$. To calculate $\mathbb{P}(H_2 \mid H_1)$, write

$$\begin{aligned}\mathbb{P}(H_2 \mid H_1) &= \frac{\mathbb{P}(H_2 H_1)}{\mathbb{P}(H_1)} \\ &= \frac{1/5 \cdot \sum p_i^2}{1/2} \\ &= \frac{3/8}{1/2} \\ &= \boxed{3/4}\end{aligned}$$

Empirical simulation:

```
def coinselect():
  coin = random.randint(1,5)
  return coin

def coinprob(i, k):
  coinface = ['H', 'T']
  headsprob=[0, 0.25, 0.5, 0.75, 1]
  randomFace = choice(
    coinface, k, p=[headsprob[i-1], 1-headsprob[i-1]])
  return randomFace.tolist(), i

def iteration(n):
  sample, probs = [], []
  for i in range(1,n):
```

```
    sample.append(coinprob(coinselect(), k))
  firsttossheads = len(
    [i for i in sample if ['H', 'H'] in i or ['H', 'T'] in i])
  for i in range(1,6):
    probs.append(sample.count((['H','H'],i)) / firsttossheads)
  return sum(probs)

n, k, probs = 10000, 2, []

for i in range(10,n+1):
  probs.append(iteration(n))
```



(c) Here, $\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i) = 1/5 \cdot \sum (1-p_i)^3 p_i = 0.0359375$.

$$\mathbb{P}(C_1 \mid B_4) = \frac{\mathbb{P}(B_4 \mid C_1)\mathbb{P}(C_1)}{\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i)} = \boxed{0}$$

$$\mathbb{P}(C_2 \mid B_4) = \frac{\mathbb{P}(B_4 \mid C_2)\mathbb{P}(C_2)}{\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i)} = \frac{(3/4)^3(1/4)(1/5)}{0.0359375} = \boxed{0.587}$$

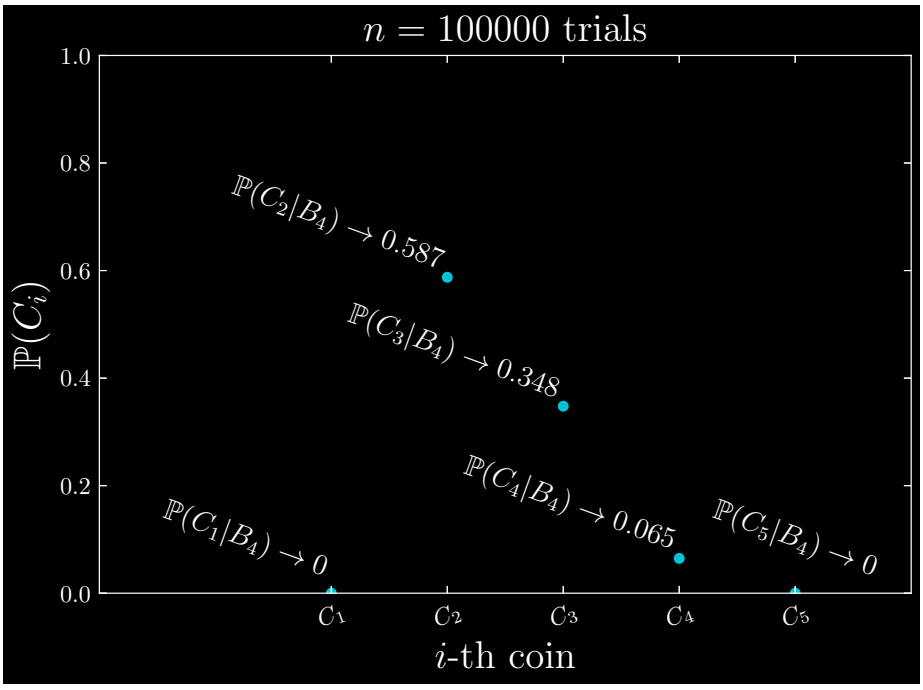
$$\mathbb{P}(C_3 \mid B_4) = \frac{\mathbb{P}(B_4 \mid C_3)\mathbb{P}(C_3)}{\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i)} = \frac{(1/2)^4(1/5)}{0.0359375} = \boxed{0.348}$$

$$\mathbb{P}(C_4 \mid B_4) = \frac{\mathbb{P}(B_4 \mid C_4)\mathbb{P}(C_4)}{\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i)} = \frac{(1/4)^3(3/4)(1/5)}{0.0359375} = \boxed{0.065}$$

$$\mathbb{P}(C_5 \mid B_4) = \frac{\mathbb{P}(B_4 \mid C_5)\mathbb{P}(C_5)}{\sum \mathbb{P}(B_4 \mid C_i)\mathbb{P}(C_i)} = \boxed{0}$$

Empirical simulation:

```
def coinselect():
  coin = random.randint(1,5)
  return coin

def coinprob(i, k):
  coinface = ['H', 'T']
  headsprob=[0, 0.25, 0.5, 0.75, 1]
  randomFace = choice(
    coinface, k, p=[headsprob[i—1], 1—headsprob[i—1]])
  return randomFace.tolist(), i

def iteration(n):
  sample, probs = [], []
  for i in range(1,n):
    sample.append(coinprob(coinselect(),k))
  threetails = len([i for i in sample if ['T', 'T', 'T', 'H'] in i])
  for i in range(1,6):
    probs.append(
      (i, sample.count((['T', 'T', 'T', 'H'],i)) / threetails))
  return probs

n, k = 100000, 4
iteration(n)
```

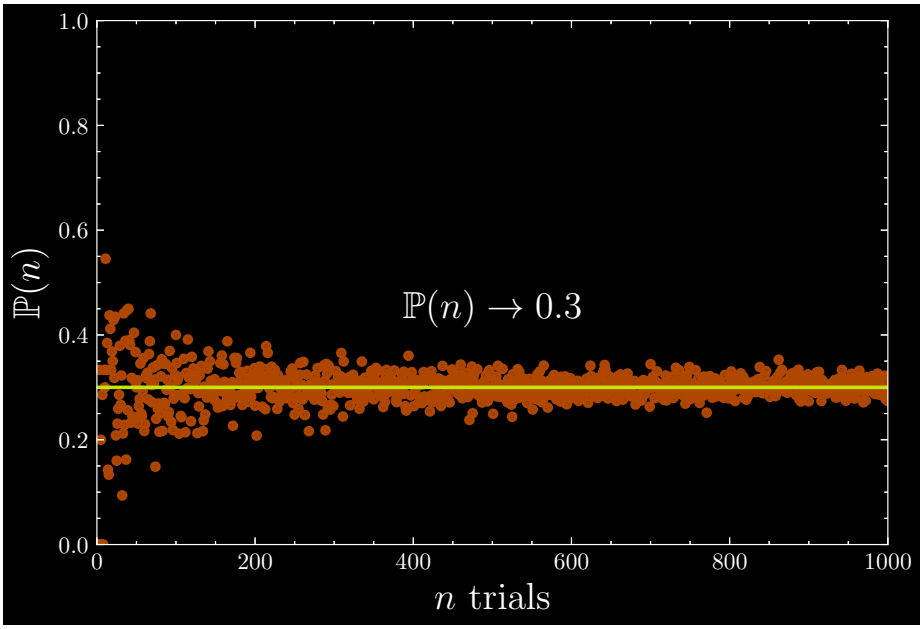


Question: 1.10.21

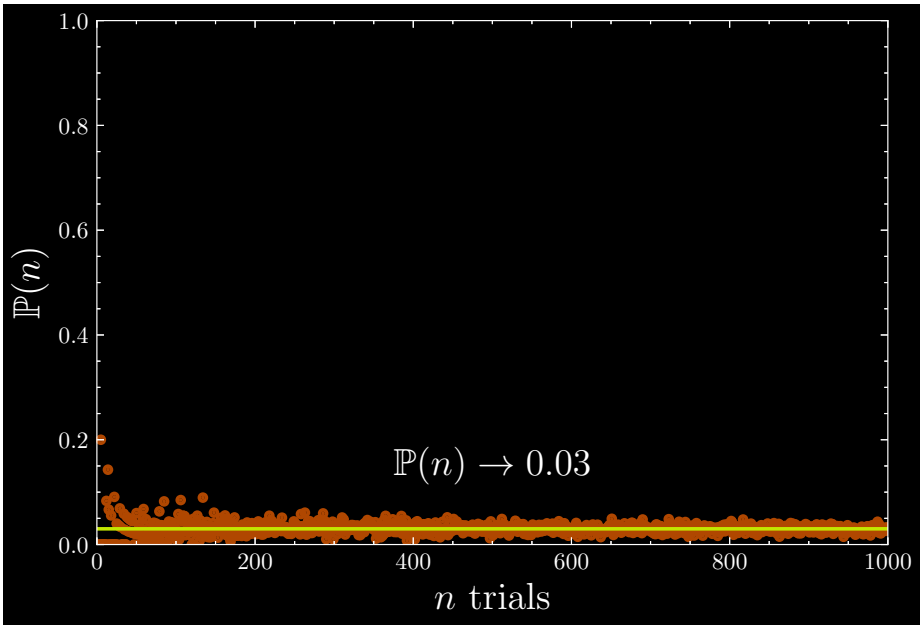
(Computer Experiment.) Suppose a coin has probability p of falling heads up. If we flip the coin many times, we would expect the proportion of heads to be near p. We will make this formal later. Take $p = .3$ and $n = 1,000$ and simulate n coin flips. Plot the proportion of heads as a function of n. Repeat for $p = .03$.

As intuitively anticipated, for large n, the probability of heads tends towards 0.3 or 0.03.

```
def cointoss():
    coinface = ['H', 'T']
    randomFace = choice(
        coinface, 1, p=[0.3, 0.7])
    return randomFace.tolist()[0]

def iteration(n):
    prob = []
    for i in range(1,n+1):
        prob.append(cointoss())
    return prob

def trials(n):
    prob = []
    for i in range(1,n+1):
        prob.append(iteration(i).count('H')/i)
    return prob

n = 1000
```




```
def cointoss():
    coinface = ['H', 'T']
    randomFace = choice(
        coinface, 1, p=[0.03, 0.97])
    return randomFace.tolist()[0]

def iteration(n):
    prob = []
    for i in range(1,n+1):
        prob.append(cointoss())
    return prob

def trials(n):
    prob = []
    for i in range(1,n+1):
        prob.append(iteration(i).count('H')/i)
    return prob

n = 1000
```



Question: 1.10.22

(Computer Experiment.) Suppose we flip a coin n times and let p denote the probability of heads. Let X be the number of heads. We call X a binomial random variable, which is discussed in the next chapter. Intuition suggests that X will be close to np. To see if this is true, we can repeat this experiment many times and average the X values. Carry out a simulation and compare the average of the X's to np. Try this for $p = .3$ and $n = 10, n = 100$, and $n = 1,000$.

We run 1,000 trials for 10, 100, and 1,000 flips. The average number of heads for each number of flips are 2.996, 29.958, and 299.92, compared to $np = 3$, 30, and 300.

```
from statistics import mean

def cointoss():
    coinface = ['H', 'T']
    randomFace = choice(
        coinface, 1, p=[0.3, 0.7])
    return randomFace.tolist()[0]

def trials(n, k):
    trials = []
    for j in range(1,k+1):
        probs = []
        for i in range(1,n+1):
            probs.append(cointoss())
        trialcount = probs.count('H')
        trials.append(trialcount)
```

```
    return trials

p, n1, n2, n3, k = 0.3, 10, 100, 1000, 1000
print("Running an experiment of", n1, "flips over", k,
        "trials, we average", mean(trials(n1,k)),
        "heads, compared to np =", n1*p)
print("Running an experiment of", n2, "flips over", k,
        "trials, we average", mean(trials(n2,k)),
        "heads, compared to np =", n2*p)
print("Running an experiment of", n3, "flips over", k,
        "trials, we average", mean(trials(n3,k)),
        "heads, compared to np =", n3*p)
```

Question: 1.10.23

(Computer Experiment.) Here we will get some experience simulating conditional probabilities. Consider tossing a fair die. Let $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Then, $\mathbb{P}(A) = 1/2, \mathbb{P}(B) = 2/3$ and $\mathbb{P}(AB) = 1/3$. Since $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$, the events A and B are independent. Simulate draws from the sample space and verify that $\widehat{\mathbb{P}}(AB) = \widehat{\mathbb{P}}(A)\widehat{\mathbb{P}}(B)$ where $\widehat{\mathbb{P}}(A)$ is the proportion of times A occurred in the simulation and similarly for $\widehat{\mathbb{P}}(AB)$ and $\widehat{\mathbb{P}}(B)$. Now find two events A and B that are not independent. Compute $\widehat{\mathbb{P}}(A), \widehat{\mathbb{P}}(B)$ and $\widehat{\mathbb{P}}(AB)$. Compare the calculated values to their theoretical values. Report your results and interpret.

```
def dice():
    return random.randint(1,6)

def rolls(n):
    sample = []
    for i in range(1,n+1):
        sample.append(dice())
    return sample

n = 10000

# P(A)
prob_A = (rolls(n).count(2) + rolls(n).count(4) + rolls(n).count(6))/n

# P(B)
prob_B = (rolls(n).count(1) + rolls(n).count(2) + rolls(n).count(3)
         + rolls(n).count(4))/n

# P(AB)
prob_AB = (rolls(n).count(2) + rolls(n).count(4))/n

print("P(A)P(B) =", prob_A*prob_B, "and P(AB) =", prob_AB,
        ". Thus A and B are independent.")
```

Now suppose we have $A = B = \{k\}$ where k is any integer $1, \ldots, 6$. Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(AB) = 1/6$, but $\mathbb{P}(A)\mathbb{P}(B) = 1/36$. Thus A and B are not independent. In general, we can conclude that any event cannot be independent of itself unless the probability of the event is either 0 or 1.

```
def dice():
    return random.randint(1,6)

def rolls(n):
    sample = []
    for i in range(1,n+1):
        sample.append(dice())
    return sample

n = 10000

# P(A)
prob_A = rolls(n).count(1)/n

# P(B)
prob_B = rolls(n).count(1)/n

# P(AB)
prob_AB = rolls(n).count(1)/n

print("P(A)P(B) =", prob_A*prob_B, "and P(AB) =", prob_AB,
        ". Thus A and B are not independent.")
```

Chapter 2: Random Variables

Import Packages

Please see the associated GitHub repo for all code and comments to simulations. [LINK HERE]

```
import numpy as np
from numpy.random import choice
import random
import matplotlib.pyplot as plt
import scienceplots
```

Question: 2.14.1

Show that

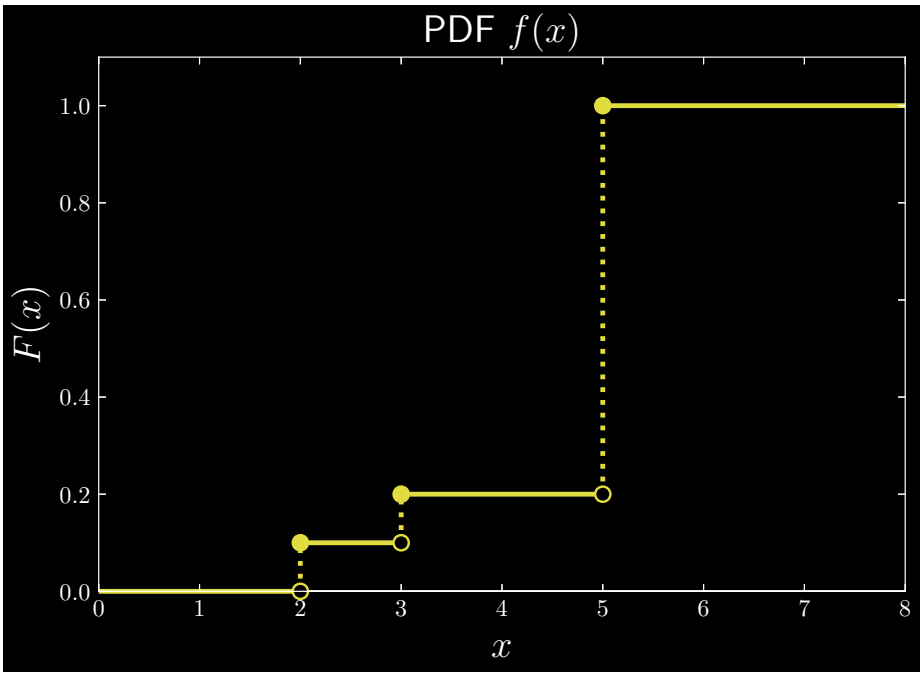
$$\mathbb{P}(X = x) = F(x^+) - F(x^-).$$

Proof. By Lemma 2.15(1), $\mathbb{P}(X = x) = F(x) - F(x^-)$ where $F(x^-) = \lim_{y \uparrow x} F(y)$. By Theorem 2.8(iii), F is right-continuous, or $F(x) = F(x^+)$, where $F(x^+) = \lim_{\substack{y \to x \\ y > x}} F(y)$. Thus $\mathbb{P}(X = x) = F(x^+) - F(x^-)$. □

Question: 2.14.2

Let X be such that $\mathbb{P}(X = 2) = \mathbb{P}(X = 3) = 1/10$ and $\mathbb{P}(X = 5) = 8/10$. Plot the CDF F. Use F to find $\mathbb{P}(2 < X \leq 4.8)$ and $\mathbb{P}(2 \leq X \leq 4.8)$.

The CDF is given by



And the respective probabilities are

$$\begin{aligned}\mathbb{P}(2 < X \leq 4.8) &= \mathbb{P}(X \leq 4.8) - \mathbb{P}(X \leq 2)\\ &= 2/10 - 1/10\\ &= \boxed{1/10}\\ \mathbb{P}(2 \leq X \leq 4.8) &= \mathbb{P}(X \leq 4.8) - \mathbb{P}(X < 2)\\ &= 2/10 - 0\\ &= \boxed{1/5}\end{aligned}$$

Question: 2.14.3

Prove Lemma 2.15.

Lemma (Wasserman 2.15)

Let F be the CDF for a random variable X. Then:

1. $\mathbb{P}(X = x) = F(x) - F(x^-)$ where $F(x^-) = \lim_{y \uparrow x} F(y)$;
2. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$;
3. $\mathbb{P}(X > x) = 1 - F(x)$;
4. If X is continuous then

$$\begin{aligned}F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b)\\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b).\end{aligned}$$

Proof. 1. Let $F(x^-) = \lim_{y \uparrow x} F(y)$. Then let $y_1, y_2, \ldots$ be a sequence such that $y_1 < y_2 < \cdots$ and $\lim_i y_i = x$. Then let $A_i = (-\infty, y_i]$ and $A = (-\infty, x]$, so $A = \bigcup_{i=1}^{\infty} A_i$ and $A_1 \subset A_2 \subset \cdots$. Thus by continuity of probability, $\lim_i \mathbb{P}(A_i) = \mathbb{P}(\bigcup_i A_i) = \mathbb{P}(A) = F(x)$. But $\lim_i \mathbb{P}(A_i) = F(x^-)$ by definition, so we have $F(x) = F(x^-)$, implying $\mathbb{P}(X = x) = F(x) - F(x^-) = 0$.

2. We have $\mathbb{P}(X \leq y) = F(y)$ and $\mathbb{P}(X \leq x) = F(x)$. Using mutual exclusivity of the events $\{ X \colon X \leq x \}$ and $\{ X \colon x < X \leq y \}$, we can conclude that $\mathbb{P}(X \leq x) + \mathbb{P}(x < X \leq y) = \mathbb{P}(X \leq y)$ implies $\mathbb{P}(x < X \leq y) = F(y) - F(x)$.

3. By mutual exclusivity of the events $\{ X \colon X > x \}$ and $\mathbb{P}(X \leq x)$, we must have $\mathbb{P}(X \leq x) + \mathbb{P}(X > x) = 1$, which implies $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F(x)$.

4. Assume X is continuous. Then $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$. Since the events $\{X = a\}, \{ X \colon a < X < b \}, \{X = b\}$ are all disjoint, the desired equalities can be derived from summing the probabilities of the corresponding mutually exclusive events. □

Question: 2.14.4

Let X have the probability density function

$$f_X(x) = \begin{cases} 1/4 & 0 < x < 1 \\ 3/8 & 3 < x < 5 \\ 0 & \text{otherwise.} \end{cases}$$

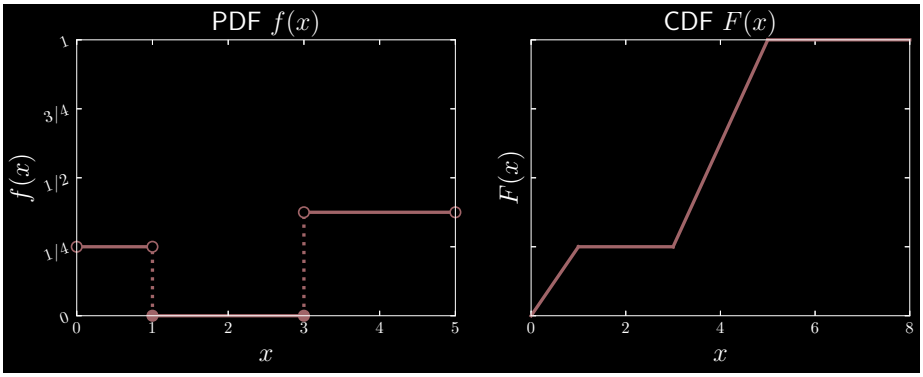
(a) Find the cumulative distribution function of X.

(b) Let $Y = 1/X$. Find the probability density function $f_Y(y)$ for Y. Hint: Consider three cases $\frac{1}{5} \leq y \leq \frac{1}{3}$, $\frac{1}{3} \leq y \leq 1$, and $y \geq 1$.

(a) To find the CDF, we find how the area under the PDF changes as x increases. The direct approach is to integrate each segment of the piecewise PDF, with some nuance – as x moves between 1 and 3, since the PDF is zero here, there is no contribution to the CDF, and so $F(x)$ stays at $1/4$ on this segment. Additionally, once we reach the $(3, 5)$ segment, we not only integrate $3/8$ over the aforementioned interval, but add $1/4$ to reflect the *cumulative* probability. Proceeding in this manner, the CDF is

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ x/4 & 0 < x < 1 \\ 1/4 & 1 \leq x \leq 3 \\ 1/4 + 3(x-3)/8 & 3 < x < 5 \\ 1 & 5 \leq x \end{cases}$$

Graphing the PDF and CDF:



(b) To derive $F_Y(y)$, we begin with

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(1/X \leq y) \\ &= \mathbb{P}(1/y \leq X) \\ &= 1 - \mathbb{P}(X < 1/y) \end{aligned}$$

Since $f_X(x)$ is given as a piecewise function, we must calculate the segments of the CDF $F_Y(y)$, and consequently the PDF $f_Y(y)$, separately.

Note also that

$$\frac{d}{dy}F_Y(y) = f_Y(y)$$

Case 1: $1/5 < y < 1/3$ corresponding to $3 < x < 5$

$$\begin{aligned} F_Y(y) &= 1 - \mathbb{P}(X < 1/y) \\ &= 1 - \left(\frac{1}{4} + \frac{3(1/y - 3)}{8}\right) \\ &= \frac{15}{8} - \frac{3}{8y} \\ \implies f_Y(y) &= \boxed{\frac{3}{8y^2}} \end{aligned}$$

Case 2: $1/3 \leq y \leq 1$ corresponding to $1 \leq x \leq 3$

$$\begin{aligned} F_Y(y) &= 1 - \mathbb{P}(X < 1/y) \\ &= 1 - 1/4 \\ &= 3/4 \\ \implies f_Y(y) &= \boxed{0} \end{aligned}$$

Intuitively, since $f_X(x) = 0$ here, $f_Y(y) = \boxed{0}$.

Case 3: $y > 1$ corresponding to $0 < x < 1$

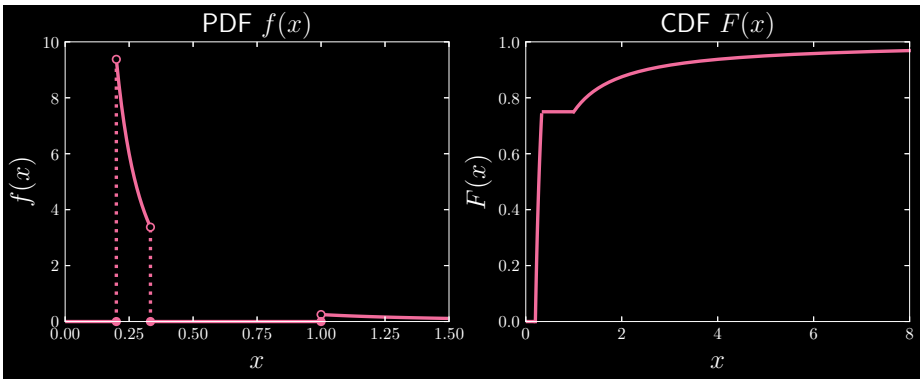
$$\begin{aligned} F_Y(y) &= 1 - \mathbb{P}(X < 1/y) \\ &= 1 - \frac{1}{4y} \\ \implies f_Y(y) &= \boxed{\frac{1}{4y^2}} \end{aligned}$$

Thus the CDF and PDF of Y is given by

$$F_Y(y) = \begin{cases} 0 & y \leq 1/5 \\ 15/8 - 3/8y & 1/5 < y < 1/3 \\ 3/4 & 1/3 \leq y \leq 1 \\ 1 - 1/4y & y > 1 \end{cases}$$

$$f_Y(y) = \begin{cases} 3/8y^2 & 1/5 < y < 1/3 \\ 1/4y^2 & y > 1 \\ 0 & \text{otherwise} \end{cases}$$

Graphing the PDF and CDF:



Question: 2.14.5

Let X and Y be discrete random variables. Show that X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x and y.

PROOF. ($\implies$) Let $X \sqcup Y$. Then for all x, y, $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$. Since $f_X(x) = \mathbb{P}(X = x), f_Y(y) = \mathbb{P}(Y = y)$, and $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$, we have

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

($\impliedby$) Let $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y. Since $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$, $f_X(x) = \mathbb{P}(X = x)$ and $f_Y(y) = \mathbb{P}(Y = y)$, we must have

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all x, y. Thus $X \sqcup Y$. □

Question: 2.14.6

Let X have distribution F and density function f and let A be a subset of the real line. Let $I_A(x)$ be the indicator function for A:

$$I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}$$

Let $Y = I_A(X)$. Find an expression for the cumulative distribution of Y. (Hint: first find the probability mass function for Y.)

Let $Y = I_A(x)$. Then

$$f_Y(y) = \begin{cases} \mathbb{P}(Y = 1) = \int_A f\ dx & x \in A \\ \mathbb{P}(Y = 0) = 1 - \int_A f\ dx & x \notin A \\ 0 & \text{otherwise} \end{cases}$$

which implies

$$F_Y(y) = \begin{cases} 0 & y < 0 \\ 1 - \int_A f\ dx & 0 \le y < 1 \\ 1 & y \ge 1 \end{cases}$$

Question: 2.14.7

Let X and Y be independent and suppose that each has a Uniform$(0,1)$ distribution. Let $Z = \min\{X, Y\}$. Find the density $f_Z(z)$ for Z. Hint: It might be easier to first find $\mathbb{P}(Z > z)$.

One can follow the hint and observe that $\mathbb{P}(Z > z)$ if and only if

$$\mathbb{P}(X > z)\mathbb{P}(Y > z) = (1 - \mathbb{P}(X < z))(1 - \mathbb{P}(Y < z)) = (1 - F_X(z))(1 - F_Y(z))$$

and applying the fact that $F_X(x), F_Y(y)$ both follow a standard uniform distribution. We will also attempt to derive the density $f_Z(z)$ by modeling the mutual exclusivity of either X or Y being the minimum.

We define Z as

$$Z = \min\{X, Y\} = \begin{cases} X & x < y \\ Y & y < x \end{cases}$$

Since X, Y have standard uniform distribution, it must also be the case that $0 < z < 1$ has non-zero probability, and zero probability otherwise.

Our task is to derive $\mathbb{P}(0 < Z < z)$. Either $X \leq Y$ or $Y \leq X$, and in those cases, we must also have $X \leq z$ and $Y \leq z$, respectively. Thus we derive $F_Z(z)$ on this interval:

$$\begin{aligned} \mathbb{P}(0 < Z \leq z) &= \mathbb{P}(X \leq z)\mathbb{P}(X \leq Y) + \mathbb{P}(Y \leq z)\mathbb{P}(Y \leq X) \\ &= \int_0^z \int_x^1 dy\, dx + \int_0^z \int_y^1 dx\, dy \\ &= \boxed{2z - z^2} \end{aligned}$$

Similarly, we look at $\mathbb{P}(z < Z < 1)$, which corresponds to $1 - F_Z(z)$. As in the previous case, either $X \leq Y$ or $Y \leq X$, but this time these cases correspond to $X > z$ and $Y > z$. We derive

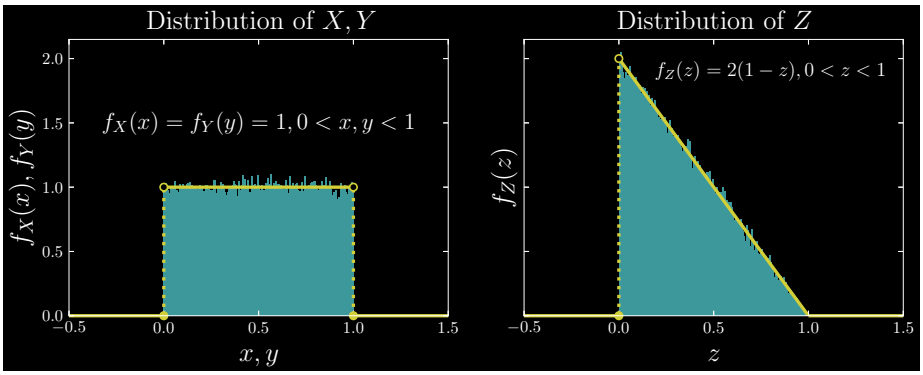
$$\begin{aligned} \mathbb{P}(z < Z < 1) &= \mathbb{P}(X > z)\mathbb{P}(X \leq Y) + \mathbb{P}(Y > z)\mathbb{P}(Y \leq X) \\ &= \int_z^1 \int_x^1 dy\, dx + \int_z^1 \int_y^1 dx\, dy \\ &= \boxed{1 - 2z + z^2} \end{aligned}$$

Either way, we can conclude

$$F_Z(z) = \begin{cases} 0 & z \leq 0 \\ 2z - z^2 & 0 < z < 1 \\ 1 & z \geq 1 \end{cases}$$

$$f_Z(z) = \begin{cases} 2(1 - z) & 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$$

And as anticipated, our theoretical work aligns with the empirical:



Question: 2.14.8

Let X have CDF F. Find the CDF of $X^+ = \max\{0, X\}$.

Observe that $\mathbb{P}(X^+ = 0) = \mathbb{P}(X \leq 0) = F(0)$. For $X > 0$, we have $\mathbb{P}(X^+ = X) = \mathbb{P}(0 < X \leq x) = F(x) - F(0)$. Then the CDF of X^+ is

$$F_{X^+}(X^+) = \begin{cases} F_X(0) & X \leq 0, X^+ = 0 \\ F_X(x) & X > 0, X^+ = X > 0 \end{cases}$$

Question: 2.14.9

Let $X \sim \text{Exp}(\beta)$. Find $F(x)$ and $F^{-1}(q)$.

$$F(x) = \int_0^x \frac{1}{\beta} \exp(-x/\beta)\, dx = 1 - \exp(-x/\beta), \quad x > 0$$

Let $q = F(x)$. Then $q = 1 - \exp(-x/\beta)$ implies

$$\implies \exp(-x/\beta) = 1 - q$$
$$\implies \boxed{x = -\beta \ln(1-q)}, \quad 0 \leq q < 1$$

Question: 2.14.10

Let X and Y be independent. Show that $g(X)$ is independent of $h(Y)$ where g and h are functions.

Proof. Let $X \sqcup Y$. Consider the joint probability $\mathbb{P}(X, Y)$. By premise, $\mathbb{P}(X, Y) = \mathbb{P}(X)\mathbb{P}(Y)$ for all x, y. Now consider $g(X), h(Y)$. Let

$$A = \{\, X : g(X) \in A' \,\}$$
$$B = \{\, Y : h(Y) \in B' \,\}$$

where A', B' are the sets containing the respective transformations of $X \in A, Y \in B$. The intuition to have here is that the "link" between $X \in A$ and $g(X) \in A'$ preserves the independence of the former in the latter under transformation.

Then we have

$$\begin{aligned} \mathbb{P}(g(X) \in A') &= \mathbb{P}(X \in A) \\ \mathbb{P}(h(Y) \in B') &= \mathbb{P}(Y \in B) \\ \implies \mathbb{P}(g(X) \in A', h(Y) \in B') &= \mathbb{P}(X \in A, Y \in B) \end{aligned}$$

Applying independence on the last equality,

$$\mathbb{P}(g(X \in A'), h(y \in B')) = \mathbb{P}(g(X \in A'))\mathbb{P}(h(Y \in B'))$$

ascertaining that $g(X) \sqcup h(X)$. □

Question: 2.14.11

Suppose we toss a coin once and let p be the probability of heads. Let X denote the number of heads and let Y denote the number of tails.

(a) Prove that X and Y are dependent.

(b) Let $N \sim \text{Poisson}(\lambda)$ and suppose we toss a coin N times. Let X and Y be the number of heads and tails. Show that X and Y are independent.

Proof. (a) We have $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(Y = 1) = 1 - p$, where X, Y are the number of heads and tails. Note that $\mathbb{P}(X = 1, Y = 1) = 0$, as we can only have one head or one tail after a single toss. But clearly $\mathbb{P}(X = 1)\mathbb{P}(Y = 1) = p(1 - p) = 0$ for non-zero p, so X, Y are dependent.

(b) The intuition is to decompose the joint probability $f_{X,Y}(x, y)$ into $g(x)h(y)$, where $g(x), h(y)$ are *any* functions (not necessarily densities). We can equivalently find the joint probability that we toss $X = x$ heads *and* toss the coin N times:

$$\begin{aligned} \mathbb{P}(X = x, Y = y) &= \mathbb{P}(X = x, N = x + y) \\ &= \binom{N}{x} p^x (1 - p)^{N-x} e^{-\lambda} \frac{\lambda^N}{N!} \\ &= \binom{x + y}{x} p^x (1 - p)^y e^{-\lambda} \frac{\lambda^{x+y}}{(x + y)!} \\ &= \frac{(x + y)!}{x!y!} p^x (1 - p)^y e^{-\lambda} \frac{\lambda^{x+y}}{(x + y)!} \\ &= \underbrace{\left(e^{-\lambda} \frac{\lambda^x}{x!} p^x \right)}_{g(x)} \underbrace{\left(\frac{\lambda^y}{y!} (1 - p)^y \right)}_{h(y)} \end{aligned}$$

□

Question: 2.14.12

Prove Theorem 2.33.

Theorem (Wasserman 2.33)

Suppose that the range of X and Y is a (possibly infinite) rectangle. If $f(x,y) = g(x)h(y)$ for some functions g and h (not necessarily probability density functions) then X and Y are independent.

PROOF. Let X, Y have density $f(x,y)$, and define

$$A = \{\,(x,y) \colon x > 0, y > 0\,\}$$

Then

$$\begin{aligned} 1 &= \int\int_A f(x,y)\; dx\; dy \\ &= \underbrace{\int_0^\infty g(x)\; dx}_{c} \underbrace{\int_0^\infty h(y)\; dy}_{1/c} \end{aligned}$$

where $0 < c < 1$. Let $f_X(x) = \frac{1}{c} g(x), x > 0$ and $f_Y(y) = ch(y), y > 0$. Then

$$f(x,y) = g(x)h(x) = \left(\frac{1}{c} g(x)\right)(ch(y)) = f_X(x) f_Y(y)$$

for all x, y. Thus $X \sqcup Y$. □

Question: 2.14.13

Let $X \sim N(0,1$ and let $Y = e^X$.

(a) Find the PDF for Y. Plot it.

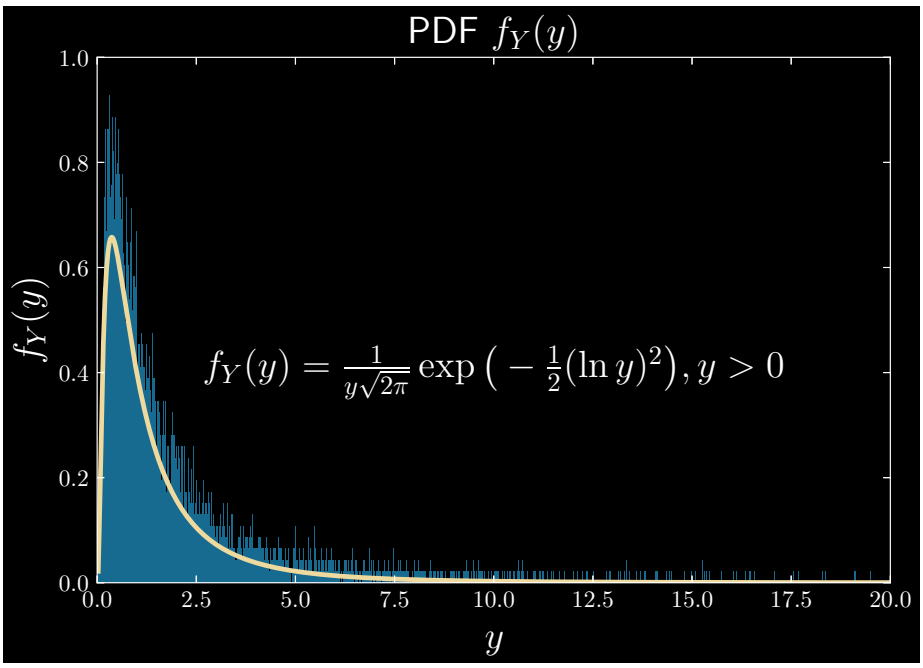
(b) (Computer Experiment.) Generate a vector $x = (x_1, \ldots, x_{10,000})$ consisting of $10,000$ random standard Normals. Let $y = (y_1, \ldots, y_{10,000})$ where $y_i = e^{x_i}$. Draw a histogram of y and compare it to the PDF you found in part (a).

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) \\ &= \mathbb{P}(e^X \le y) \\ &= \mathbb{P}(X \le \ln y) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\ln y} \exp\left(-\frac{x^2}{2}\right)\; dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(-\frac{x^2}{2}\right)\; dx + \frac{1}{2\pi} \int_0^{\ln y} \exp\left(-\frac{x^2}{2}\right)\; dx \end{aligned}$$

Applying the Leibniz rule, we find

$$F'_Y(y) = f_Y(y) = \frac{1}{y\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\ln y)^2\right), \quad y > 0$$

(b) Our theoretical and empirical results agree:



Question: 2.14.14

Let (X, Y) be uniformly distributed on the unit disk $\{(x, y) : x^2+y^2 \leq 1\}$. Let $R = \sqrt{X^2 + Y^2}$. Find the CDF and PDF of R.

For a joint density involving uniformly distributed variables, $f(x, y) = 1/\text{area}(A)$, where A is the region over which the densities of the constituent variables are defined and non-zero. Since we choose uniformly distributed (x, y) over the unit circle, its area is $A = \pi$.

Integrating in polar coordinates, we derive for $0 < r < 1$:

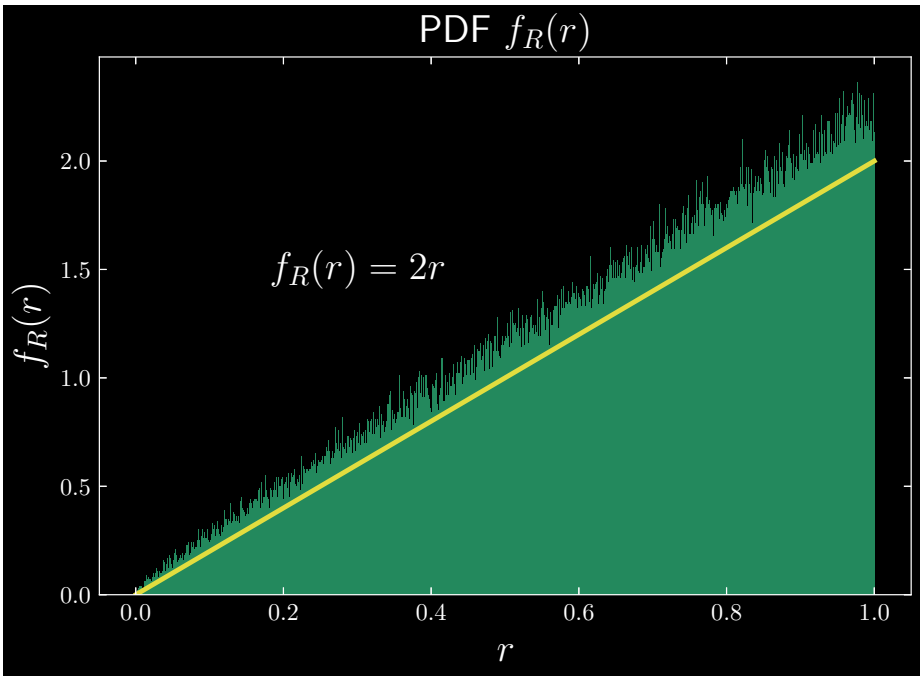
$$F_R(r) = \int_0^{2\pi} \int_0^r \frac{1}{\pi} s \; ds \; d\theta = \int_0^{2\pi} \frac{1}{\pi} \frac{r^2}{2} \; d\theta = r^2$$

giving us the CDF and PDF:

$$F_R(r) = \begin{cases} 0 & r \leq 0 \\ r^2 & 0 < r < \\ 1 & r \geq 1 \end{cases}$$

$$F'_R(r) = f_R(r) = 2r, \quad 0 < r < 1$$

Empirically generating the (x, y)'s agrees with the theoretically derived density:



Question: 2.14.15

(A universal random number generator.) Let X have a continuous, strictly increasing CDF F. Let $Y = F(X)$. Find the density of Y. This is called the probability integral transform. Now let $U \sim \text{Uniform}(0, 1)$ and let $X = F^{-1}(U)$. Show that $X \sim F$. Now write a program that takes $\text{Uniform}(0, 1)$ random variables and generates random variables from an $\text{Exp}(\beta)$ distribution.

PROOF. By assumption, $F_X(x)$ is strictly increasing and continuous. This means $F_X(x)$ is a bijection, and has an inverse $F_X^{-1}(x)$. We use this fact to derive

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(F_X(x) \leq y) \\ &= F_X(F_X^{-1}(y)) \\ &= y \\ \implies F_Y^{-1}(y) = f_Y(y) &= 1 \end{aligned}$$

A shocking result! All probability integral transforms of *any* continuous distribution give us the density to the standard uniform distribution.

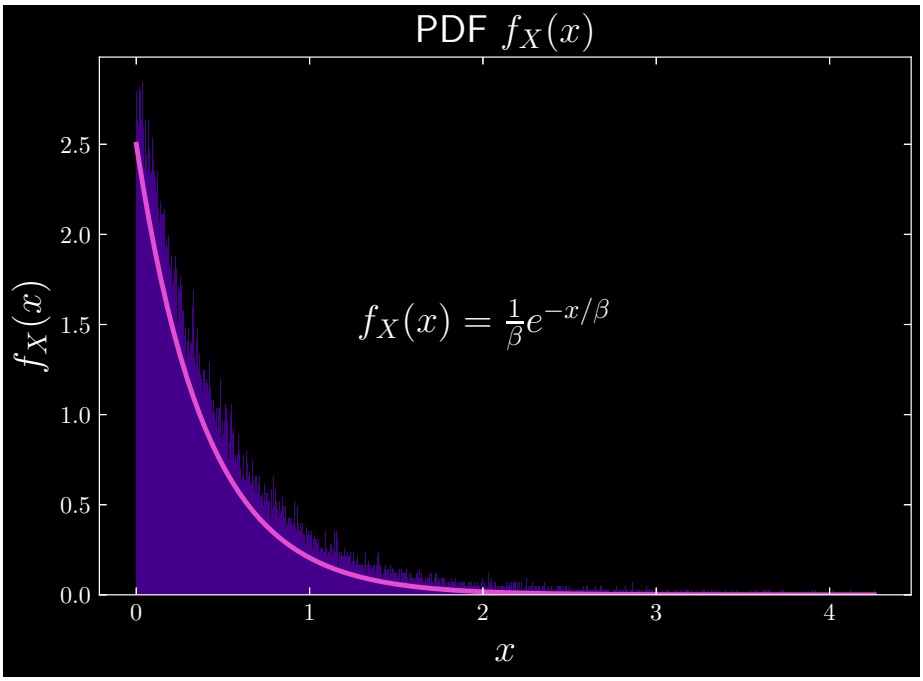
For the next result, let $U \sim \text{Uniform}(0, 1)$ and $X = F^{-1}(U)$. The intuition to have here is that F^{-1} is a quantile function! Again appealing to the bijectivity of F, we have $F(X) = U$. Since $F : X \to [0, 1]$ and invoking the preservation of

monotonocity under an inverse operation ($x_1 < x_2 \iff F(x_1) < F(x_2)$), we have

$$\mathbb{P}(X \leq x) = \mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x)$$

Thus we must have $X \sim F$. □

The quantile function for the Exp(β) distribution is $F^{-1}(q) = -\beta \ln(1-q)$. Generating the exponential distribution via the probability integral transform:



Question: 2.14.16

Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ and assume that X and Y are independent. Show that the distribution of X given that $X + Y = n$ is Binomial(n, π) where $\pi = \lambda/(\lambda + \mu)$.
Hint 1: You may use the following fact: If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, and X and Y are independent, then $X + Y \sim \text{Poisson}(\mu + \lambda)$.
Hint 2: Note that $\{X = x, X + Y = n\} = \{X = x, Y = n - x\}$.

Proof. By definition of the Poisson distribution:

$$f(X + Y = n) = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^n}{n!}$$
$$f(X = x, Y = n - x) = e^{-\lambda} \frac{\lambda^x}{x!} e^{-\mu} \frac{\mu^{n-x}}{(n-x)!}$$

Proceeding to set up the conditional probability:

$$\begin{aligned} f(X \mid X+Y=n) &= \frac{e^{-\lambda}\frac{\lambda^x}{x!}e^{-\mu}\frac{\mu^{n-x}}{(n-x)!}}{e^{-(\lambda+\mu)}\frac{(\lambda+\mu)^n}{n!}} \\ &= \frac{\lambda^x\mu^{n-x}}{(\lambda+\mu)^n}\frac{n!}{(n-x)!x!} \\ &= \binom{n}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(\frac{\mu}{\lambda+\mu}\right)^{n-x} \\ &= \binom{n}{x}\left(\frac{\lambda}{\lambda+\mu}\right)^x\left(1-\frac{\lambda}{\lambda+\mu}\right)^{n-x} \end{aligned}$$

ascertaining that $X \mid X+Y=n \sim \text{Binomial}(n,\pi)$ for $\pi = \lambda/(\lambda+\mu)$. □

Question: 2.14.17

Let

$$f_{X,Y}(x,y) = \begin{cases} c(x+y^2) & 0 \le x < 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{P}(X < \frac{1}{2} \mid Y = \frac{1}{2})$.

First we find the marginal density function of y:

$$f_Y(y) = \int_0^1 c(x+y^2)\,dx = c(1/2+y^2)$$

Then the conditional density function is

$$f_{X|Y}(x \mid y) = \frac{x+y^2}{1/2+y^2}$$

Now we calculate the desired probability:

$$\begin{aligned} \mathbb{P}(X < 1/2 \mid Y = 1/2) &= \int_0^{1/2} f_{X|Y}(x \mid 1/2)\,dx \\ &= \int_0^{1/2} \frac{x+1/4}{1/2+1/4}\,dx \\ &= \frac{1/8+1/8}{1/2+1/4} \\ &= \boxed{1/3} \end{aligned}$$

Question: 2.14.18

Let $X \sim N(3, 16)$. Solve the following using the Normal table and using a computer package.

(a) Find $\mathbb{P}(X < 7)$.

(b) Find $\mathbb{P}(X > -2)$.

(c) Find x such that $\mathbb{P}(X > x) = .05$.

(d) Find $\mathbb{P}(0 \leq X < 4)$.

(e) Find x such that $\mathbb{P}(|\ X\ | > |\ x\ |) = .05$.

It is best to use scipy.stats here:

```
from scipy.stats import norm

normal_dist = norm(3, 4)

# (a)
prob_a = normal_dist.cdf(7)
print("The probability for (a) is", prob_a)

# (b)
prob_b = 1 - normal_dist.cdf(-2)
print("The probability for (b) is", prob_b)

# (c)
x_c = normal_dist.ppf(0.95)
print("The value of x for (c) is", x_c)

# (d)
prob_d = normal_dist.cdf(4) - normal_dist.cdf(0)
print("The probability for (d) is", prob_d)

# (e)
x = np.arange(9.5,9.7,0.00000001)
array = np.round(normal_dist.cdf(x) - normal_dist.cdf(-x), decimals=9)
index = np.where(array == 0.95)
x_e = x[index[0]]
print("Some values of x for (e) are", x_e)
```

Using the above calculations:

(a) $\mathbb{P}(X < 7) = 0.841$

(b) $\mathbb{P}(X > -2) = 0/894$

(c) $x = 9.579$

(d) $\mathbb{P}(0 \leq X < 4) = 0.372$

(e) $x = 9.611$

For (e), the intuition behind the absolute values is to evaluate the probabilities at the tails of the distribution. Given $\mid x \mid > 0$, we can equivalently write

$$\begin{aligned}\mathbb{P}(\mid X \mid > \mid x \mid) &= \mathbb{P}(X > \mid x \mid) + \mathbb{P}(X < - \mid x \mid) \\ &= 1 - \mathbb{P}(X < \mid x \mid) + \mathbb{P}(X < - \mid x \mid) = 0.05 \\ \implies 0.95 &= \mathbb{P}(X < \mid x \mid) - \mathbb{P}(X < - \mid x \mid)\end{aligned}$$

Numerically iterating enables us to identify the value of $\mid x \mid = x$ without needing to convert to a standard normal distribution.

Question: 2.14.19

Prove formula (2.12).

Property (Wasserman Formula 2.12)

For a strictly monotone increasing or strictly monotone decreasing function $r(X)$ with an inverse $s = r^{-1}$, we have

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|.$$

Proof. Case 1: Increasing $r(X)$. Derive

$$\begin{aligned}F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(r(X) \leq y) \\ &= \mathbb{P}(X \leq r^{-1}(y)) \\ &= \mathbb{P}(X \leq s(y)) \\ &= \int_{-\infty}^{s(y)} f_X(x)\, dx\end{aligned}$$

Implying $F'_Y(y) = f_Y(y) = f_X(s(y)) \frac{ds(y)}{dy}$. Note that since $r(X)$ is strictly monotone increasing by premise, $ds(y)/dy > 0$.

Case 2: Decreasing $r(X)$. Derive

$$\begin{aligned}F_Y(y) = \mathbb{P}(Y \leq y) &= \mathbb{P}(r(X) \leq y) \\ &= \mathbb{P}(X \geq r^{-1}(y)) \\ &= \mathbb{P}(X \geq s(y)) \\ &= 1 - \mathbb{P}(X \leq s(y)) \\ &= 1 - \int_{-\infty}^{s(y)} f_X(x)\, dx \\ F'_Y(y) = f_Y(y) &= -f_X(s(y)) \frac{ds(y)}{dy}\end{aligned}$$

where the second equality follows because $r(X)$ is strictly monotonically *decreasing*. That same fact also allows us to say that $ds(y)/dy < 0$. To encapsulate both the increasing and decreasing cases, we have

$$f_Y(y) = f_X(s(y))\left|\frac{ds(y)}{dy}\right|$$

□

Question: 2.14.20

Let $X, Y \sim \text{Uniform}(0,1)$ be independent. Find the PDF for $X - Y$ and X/Y.

The joint probability density function is given by

$$f_{X,Y}(x,y) = \begin{cases} 1 & 0 < x, y < 1 \\ 0 & \text{otherwise} \end{cases}$$

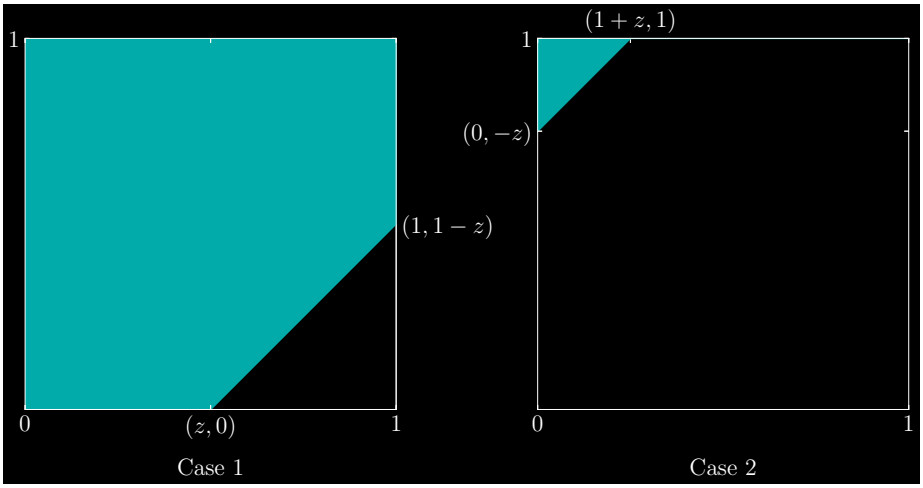
$\underline{X - Y}$

Illustrations are particularly useful here to understand where the piecewise segments of $f_Z(z)$ are defined.

Given $Z = X - Y$, it must be the case that either $0 < z < 1$ or $-1 < z < 0$. To find A_z, the support for the probability density function, we find all x, y satisfying

$$A_z = \{\,(x,y) \colon x - y \le z\,\}$$

Now, there is a nice geometric intuition in that the piecewise CDF is given by the area of the support on each interval $0 < z < 1, -1 < z < 0$. Graphically, this support in the first and second cases is:



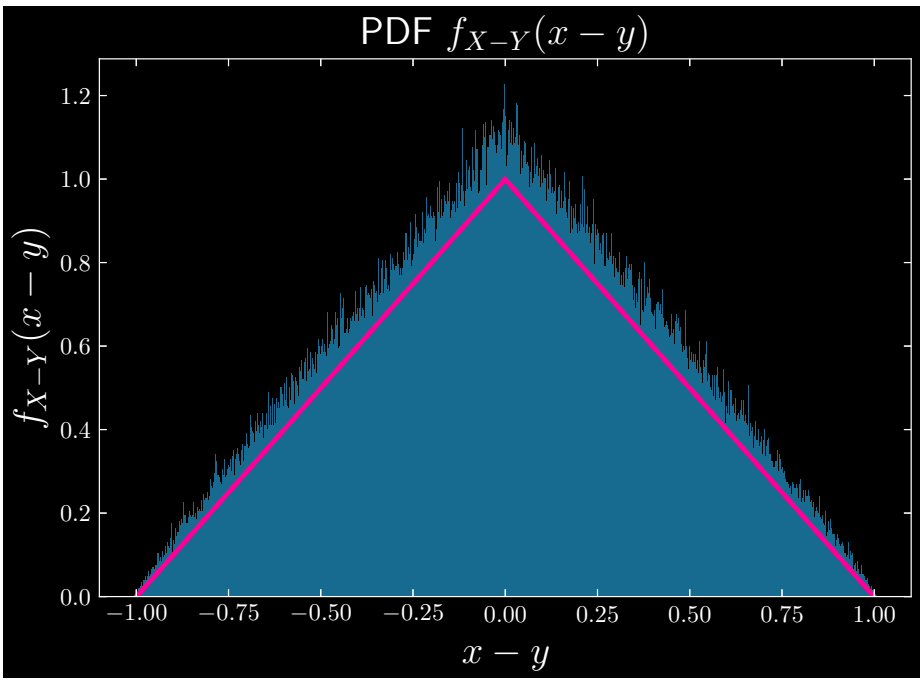
Calculating the area of the shaded parts, we can derive the CDF as

$$F_Z(z) = \begin{cases} 0 & z < -1 \\ \frac{1}{2}(1+z)^2 & -1 < z < 0 \\ 1 - \frac{1}{2}(1-z)^2 & 0 < z < 1 \\ 1 & z > 1 \end{cases}$$

and consequently, the PDF is

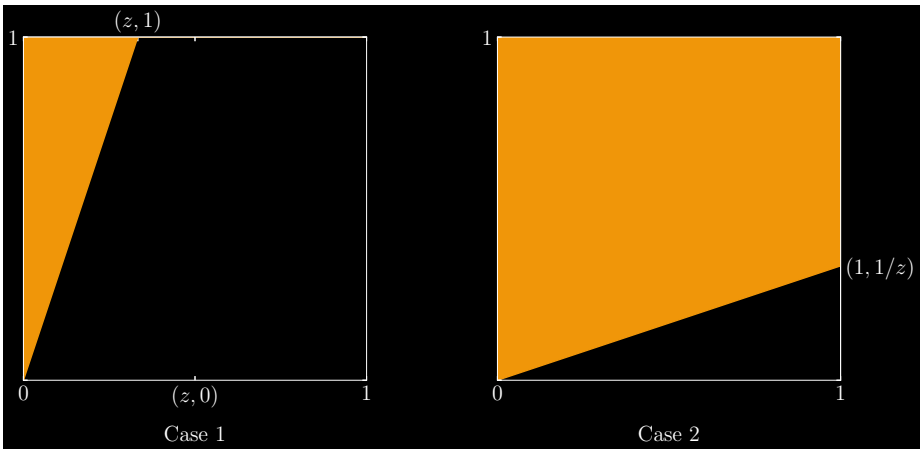
$$f_Z(z) = \begin{cases} 1+z & -1 < z < 0 \\ 1-z & 0 < z < 1 \\ 0 & \text{otherwise} \end{cases}$$

which agrees with our empirical simulation:



X/Y

Once again finding the supports for $Z = X/Y$, observe that the intervals we examine are $0 < z < 1$ and $z > 1$. The areas are graphed as



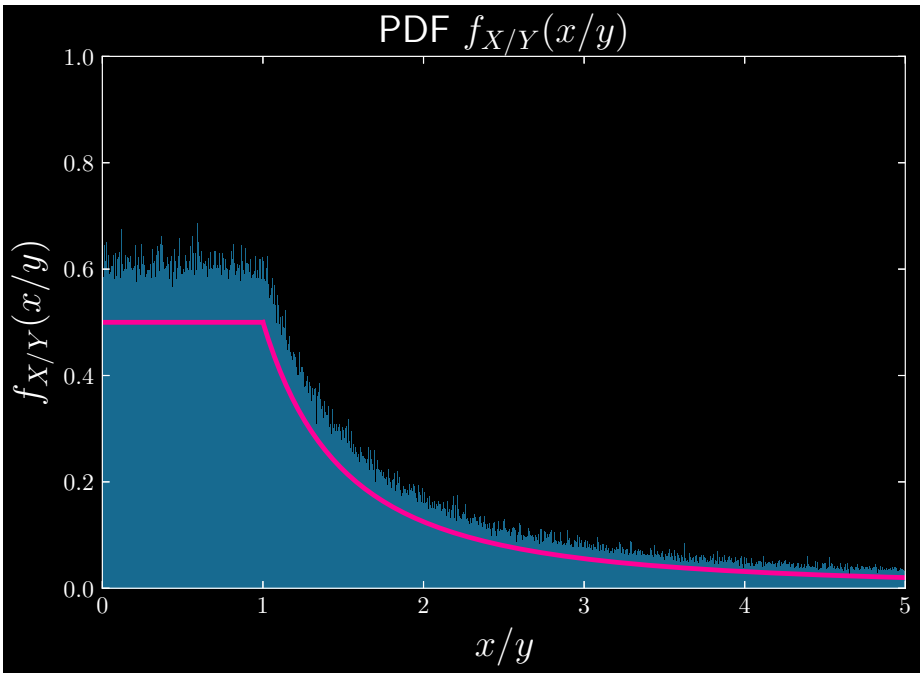
with CDF

$$F_Z(z) = \begin{cases} 0 & z < 0 \\ \frac{z}{2} & 0 < z < 1 \\ 1 - \frac{1}{2z} & z > 1 \end{cases}$$

and following that, the PDF

$$f_Z(z) = \begin{cases} \frac{1}{2} & 0 < z < 1 \\ \frac{1}{2z^2} & z > 1 \\ 0 & \text{otherwise} \end{cases}$$

Once again, the theoretical and the empirical align:



Question: 2.14.21

Let $X_1, \ldots, X_n \sim \text{Exp}(\beta)$ be IID. Let $Y = \max\{X_1, \ldots, X_n\}$. Find the PDF of Y. Hint: $Y \leq y$ if and only if $X_i \leq y$ for $i = 1, \ldots, n$.

Taking advantage of the hint and the IID random variables, observe that $\mathbb{P}(Y \leq y) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq y)$. The derivation is straightforward:

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) &= \prod_{i=1}^{n} \int_0^y \frac{1}{\beta} e^{-x_i/\beta} \, dx_i \\ &= \prod_{i=1}^{n} (1 - e^{-y/\beta}) \\ &= (1 - e^{-y/\beta})^n \\ \implies F_Y'(y) = f_Y(y) = \frac{d}{dy}(1 - e^{-y/\beta})^n &= \boxed{\frac{n}{\beta} e^{-y/\beta} (1 - e^{-y/\beta})^{n-1}} \end{aligned}$$

Chapter 3: Expectation

Import Packages

Please see the associated GitHub repo for all code and comments to simulations. [LINK HERE]

```
import numpy as np
from numpy.random import choice
import random
import statistics
import matplotlib.pyplot as plt
import scienceplots
```

Question: 3.8.1

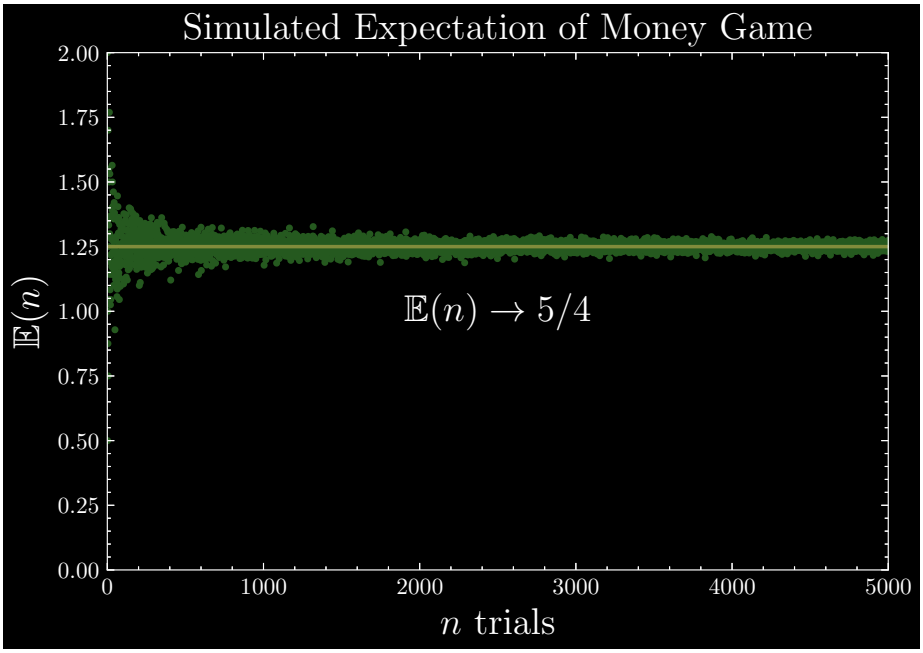
Suppose we play a game where we start with c dollars. On each play of the game you either double or halve your money, with equal probability. What is your expected fortune after n trials?

By premise, $\mathbb{P}(X = 2c) = 1/2, \mathbb{P}(X = \frac{1}{2}c) = 1/2$. By definition of expectation, we have

$$\mathbb{E}(X) = 2c\left(\frac{1}{2}\right) + \frac{1}{2}c\left(\frac{1}{2}\right) = \boxed{\frac{5}{4}c}$$

Simulating the money game, we find:

```
def moneygame(c):
    flip = random.randint(0,1)
    if flip == 0:
        c = 0.5*c
    else:
        c = 2*c
    return c;

c = 1
n = 5000

def runs(n):
    expectations = []
    for j in range(1,n+1):
        trials = []
        for i in range(1,j+1):
            trials.append(moneygame(c))
        mean = sum(trials)/len(trials)
        expectations.append(mean)
    return expectations;
```



Question: 3.8.2

Show that $\mathbb{V}(X) = 0$ if and only if there is a constant c such that $\mathbb{P}(X = c) = 1$.

PROOF. ($\implies$) Let $\mathbb{V}(X) = 0$. Since $\mathbb{V}(X) = \int (x-\mu)^2 \, dF(x)$, we must have $x = \mu$. Then $\mathbb{P}(X = \mu) = 1$.

($\impliedby$) Let $\mathbb{P}(X = c) = 1$. Then $\mathbb{E}(X) = c$ and $\mathbb{E}(X^2) = c^2$. Then $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = c^2 - c^2 = 0$. □

Question: 3.8.3

Let $X_1, \ldots, X_n \sim \text{Uniform}(0,1)$ and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $\mathbb{E}(Y_n)$.

Assume IID standard uniform variables. Note that $Y_n \leq y$ if and only if $X_1, \ldots, X_n \leq y$. We can derive

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}(Y_n \leq y) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \leq y) \\ &= \prod_{i=1}^{n} \int_0^y dx_i \\ &= y^n \end{aligned}$$

Then the PDF is

$$F'_{Y_n}(y) = f_{Y_n}(y) = ny^{n-1}$$

and the expectation is

$$\mathbb{E}(Y_n) = \int_0^1 ny^n \, dy = \boxed{\frac{n}{n+1}}$$

Question: 3.8.4

A particle starts at the origin of the real line and moves along the line in jumps of one unit. For each jump the probability is p that the particle will jump one unit to the left and the probability is $1-p$ that the particle will jump one unit to the right. Let X_n be the position of the particle after n units. Find $\mathbb{E}(X_n)$ and $\mathbb{V}(X_n)$. (This is known as a **random walk**.)

We define X as

$$X = \begin{cases} +1 & 1-p \\ -1 & p \end{cases}$$

After one jump,

$$\begin{aligned} \mathbb{E}(X_i) &= 1(1-p) + (-1)p = 1 - 2p \\ \mathbb{E}(X_i^2) &= 1 - p + p = 1 \\ \mathbb{V}(X_i) &= \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = 1 - (1-2p)^2 = 4(p - p^2) \end{aligned}$$

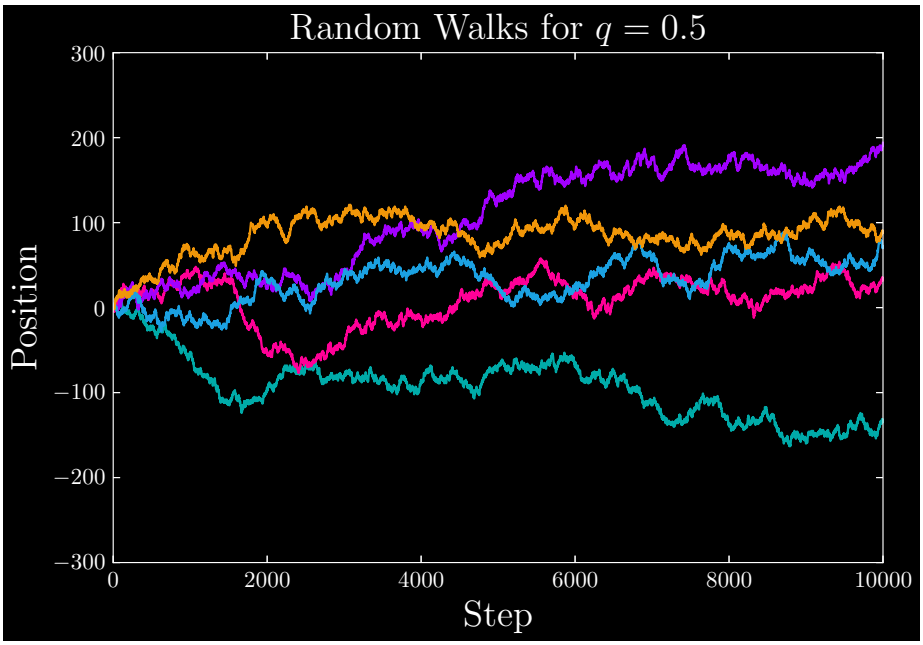
After n jumps, assuming independence of each jump,

$$\begin{aligned} \mathbb{E}\left(\sum^n X_i\right) &= \sum^n \mathbb{E}(X_i) = \boxed{n(1-2p)} \\ \mathbb{V}\left(\sum^n X_i\right) &= \sum^n \mathbb{V}(X_i) = \boxed{4n(p-p^2)} \end{aligned}$$

We can simulate some random walks for $p = 0.5$:

```
q = 0.5

def randomwalk(q):
    step = [-1,1]
    randomStep = choice(
        step, 1, p=[q, 1-q])
    return randomStep.tolist()[0]

def trials(n):
    record = []
    position = 0
    for i in range(1,n+1):
```

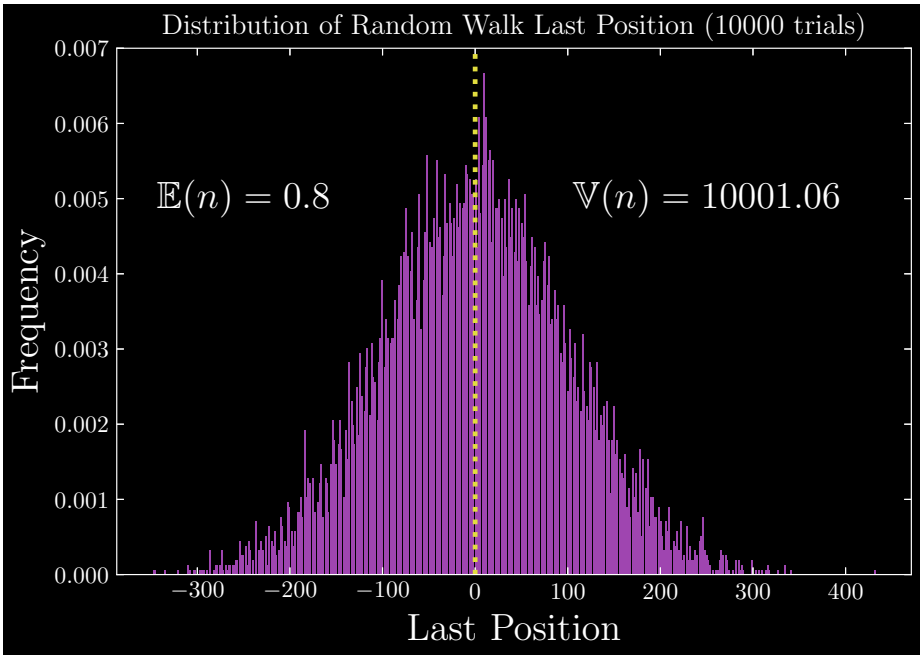
```
        position = position + randomwalk(q)
        record.append(position)
    return position, record;

def multiplewalks(k,n):
    walklist = []
    for i in range(1,k+1):
        trials(n)
        walklist.append(trials(n)[1])
    return walklist;

k = 5
n = 10000
```



Empirically simulating the expectation and variance for n steps also brings us quite close to our derived theoretical values:

```
def trials(n):
    record = []
    position = 0
    for i in range(1,n+1):
        position = position + randomwalk(q)
    return position;

n = 10000
lastposition = []
for i in range(1, n+1):
    lastposition.append(trials(n))
```

```
e = statistics.mean(lastposition)
v = statistics.variance(lastposition)
```



Question: 3.8.5

A fair coin is tossed until a head is obtained. What is the expected number of tosses that will be required?

In general, for a coin in which a head is obtained with probability p, we can model the expected number of tosses until we get a head with the geometric distribution:

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

Deriving the first moment, we first write the MGF (with $q = 1 - p$)

$$\begin{aligned}\psi_X(t) &\sum_x e^{tx} p q^{x-1} \\ &= p(e^t + qe^{2t} + q^2 e^{3t} + \cdots) \\ &= pe^t(1 + qe^t + q^2 e^{2t} + \cdots) \\ &= \frac{pe^t}{1 - qe^t}\end{aligned}$$

Differentiating gives us

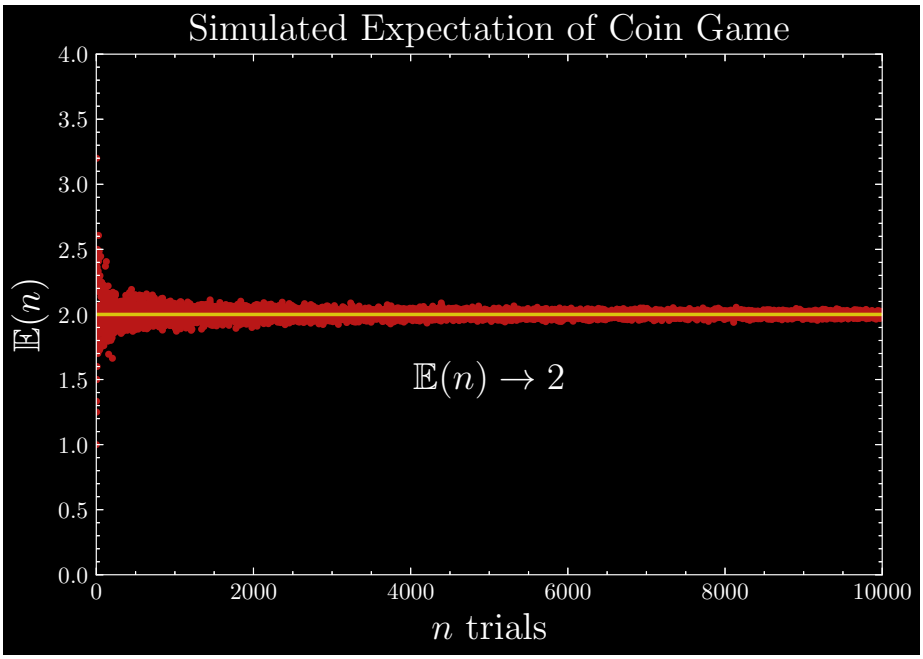
$$\psi'_X(t) = pe^t(1 - qe^t)^{-1} - pe^t(-qe^t)(1 - qe^t)^{-2}$$

And setting $t = 0$ gives us

$$\begin{aligned}&\psi'_X(0) p(1-q)^{-1} - p(-q)(1-q)^{-2} \\ &= \frac{1}{p}\end{aligned}$$

For $p = 1/2$, $\mathbb{E}(X) = \frac{1}{1/2} = \boxed{2}$. Simulating gives us the desired expectation:

```
def cointoss():
    tosses, heads = [], 0
    while (heads < 1):
        flip = random.randint(0,1)
        if (flip == 0):
            tosses.append('Heads')
            heads = tosses.count('Heads')
        else:
            tosses.append('Tails')
    return len(tosses);

def empirical(n):
    numtosses, probs = [], []
    for i in range(0,n):
        numtosses.append(cointoss())
    for x in np.unique(numtosses):
        probs.append((x, numtosses.count(x)))
    return numtosses;

def trials(n):
    trials = []
    for i in range(1,n+1):
        trials.append(statistics.mean(empirical(i)))
    return trials;

n = 10000
```



Question: 3.8.6

Prove Theorem 3.6 for discrete random variables.

PROOF. Let $Y = r(X)$. We make no assumption about whether r is a bijection, namely if it has an inverse. Therefore, it is possible for multiple values of $X = x$ to map to the same $Y = y$. For instance, if $Y = r(X) = X^2$, then $X = 1$ and $X = -1$ map to $Y = 1$, and we have $\mathbb{P}(Y = 1) = \mathbb{P}(X = 1) + \mathbb{P}(X = -1)$.

From the definition of expectation, we have

$$\mathbb{E}(Y) = \sum_i y_i \mathbb{P}(Y = y_i)$$

Now, $\mathbb{P}(Y = y_i)$ is equal to the sum of the probabilities of X taking on all values x_j such that $y_i = g(X = x_j)$. In other words,

$$\mathbb{P}(Y = y_i) = \sum_{j:y_i = g(x_j)} \mathbb{P}(X = x_j) = \mathbb{P}(X = x)$$

Thus we can conclude with

$$\begin{aligned} &= \sum_i y_i \sum_{j:y_i = g(x_j)} \mathbb{P}(X = x_j) \\ &= \sum g(x) \mathbb{P}(X = x) \\ &= \sum g(x) f_X(x) \end{aligned}$$

□

Question: 3.8.7

Let X be a continuous random variable with CDF F. Suppose that $\mathbb{P}(X > 0) = 1$ and that $\mathbb{E}(X)$ exists. Show that $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)\, dx$. Hint: Consider integrating by parts. The following fact is helpful: if $\mathbb{E}(X)$ exists then $\lim_{x\to\infty} x[1 - F(x)] = 0$.

PROOF. By definition of expectation for a continuous random variable,

$$\mathbb{E}(X) = \int_0^\infty x f_X(x)\, dx$$

Proceeding with integration by parts, let $\mu = x, dv = f_X(x)dx$. Then $du =$

$dx, v = F_X(x)$. We continue with

$$\begin{aligned}
&= xF_X(x)\Big|_0^\infty - \int_0^\infty F_X(x)\ dx \\
&= x\Big|_0^\infty - \int_0^\infty F_X(x)\ dx \\
&= \int_0^\infty dx - \int_0^\infty F_X(x)\ dx \\
&= \int_0^\infty [1 - F_X(x)]\ dx \\
&= \int_0^\infty \mathbb{P}(X > x)\ dx
\end{aligned}$$

with the second equality following from the fact that $\lim_{x\to\infty} x[1 - F_X(x)] = 0$, implying that $\lim_{x\to\infty} x = \lim_{x\to\infty} xF_X(x)$. □

Question: 3.8.8

Prove Theorem 3.17.

Theorem (Wasserman 3.17)
Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_i), \sigma^2 = \mathbb{V}(X_i)$. Then

$$\mathbb{E}(\overline{X}_n) = \mu, \quad \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n} \quad \text{and} \quad \mathbb{E}(S_n^2) = \sigma^2.$$

Proof. Use the fact that the X_i's are IID.

$$\begin{aligned}
\mathbb{E}(\overline{X}_n) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=}^{n} X_i\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \frac{1}{n}(n\mu) \\
&= \boxed{\mu} \\
\mathbb{V}(\overline{X}_n) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n}\frac{1}{n^2}\mathbb{V}(X_i) \\
&= \frac{1}{n^2}(n\sigma^2) \\
&= \boxed{\frac{\sigma^2}{n}}
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}(S_n^2) &= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2\right) \\
&= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2)\right) \\
&= \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}X_i^2 - n\overline{X}_n^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}(\overline{X}_n^2)\right) \\
&= \frac{1}{n-1}(n(\sigma^2+\mu^2)) - \frac{1}{n-1}\frac{1}{n}(n(n-1)\mu^2 + n(\sigma^2+\mu^2)) \\
&= \frac{1}{n-1}(n(\sigma^2+\mu^2)) - \frac{1}{n-1}(n\mu^2+\sigma^2) \\
&= \frac{n\sigma^2 - \sigma^2}{n-1} \\
&= \boxed{\sigma^2}
\end{aligned}
$$

□

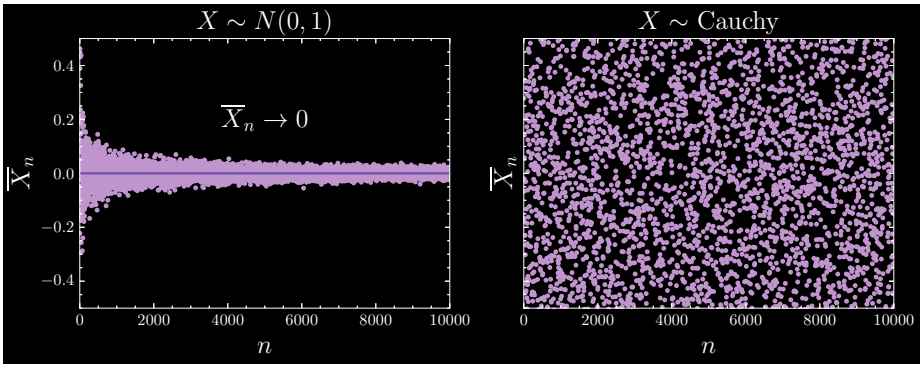
Question: 3.8.9

(Computer Experiment.) Let $X_1, X_2, \ldots, X_n$ be $N(0,1)$ random variables and let $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Plot $\overline{X}_n$ versus n for $n = 1, \ldots, 10,000$. Repeat for $X_1, X_2, \ldots, X_n \sim$ Cauchy. Explain why there is such a difference.

We can code our functions to generate $\overline{X}_n$ for the $N(0,1)$ and Cauchy distributions as follows:

```
def normal(n):
    normalmeans = []
    for i in range(1,n+1):
        normalmeans.append(statistics.mean(np.random.normal(0,1,i)))
    return normalmeans;

def cauchy(n):
    cauchymeans = []
    for i in range(1,n+1):
        cauchymeans.append(
                statistics.mean(np.random.standard_cauchy(i)))
    return cauchymeans;

n = 10000
```

The simulations yield:



A striking comparison! Qualitatively, the standard normal and standard Cauchy distributions look to be quite similar. But notably, the latter *has no mean*. Given the PDF

$$f_Z(z) = \frac{1}{\pi(1+z^2)}, \quad z \in \mathbb{R}$$

which is generated by finding the distribution of $Z = X/Y$ for $X, Y \sim N(0, \sigma^2)$, the expectation $\mathbb{E}(Z)$ turns out to diverge.

Question: 3.8.10

Let $X \sim N(0,1)$ and let $Y = e^X$. Find $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$.

By the Law of the Unconscious Statistician, we have

$$\begin{aligned}\mathbb{E}(Y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^x e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2+x}\, dx\end{aligned}$$

Using the identity

$$\int_{-\infty}^{\infty} e^{-ax^2+bx}\, dx = e^{b^2/4a}\sqrt{\frac{\pi}{a}}$$

with $a = b = 1$, we can conclude

$$\mathbb{E}(Y) = \frac{1}{\sqrt{2\pi}} e^{1/2}\sqrt{2\pi} = \boxed{e^{1/2}}$$

Similarly, we calculate $\mathbb{E}(Y^2)$:

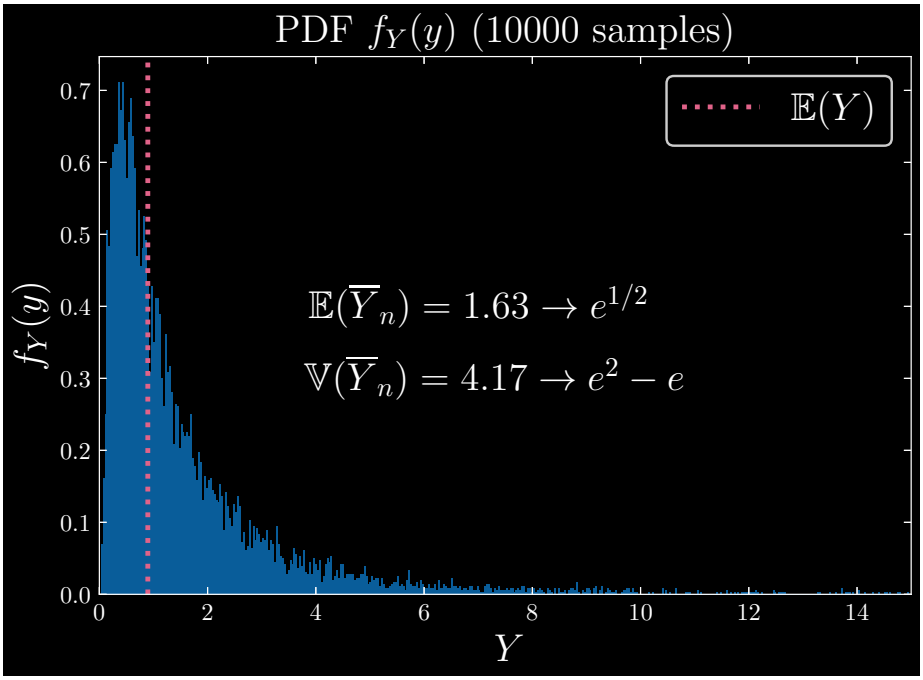
$$\begin{aligned}\mathbb{E}(Y^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{2x} e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2+2x}\, dx \\ &= e\end{aligned}$$

and derive

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \boxed{e^2 - e}$$

Simulating, we get

```
n = 10000

x = np.random.normal(0,1,n) # normal samples
y = np.exp(x) # exponential transformation
meany = round(statistics.mean(y),2)
vary = round(statistics.variance(y),2)
```



Question: 3.8.11

(Computer Experiment: Simulating the Stock Market.) Let $Y_1, Y_2, \ldots$ be independent random variables such that $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i = -1) = 1/2$. Let $X_n = \sum_{i=1}^{n} Y_i$. Think of $Y_i = 1$ as "the stock price increased by one dollar," $Y_i = -1$ as "the stock price decreased by one dollar," and X_n as the value of the stock on day n.

(a) Find $\mathbb{E}(X_n)$ and $\mathbb{V}(X_n)$.

(b) Simulate X_n and plot X_n versus n for $n = 1, 2, \ldots, 10,000$. Repeat the whole simulation several times. Notice two things. First, it's easy to "see" patterns in the sequence even though it is random. Second, you will find that the four runs look very different even though they were generated the same way. How do the calculations in (a) explain the second observation?

(a) The expectation and variance are derived as

$$\begin{aligned}
\mathbb{E}(Y_i) &= 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{2} = 0 \\
\implies \mathbb{E}(X_n) &= \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right) = \boxed{0} \\
\mathbb{E}\left(Y_i^2\right) &= 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1 \\
\mathbb{V}(Y_i) &= \mathbb{E}\left(Y_i^2\right) - \mathbb{E}(Y_i)^2 = 1 \\
\implies \mathbb{V}(X_n) &= \mathbb{V}\left(\sum_{i=1}^{n} Y_i\right) = \boxed{n}
\end{aligned}$$

(b) Refer to exercise 3.8.4, as this exercise is a special case of the random walk where $p = 0.5$. The intuition for why the paths diverge at increasing number of steps n is because of the path-dependency of the movement of the sequence – the position in the random walk or the stock price is contingent on what came before it – and that the variance is dependent on n (in fact it is n itself). For high n, the larger the spread of X_n.

Question: 3.8.12

Prove the formulas given in the table at the beginning of Section 3.4 for the Bernoulli, Poisson, Uniform, Exponential, Gamma, and Beta. Here are some hints. For the mean of the Poisson, use the fact that $e^a = \sum_{x=0}^{\infty} a^x/x!$. To compute the variance, first compute $\mathbb{E}(X(X-1))$. For the mean of the Gamma, it will help to multiply and divide by $\Gamma(\alpha+1)/\beta^{\alpha+1}$ and use the fact that a Gamma density integrates to 1. For the Beta, multiply and divide by $\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)$.

Proof. Bernoulli

$$\begin{aligned}
\mathbb{E}(X) &= 1 \cdot \mathbb{P}(X=1) + 0 \cdot \mathbb{P}(X=0) = \boxed{p} \\
\mathbb{E}(X^2) &= 1^2 \cdot \mathbb{P}(X=1) + 0^2 \cdot \mathbb{P}(X=0) = p \\
\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = \boxed{p(1-p)}
\end{aligned}$$

Poisson Using the infinite sum definition of e^x:

$$e^x = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!}$$

we can derive

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^x}{(x-1)!} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} \\
&= \boxed{\lambda} \\
\mathbb{E}(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1) e^{-\lambda} \frac{\lambda^x}{x!} \\
&= \lambda \sum_{x=1}^{\infty} (x-1) e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda^2
\end{aligned}$$

which implies $\mathbb{E}(X^2) = \lambda^2 + \lambda$, and in turn

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda^2 + \lambda - \lambda^2 = \boxed{\lambda}$$

Uniform

$$\begin{aligned}
\mathbb{E}(X) &= \int_a^b x\left(\frac{1}{b-a}\right)\,dx \\
&= \boxed{\frac{a+b}{2}} \\
\mathbb{E}(X^2) &= \int_a^b x^2\left(\frac{1}{b-a}\right)\,dx \\
&= \frac{a^2+ab+b^2}{3} \\
\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
&= \boxed{\frac{(b-a)^2}{12}}
\end{aligned}$$

Exponential We will have to carry out this integration by parts. Let $\mu = x, dv = \exp(-x/\beta)dx$. Then $du = dx, v = -\beta\exp(-x/\beta)$.

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\beta}\int_0^\infty x\exp(-x/\beta)\,dx \\
&= \frac{1}{\beta}\left(-\beta x\exp(-x/\beta)\Big|_0^\infty + \int_0^\infty \beta\exp(-x/\beta)\,dx\right) \\
&= -\beta\exp(-x/\beta)\Big|_0^\infty \\
&= \boxed{\beta}
\end{aligned}$$

To calculate $\mathbb{E}(X^2)$, let $\mu = x, dv = x\exp(-x/\beta)dx$, then $du = dx, v = -(\beta x + \beta^2)\exp(-x/\beta)$.

$$\begin{aligned}
\mathbb{E}(X^2) &= \frac{1}{\beta}\int_0^\infty x^2\exp(-x/\beta)\,dx \\
&= \frac{1}{\beta}\Bigg(-(\beta x+\beta^2)x\exp(-x/\beta)\Big|_0^\infty \\
&\quad + \beta\int_0^\infty x\exp(-x/\beta)\,dx + \beta^2\int_0^\infty \exp(-x/\beta)\,dx\Bigg) \\
&= \int_0^\infty x\exp(-x/\beta)\,dx + \beta\int_0^\infty \exp(-x/\beta)\,dx \\
&= \beta^2 + \beta^2 \\
&= 2\beta^2
\end{aligned}$$

and now the variance:

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \boxed{\beta^2}$$

Gamma Use the identity

$$\Gamma(\alpha+1) = \alpha \int_0^\infty t^{\alpha-1} e^{-t}\ d\alpha = \alpha\Gamma(\alpha)$$

Then we can calculate

$$\mathbb{E}(X) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^\alpha e^{-x/\beta}\ dx$$

Let $X = \beta y$. Then $dx = \beta dy$. (Use this substitution for $\mathbb{E}(X^2)$ too.)

$$\begin{aligned}
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty (\beta y)^\alpha e^{-y} \beta\ dy \\
&= \beta \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \beta \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \boxed{\alpha\beta} \\
\mathbb{E}(X^2) &= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha+1} e^{-x/\beta}\ dx \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty (\beta y)^{\alpha+1} e^{-y} \beta\ dy \\
&= \beta^2 \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \\
&= \beta^2 \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \alpha(\alpha+1)\beta^2 \\
\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 \\
&= \boxed{\alpha\beta^2}
\end{aligned}$$

Beta

$$\begin{aligned}
\mathbb{E}(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\, dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\, dx \\
&= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+1)} \\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha)\Gamma(\alpha+\beta)} \\
&= \boxed{\frac{\alpha}{\alpha+\beta}} \\
\mathbb{E}(X^2) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\, dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+2)\Gamma(\beta)/\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+2)\Gamma(\beta)/\Gamma(\alpha+\beta+2)} \int_0^1 x^{\alpha+1}(1-x)^{\beta-1}\, dx \\
&= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+2)} \\
&= \frac{\alpha(\alpha+1)\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)(\alpha+\beta+1)\Gamma(\alpha)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \boxed{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}
\end{aligned}$$

□

Question: 3.8.13

Suppose we generate a random variable X in the following way. First we flip a fair coin. If the coin is heads, take X to have a Uniform$(0,1)$ distribution. If the coin is tails, take X to have a Uniform$(3,4)$ distribution.

(a) Find the mean of X.

(b) Find the standard deviation of X.

(a) If heads, then $X \sim \text{Uniform}(0,1)$, and if tails, $X \sim \text{Uniform}(3,4)$. By the definition of expectation:

$$\begin{aligned}\mathbb{E}(X) &= \frac{1}{2}\left(\int_0^1 x\ dx + \int_3^4 x\ dx\right) \\ &= \frac{1}{2}\left(\frac{1}{2} + \frac{7}{2}\right) \\ &= \boxed{2}\end{aligned}$$

(b) By definition, the standard deviation $\text{sd}(X)$ is $\sqrt{\mathbb{V}(X)}$. Calculate

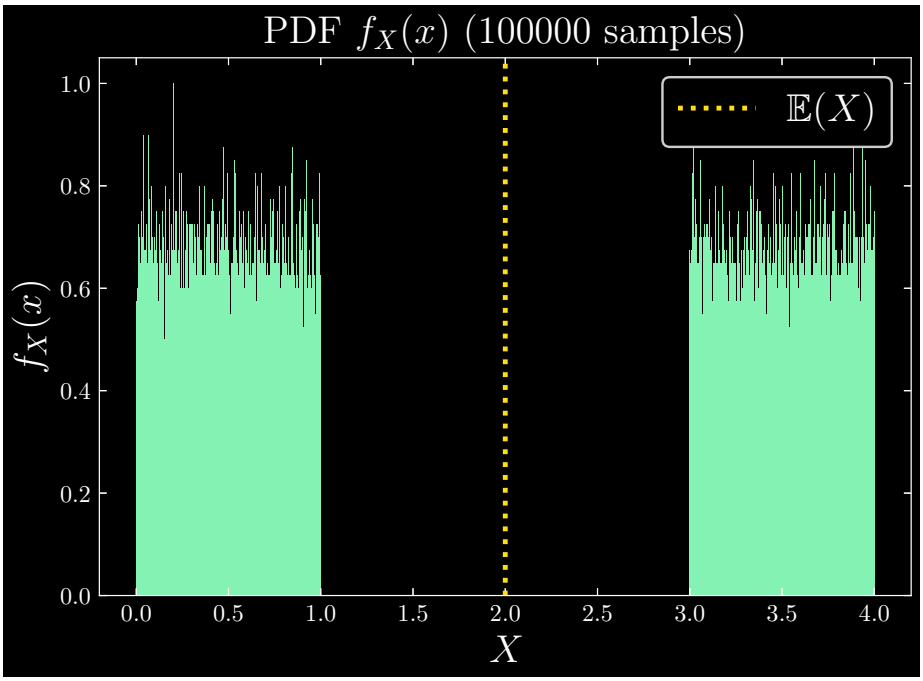
$$\begin{aligned}\mathbb{E}(X^2) &= \frac{1}{2}\left(\int_0^1 x^2\ dx + \int_3^4 x^2\ dx\right) \\ &= \frac{1}{2}\left(\frac{1}{3} + \frac{37}{3}\right) \\ &= \frac{19}{3} \\ \mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= \frac{19}{3} - \frac{12}{3} \\ &= \frac{7}{3} \\ \implies \text{sd}(X) &= \boxed{\sqrt{\frac{7}{3}}}\end{aligned}$$

The simulation gives us:

```
def coingame():
    flip = random.randint(0,1)
    if flip == 0:
        x = np.random.uniform(0,1)
    else:
        x = np.random.uniform(3,4)
    return x;

def samples(n):
    samples = []
    for i in range(1,n+1):
        samples.append(coingame())
```

```
    return samples;

n = 100000

exp = round(statistics.mean(samples(n)),3)
sd = round(np.sqrt(statistics.variance(samples(n))),3)
print('The expected value is ', exp, '
      and the standard deviation is ', sd)
```

The expected value is 2.004 and the standard deviation is 1.528. The distribution looks like:



Question: 3.8.14

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random variables and let $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ be constants. Show that

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

Proof. By definition of $\mathsf{Cov}(X, Y)$:

$$\mathsf{Cov}(X, Y) = \mathbb{E}\left((X - \mu_X)(Y - \mu_Y)\right)$$

First, we will calculate the means of the two sums.

$$\mathbb{E}\left(\sum_{i=1}^{m} a_i X_i\right) = \sum_{i=1}^{m} a_i \mathbb{E}(X_i) = \sum_{i=1}^{m} a_i \mu_{X_i}$$
$$\mathbb{E}\left(\sum_{j=1}^{n} b_j Y_j\right) = \sum_{j=1}^{n} b_j \mathbb{E}(Y_j) = \sum_{j=1}^{n} b_j \mu_{Y_j}$$

Proceeding from the definition of covariance, we can continue to derive

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

$$= \mathbb{E}\left(\left(\sum_{i=1}^{m} a_i X_i - \sum_{i=1}^{m} a_i \mu_{X_i}\right)\left(\sum_{j=1}^{n} b_j Y_j - \sum_{j=1}^{n} b_j \mu_{Y_j}\right)\right)$$
$$= \mathbb{E}\left(\left(\sum_{i=1}^{m} a_i (X_i - \mu_{X_i})\right)\left(\sum_{j=1}^{n} b_j (Y_j - \mu_{Y_j}\right)\right)$$
$$= \mathbb{E}\left(a_1(X_1 - \mu_{X_1})\sum_{j=1}^{n} b_j(Y_j - \mu_{Y_j}) + \cdots + a_n(X_n - \mu_{X_n})\sum_{j=1}^{n} b_j(Y_j - \mu_{Y_j})\right)$$
$$= \mathbb{E}(a_1(X_1 - \mu_{X_1})b_1(Y_1 - \mu_{Y_1}) + \cdots + a_1(X_1 - \mu_{X_1})b_n(Y_n - \mu_{Y_n}) + \cdots$$
$$+ a_n(X_1 - \mu_{X_1})b_1(Y_1 - \mu_{Y_1}) + \cdots + a_n(X_1 - \mu_{X_1})b_n(Y_n - \mu_{Y_n}))$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbb{E}(a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}))$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \,\mathsf{Cov}(X_i, Y_j)$$

□

Question: 3.8.15

Let

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{3}(x+y) & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{V}(2X - EY + 8)$.

The two approaches to this problem are to either use the definition $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, or to derive the marginal densities and individually calculate the expectations and variances for X, Y. The first is far shorter, but we will solve using both methods.

Given $\mathbb{V}(2X-3Y+8)$, observe that this is equivalent to $\mathbb{V}(2X-3Y)$, using the fact that constant terms do not affect the variance of a sum of random variables. Therefore we calculate

$$\mathbb{V}(2X-3Y) = \mathbb{E}((2X-3Y)^2) - \mathbb{E}(2X-3Y)^2$$

$$\begin{aligned}
\mathbb{E}((2X-3Y)^2) &= \int_0^2 \int_0^1 (2x-3y)^2 \frac{1}{3}(x+y)\ dx\ dy = \frac{86}{9} \\
\mathbb{E}(2X-3Y) &= \int_0^2 \int_0^1 (2x-3y) \frac{1}{3}(x+y)\ dx\ dy = -\frac{23}{9} \\
\mathbb{V}(2X-3Y) &= \frac{86}{9} - \left(-\frac{23}{9}\right)^2 \\
&= \boxed{\frac{245}{81}}
\end{aligned}$$

For the second method, begin with deriving the marginal densities:

$$\begin{aligned}
f_X(x) &= \int_0^2 \frac{1}{3}(x+y)\ dy = \frac{2}{3}(x+1), \quad 0 \le x \le 1 \\
f_Y(y) &= \int_0^1 \frac{1}{3}(x+y)\ dx = \frac{1}{3}\left(\frac{1}{2}+y\right), \quad 0 \le y \le 2
\end{aligned}$$

The constituent expectations for calculations down the line are

$$\begin{aligned}
\mathbb{E}(X) &= \int_0^1 \frac{2}{3}(x^2+x)\ dx = \frac{5}{9} \\
\mathbb{E}(X^2) &= \int_0^1 \frac{2}{3}(x^3+x^2)\ dx = \frac{7}{18} \\
\mathbb{E}(Y) &= \int_0^2 \frac{1}{3}\left(y^2+\frac{y}{2}\right)\ dy = \frac{11}{9} \\
\mathbb{E}(Y^2) &= \int_0^2 \frac{1}{3}\left(y^3+\frac{y^2}{2}\right)\ dy = \frac{16}{9} \\
\mathbb{E}(XY) &= \int_0^2 \int_0^1 xy f_{X,Y}(x,y)\ dx\ dy \\
&= \int_0^2 \int_0^1 \frac{1}{3}(x^2y+xy^2\ dx\ dy = \frac{2}{3}
\end{aligned}$$

For the last step, the desired variance is calculated as

$$\mathbb{V}(2X-3Y+8) = 4\mathbb{V}(X) + 9\mathbb{V}(Y) - 12\,\mathsf{Cov}(X,Y)$$

and the remaining quantities are

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162} \\
\mathbb{V}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \frac{16}{9} - \left(\frac{11}{9}\right)^2 = \frac{23}{81} \\
\mathsf{Cov}(X,Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \frac{2}{3} - \left(\frac{5}{9}\right)\left(\frac{11}{9}\right) = -\frac{1}{81}
\end{aligned}$$

At last, we combine the quantities to get

$$\begin{aligned}
\mathbb{V}(2X - 3Y + 8) &= 4\mathbb{V}(X) + 9\mathbb{V}(Y) - 12\,\mathsf{Cov}(X,Y) \\
&= 4\left(\frac{13}{162}\right) + 9\left(\frac{23}{81}\right) - 12\left(-\frac{1}{81}\right) \\
&= \boxed{\frac{245}{81}}
\end{aligned}$$

Question: 3.8.16

Let $r(x)$ be a function of x and let $s(y)$ be a function of y. Show that

$$\mathbb{E}(r(X)s(Y) \mid X) = r(X)\mathbb{E}(s(Y) \mid X).$$

Also, show that $\mathbb{E}(r(X) \mid X) = r(X)$.

Proof. By definition of conditional expectation,

$$\begin{aligned}
\mathbb{E}(r(X)s(Y)) &= \int r(x)s(y)f_{Y|X}(y \mid x)\, dy \\
&= r(X)\int s(y)f_{Y|X}(y \mid x)\, dy \\
&= r(X)\mathbb{E}(s(Y) \mid X)
\end{aligned}$$

For the second result, use the definition of conditional density.

$$\begin{aligned}
\mathbb{E}(r(X) \mid X) &= \int r(X)f_{Y|X}(y \mid x)\, dy \\
&= r(X)\int f_{Y|X}(y \mid x)\, dy \\
&= r(X)\int \frac{f_{X,Y}(x,y)}{f_X(x)}\, dy \\
&= r(X)\frac{f_X(x)}{f_X(x)} \\
&= r(X)
\end{aligned}$$

□

Question: 3.8.16

Prove that

$$\mathbb{V}(Y) = \mathbb{E}\mathbb{V}(Y \mid X) + \mathbb{V}\mathbb{E}(Y \mid X).$$

Hint: Let $m = \mathbb{E}(Y)$ and let $b(x) = \mathbb{E}(Y \mid X = x)$. Note that $\mathbb{E}(b(X)) = \mathbb{E}\mathbb{E}(Y \mid X) = \mathbb{E}(Y) = m$. Bear in mind that b is a function of x. Now write $\mathbb{V}(Y) = \mathbb{E}(Y - m)^2 = \mathbb{E}((Y - b(X)) + (b(X) - m))^2$. Expand the square and take the expectation. You then have to take the expectation of three terms. In each case, use the rule of the iterated expectation: $\mathbb{E}(\text{stuff}) = \mathbb{E}(\mathbb{E}(\text{stuff} \mid X))$.

Proof. Here we prove the Law of Total Variance. Together with the Law of Total Expectation, we establish Adam (expectation) and Eve's (variance) laws. Following the hints, derive

$$\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y - m)^2 \\
&= \mathbb{E}((Y - b(x)) + (b(x) - m))^2 \\
&= \mathbb{E}((Y - b(x))^2 + 2(Y - b(x))(b(x) - m) + (b(x) - m)^2) \\
&= \underbrace{\mathbb{E}(Y - b(x))^2}_{I} + \underbrace{2\mathbb{E}((Y - b(x))(b(x) - m))}_{II} + \underbrace{\mathbb{E}(b(x) - m)^2}_{III}
\end{aligned}$$

Terms I and III are easier to evaluate, applying the Law of Total Expectation in both:

$$\begin{aligned}
\mathbb{E}(Y - b(x))^2 &= \mathbb{E}\mathbb{E}((Y - b(x))^2 \mid X) \\
&= \mathbb{E}\left(\int (y - b(x))^2 f_{Y|X}(y \mid x)\, dy\right) \\
&= \mathbb{E}\mathbb{V}(Y \mid X) \\
\mathbb{E}(b(X) - m)^2 &= \mathbb{E}(b(x) - \mathbb{E}(b(x)))^2 \\
&= \mathbb{E}\mathbb{E}((b(x) - \mathbb{E}(b(x)))^2 \mid X) \\
&= \mathbb{E}\left(\int (b(x) - \mathbb{E}(b(x)))^2 f_{Y|X}(y \mid x)\, dy\right) \\
&= \mathbb{E}\mathbb{V}(\mathbb{E}(Y \mid X) \mid X) \\
&= \mathbb{V}\mathbb{E}(Y \mid X)
\end{aligned}$$

where $b(x) = \mathbb{E}(Y \mid X) = \mu(x)$ in term I. To evaluate II, if we are to prove the desired result, we should end up with this term vanishing. Begin with

$$\begin{aligned}
\mathbb{E}((Y - b(x))(b(x) - m)) &= \mathbb{E}(Y\mathbb{E}(Y \mid X)) - \mathbb{E}(Y)^2 \\
&\quad - \mathbb{E}(\mathbb{E}(Y \mid X)^2) + \mathbb{E}(Y)\mathbb{E}(\mathbb{E}(Y \mid X)) \\
&= \mathbb{E}(Y\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)^2) - \mathbb{E}(Y)^2 + \mathbb{E}(Y)^2 \\
&= \mathbb{E}(\mathbb{E}(Y \mid X)(Y - \mathbb{E}(Y \mid X)))
\end{aligned}$$

By the Law of Total Expectation, we can write as our next step

$$= \mathbb{E}(\mathbb{E}(\underbrace{\mathbb{E}(Y \mid X)}(Y - \mathbb{E}(Y \mid X)) \mid X))$$

Now, using the results of exercise 3.8.16, we can pull the bracketed term out of the second nested expectation to get

$$\begin{aligned}
&= \mathbb{E}(\mathbb{E}(Y \mid X)\mathbb{E}(Y - \mathbb{E}(Y \mid X) \mid X)) \\
&= \mathbb{E}(\mathbb{E}(Y \mid X)[\mathbb{E}(Y \mid X) - \mathbb{E}(\mathbb{E}(Y \mid X) \mid X)]) \\
&= \mathbb{E}(\mathbb{E}(Y \mid X)[\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)]) \\
&= 0
\end{aligned}$$

Thereby proving

$$\begin{aligned}
\mathbb{V}(Y) = \mathbb{E}(Y - m)^2 &= \mathbb{E}(Y - b(x))^2 + \mathbb{E}(b(x) - m)^2 \\
&= \mathbb{E}\mathbb{V}(Y \mid X) + \mathbb{V}\mathbb{E}(Y \mid X)
\end{aligned}$$

Note that this is a rather convoluted way to prove the Law of Total Variance; I am not sure why the author chose to gave the hints he did. □

Question: 3.8.18

Show that if $\mathbb{E}(X \mid Y = y) = c$ for some constant c, then X and Y are uncorrelated.

Proof. Let $\mathbb{E}(X \mid Y = y) = c$. Then the following are true:

(1) $\mathbb{E}(\mathbb{E}(X \mid Y = y)) = \mathbb{E}(X) = \int x f_X(x)\ dx = \mathbb{E}(c) = c$

(2) $\mathrm{c} = \int x f_{X|Y}(x \mid y)\ dx = \int x \frac{f_{X,Y}(x,y)}{f_Y(y)}\ dx$

Combining these facts, we must have

$$\begin{aligned}
&\implies && \int x f_X(x)\ dx = \int x \frac{f_{X,Y}(x,y)}{f_Y(y)}\ dx \\
&\implies f_X(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} &&\implies f_X(x) f_Y(y) = f_{X,Y}(x,y) \\
&\implies && X \sqcup Y \\
&\implies && \rho_{X,Y} = 0
\end{aligned}$$

where the second implication follows from reasoning that

$$\int \left(x f_X(x) - x \frac{f_{X,Y}(x,y)}{f_Y(y)} \right)\ dx = \int x \left(f_X(x) - \frac{f_{X,Y}(x,y)}{f_Y(y)} \right)\ dx = 0$$

But this equality holds only when the integrand is an odd function, or 0. Since we mandate no condition on the former, it must be the case that

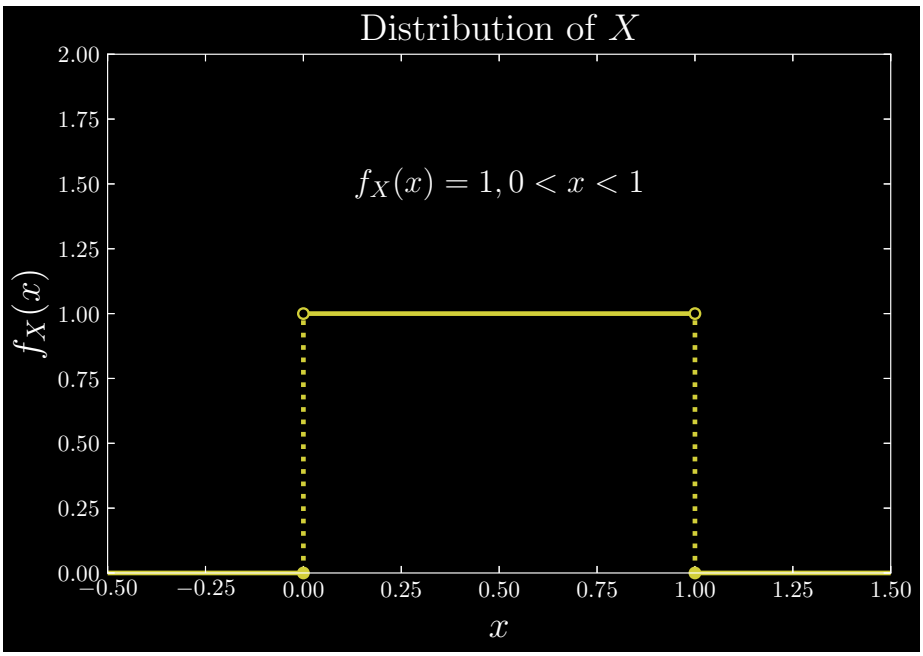
$$f_X(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

□

Question: 3.8.19

This question is to help you understand the idea of a **sampling distribution**. Let $X_1, \ldots, X_n$ be IID with mean μ and variance σ^2. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then $\overline{X}_n$ is a **statistic**, that is, a function of the data. Since $\overline{X}_n$ is a random variable, it has a distribution. This distribution is called the *sampling distribution of the statistic*. Recall from Theorem 3.17 that $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n$. Don't confuse the distribution of the data $f_X(x)$ and the distribution of the statistic $f_{\bar{X}_n}$. To make this clear, let $X_1, \ldots, X_n \sim \text{Uniform}(0,1)$. Let f_X be the density of the Uniform$(0,1)$. Plot f_X. Now let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Find $\mathbb{E}(\overline{X}_n)$ and $\mathbb{V}(\overline{X}_n)$. Plot them as a function of n. Interpret. Now simulate the distribution of $\overline{X}_n$ for $n = 1, 5, 25, 100$. Check that the simulated values of $\mathbb{E}(\overline{X}_n)$ and $\mathbb{V}(\overline{X}_n)$ agree with your theoretical calculations. What do you notice about the sampling distribution of $\overline{X}_n$ as n increases?

First, the density f_X of $X \sim \text{Uniform}\,(0,1)$ looks like



Using the definition for $\overline{X}_n$, the theoretical derivations of the expectation

and variance are

$$\begin{aligned}\mathbb{E}(\overline{X}_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^1 x_i\ dx_i\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\\ &= \boxed{\frac{1}{2}}\end{aligned}$$

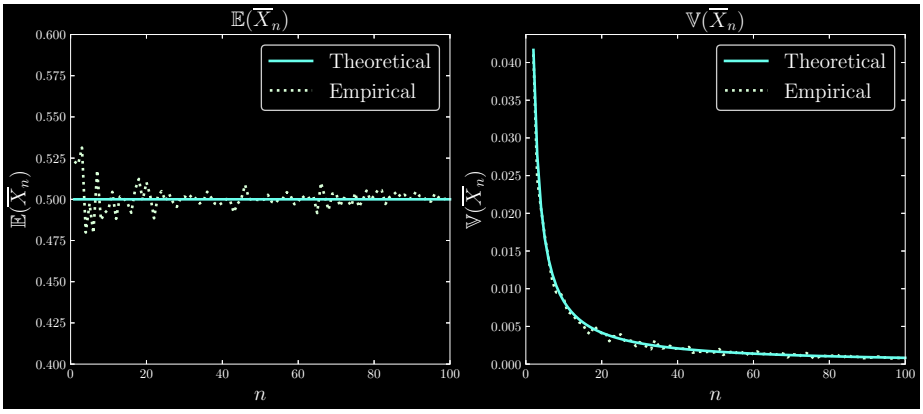
Now, $\overline{X}_n^2$ is given by

$$\overline{X}_n^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}X_i\right)^2$$
$$\frac{1}{n^2}\left(\underbrace{\sum_{i=1}^{n}X_i^2}_{n\text{ terms}} + \underbrace{\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}X_iX_j}_{n(n-1)\text{ terms}}\right)$$

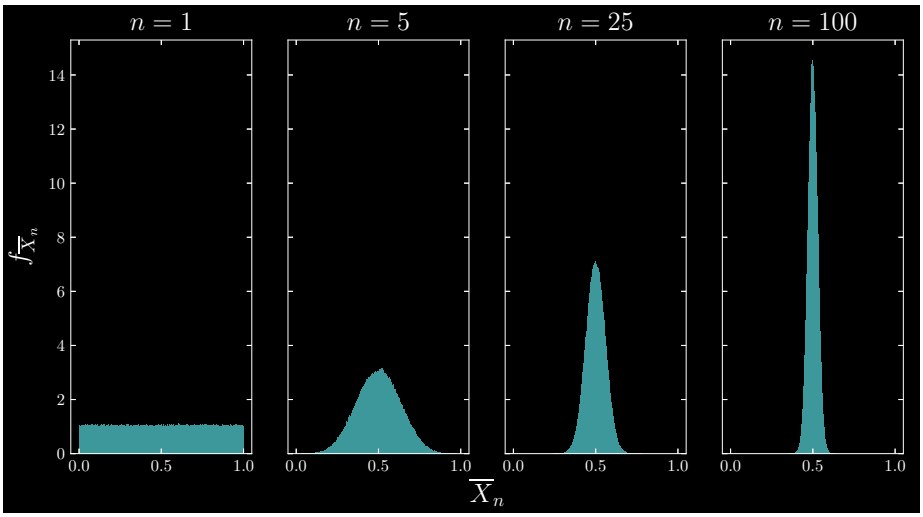
Using the fact that the random variables are IID, we can derive

$$\begin{aligned}\mathbb{E}(\overline{X}_n^2) &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\int_0^1 x_i^2\ dx_i + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\int_0^1 x_i\ dx_i\int_0^1 x_j\ dx_j\right)\\ &= \frac{1}{n^2}\left(n\left(\frac{1}{3}\right) + n(n-1)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right)\\ &= \frac{1}{n^2}\left(\frac{1}{4}n^2 - \frac{1}{4}n + \frac{1}{3}n\right)\\ &= \frac{1}{4} + \frac{1}{12n}\\ \mathbb{V}(\overline{X}_n) &= \mathbb{E}(\overline{X}_n^2) - \mathbb{E}(\overline{X}_n)^2\\ &= \frac{1}{12n} + \frac{1}{4} - \left(\frac{1}{2}\right)^2\\ &= \boxed{\frac{1}{12n}}\end{aligned}$$

Empirically, we can verify that randomly generating a sampling distribution of values of $\overline{X}_n$ and taking the expectation and variance agrees with the theory:



Lastly, we graph the sampling distributions as n increases:



which gradually increases in concentration ($\mathbb{V}\left(\overline{X}_n\right)$ goes to 0) around the mean $\mathbb{E}\left(\overline{X}_n\right) = 1/2$.

Question: 3.8.20

Prove Lemma 3.21.

Lemma (Wasserman 3.21)

If a is a vector and X is a random vector with mean μ and variance Σ, then $\mathbb{E}(a^T X) = a^T \mu$ and $\mathbb{V}(a^T X) = a^T \Sigma a$. If A is a matrix then $\mathbb{E}(AX) = A\mu$ and $\mathbb{V}(AX) = A\Sigma A^T$.

Proof. Let $a, X \in M_{n\times 1}$, where $M_{n\times 1}$ is the set of $n \times 1$ vectors, and X is a random vector with mean μ and variance Σ.

By vector multiplication and the definition of a random vector mean, we first derive

$$\mathbb{E}(a^T X) = \mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} \mathbb{E}(X_i) = \sum_{i=1}^{n} a_i \mu_i = a^T \mu$$

Now using the definition of the variance-covariance matrix Σ, we have $\mathbb{V}(a^T X)$ equal to

$$\begin{aligned}
&= \mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) \\
&= \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i) + 2\sum_{i=1}^{n}\sum_{i<j}^{n} a_i a_j \,\mathsf{Cov}(X_i, X_j) \\
&= \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \cdots & \mathbb{V}(X_n) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \\
&= a^T \Sigma a
\end{aligned}$$

Now let $A \in M_{n\times n}$. Then

$$AX = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} a_{1i} X_i \\ \vdots \\ \sum_{i=1}^{n} a_{ni} X_i \end{bmatrix}$$

which has expectation

$$\begin{aligned}
\mathbb{E}(AX) &= \mathbb{E}\begin{bmatrix} \sum_{i=1}^{n} a_{1i} X_i \\ \vdots \\ \sum_{i=1}^{n} a_{ni} X_i \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} a_{1i} \mathbb{E}(X_i) \\ \vdots \\ \sum_{i=1}^{n} a_{ni} \mathbb{E}(X_i) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} a_{1i} \mu_i \\ \vdots \\ \sum_{i=1}^{n} a_{ni} \mu_i \end{bmatrix} \\
&= \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \\
&= A\mu
\end{aligned}$$

and variance $\mathbb{V}(AX)$ equal to

$$\begin{bmatrix} \mathbb{V}(\sum a_{1i}X_i) & \mathsf{Cov}(\sum a_{1i}X_i) & \cdots & \mathsf{Cov}(\sum a_{1i}X_i, \sum a_{ni}X_i) \\ \mathsf{Cov}(\sum a_{2i}X_i, \sum a_{1i}X_i) & \mathbb{V}(\sum a_{2i}X_i) & \cdots & \mathsf{Cov}(\sum a_{2i}X_i, \sum a_{ni}X_i) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(\sum a_{ni}X_i, \sum a_{1i}X_i) & \mathsf{Cov}(\sum a_{ni}X_i, \sum a_{2i}X_i) & \cdots & \mathbb{V}(\sum a_{ni}X_i) \end{bmatrix}$$

To further simplify, use the result of exercise 14 to get

$$\begin{bmatrix} \sum_{i=1}^{n} a_{1i}^2 \mathbb{V}(X_i) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n} a_{1i}a_{1j}\,\mathsf{Cov}(X_i, X_j) \\ +2\sum_{i=1}^{n-1}\sum_{i<j}^{n} a_{1i}a_{1j}\,\mathsf{Cov}(X_i, X_j) & & \\ \sum_{i=1}^{n}\sum_{j=1}^{n} a_{2i}a_{1j}\,\mathsf{Cov}(X_i, X_j) & \cdots & \sum_{i=1}^{n}\sum_{j=1}^{n} a_{2i}a_{nj}\,\mathsf{Cov}(X_i, X_j) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ni}a_{1j}\,\mathsf{Cov}(X_i, X_j) & \cdots & \sum_{i=1}^{n} a_{ni}^2 \mathbb{V}(X_i) \\ & & +2\sum_{i=1}^{n-1}\sum_{i<j}^{n} a_{ni}a_{nj}\,\mathsf{Cov}(X_i, X_j) \end{bmatrix}$$

for which each entry is precisely the value of each entry in the matrix product

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} \mathbb{V}(X_1) & \mathsf{Cov}(X_1, X_2) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \cdots & \mathbb{V}(X_n) \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}$$

which is $\mathbb{V}(AX) = A\Sigma A^T$, as desired. □

Question: 3.8.21

Let X and Y be random variables. Suppose that $\mathbb{E}(Y \mid X) = X$. Show that $\mathsf{Cov}(X, Y) = \mathbb{V}(X)$.

Proof. Let $\mathbb{E}(Y \mid X) = X$. Then $\mathbb{E}(\mathbb{E}(Y \mid X)) = \mathbb{E}(Y) = \mathbb{E}(X)$. Moreover,

$$x = \int y f_{Y|X}(y \mid x)\, dy = \int y \frac{f_{Y,X}(y, x)}{f_X(x)}\, dy$$

This implies

$$\begin{aligned} \implies \quad x^2 f_X(x) &= \int xy f_{Y,X}(y, x)\, dy \\ \implies \int x^2 f_X(x)\, dx &= \int\int xy f_{Y,X}(y, x)\, dy\, dx \\ \implies \quad \mathbb{E}(X^2) &= \mathbb{E}(XY) \\ \implies \quad \mathsf{Cov}(X, Y) &= \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= \mathbb{V}(X) \end{aligned}$$

□

Question: 3.8.22

Let $X \sim \text{Uniform}(0,1)$. Let $0 < a < b < 1$. Let

$$Y = \begin{cases} 1 & 0 < x < b \\ 0 & \text{otherwise} \end{cases}$$

and let

$$Z = \begin{cases} 1 & a < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Are Y and Z independent? Why/Why not?

(b) Find $\mathbb{E}(Y \mid Z)$. Hint: What values z can Z take? Now find $\mathbb{E}(Y \mid Z = z)$.

(a) The probabilities for all values of Y, Z are

$$\begin{aligned} \mathbb{P}(Y = 1) &= \int_0^b dx = b \\ \mathbb{P}(Y = 0) &= \int_b^1 dx = 1 - b \\ \mathbb{P}(Z = 1) &= \int_a^1 dx = 1 - a \\ \mathbb{P}(Z = 0) &= \int_0^a dx = a \end{aligned}$$

Now, consider the following case

$$\mathbb{P}(Y = 1, Z = 1) = \int_a^b dx = b - a$$

However,

$$\mathbb{P}(Y = 1)\mathbb{P}(Z = 1) = b(1 - a)! = b - a = \mathbb{P}(Y = 1, Z = 1)$$

Thus Y and Z are not independent.

(b) The conditional expectations are

$$\mathbb{E}(Y \mid Z) = \begin{cases} 1 \cdot \mathbb{P}(Y \mid Z = 0) = \dfrac{\int_0^a dx}{\int_0^a dx} = 1 & z = 0 \\ 1 \cdot \mathbb{P}(Y \mid Z = 1) = \dfrac{\int_a^b dx}{\int_a^1 dx} = \dfrac{b - a}{1 - a} & z = 1 \end{cases}$$

Question: 3.8.23

Find the moment generating function for the Poisson, Normal, and Gamma distributions.

Poisson
For $X \sim \text{Poisson}(\lambda)$, we have mass

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

with moment generating function

$$\begin{aligned}\psi_X(t) &= \sum_{x-0}^{\infty} e^{tx}e^{-\lambda}\frac{\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda}e^{\lambda e^t} \\ &= \boxed{e^{\lambda(e^t-1)}}\end{aligned}$$

Normal
For $X \sim N(0,1)$, we have density

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

with moment generating function

$$\begin{aligned}\psi_X(t) = \mathbb{E}(e^{tx}) &= \frac{1}{\sigma\sqrt{2\pi}}\int \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 + tx\right)\,dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int \exp\left(-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2) + tx\right)\,dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int \exp\left(-\frac{1}{2\sigma^2}x^2 + \left(\frac{\mu}{\sigma^2}+t\right)x - \frac{\mu^2}{2\sigma^2}\right)\,dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\int \exp\left(-\frac{1}{2\sigma^2}x^2 + \left(\frac{\mu}{\sigma^2}+t\right)x\right)\,dx\end{aligned}$$

Here we complete the square by adding and subtracting $(\sigma^2/2)(\mu/\sigma^2+t)^2$, and writing

$$= \frac{1}{\sigma\sqrt{2\pi}}\int \exp\left(-\frac{\mu^2}{2\sigma^2}(x-(\mu+\sigma^2 t))^2 + \frac{\sigma^2}{2}\left(\frac{\mu}{\sigma^2}+t\right)^2\right)\,dx$$

Now substitute $u = x - (\mu + \sigma^2 t)$. Then $du = dx$, and

$$\begin{aligned}
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2} + \frac{\sigma^2}{2}\left(\frac{\mu}{\sigma^2} + t\right)^2\right) \int \exp\left(-\frac{1}{2\sigma^2}\mu^2\right)\, du \\
&= \exp\left(-\frac{\mu^2}{2\sigma^2} + \frac{\sigma^2}{2}\left(\frac{\mu}{\sigma^2} + t\right)^2\right) \\
&= \exp\left(-\frac{\mu^2}{2\sigma^2} + \frac{\sigma^2}{2}\left(\frac{\mu^2}{\sigma^4} + \frac{2\mu}{\sigma^2}t + t^2\right)\right) \\
&= \boxed{\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)}
\end{aligned}$$

Gamma

For $X \sim \text{Gamma}(\alpha, \beta)$, we have density

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x, \alpha, \beta > 0$$

where

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

and with moment generating function

$$\psi_X(t) = \mathbb{E}(e^{tx}) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(1/\beta - t)}\, dx$$

Let $y = x(1/\beta - t)$. Then $dy = (1/\beta - t)dx$, and we have for $t < 1/\beta$

$$\begin{aligned}
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \frac{1}{(1/\beta - t)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y}\, dy \\
&= \frac{1}{\beta^\alpha \Gamma(\alpha)} \frac{1}{(1/\beta^\alpha)(1 - t\beta)^\alpha} \Gamma(\alpha) \\
&= \boxed{(1 - t\beta)^{-\alpha}}
\end{aligned}$$

Question: 3.8.24

Let $X_1, dots, X_n \sim \text{Exp}(\beta)$. Find the moment generating function of X_i. Prove that $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \beta)$.

Proof. For $X_i \sim \text{Exp}(\beta)$, we have density

$$f_{X_i}(x_i) = \frac{1}{\beta} \exp(-x_i/\beta), \quad x_i, \beta > 0$$

with moment generating function

$$\begin{aligned}
\psi_{X_i}(t) = \mathbb{E}(e^{tx}) &= \frac{1}{\beta} \int_0^\infty \exp(-x(1/\beta - t)\, dx \\
&= \frac{1}{1 - t\beta}
\end{aligned}$$

for $t < 1/\beta$. By Lemma 3.31(2) and assuming independence of the X_i's, for $X = \sum_{i=1}^{n} X_i$ we have

$$\psi_X(t) = \prod_i \psi_{X_i}(t) = \prod_i (1 - t\beta) = (1 - t\beta)^{-n}$$

Implying $X \sim \text{Gamma}(n, \beta)$ per the results of exercise 3.8.23.

□

Chapter 4: Inequalities

Import Packages

Please see the associated GitHub repo for all code and comments to simulations. [LINK HERE]

```
import numpy as np
from numpy.random import choice
import random
import matplotlib.pyplot as plt
import scienceplots
```

Question: 4.5.1

Let $X \sim \text{Exp}(\beta)$. Find $\mathbb{P}(|X - \mu_X| \geq k\sigma_X)$ for $k > 1$. Compare this to the bound you get from Chebyshev's inequality.

For $X \sim \text{Exp}(\beta)$, $\mu_X = \beta$ and $\sigma_X^2 = \beta^2$. Calculating the probability exactly,

$$\begin{aligned}
\mathbb{P}(|X - \mu_X| \geq k\sigma_X) &= \mathbb{P}(X - \mu_X \geq k\beta) \\
&= \mathbb{P}(X \geq (k+1)\beta) \\
&= \frac{1}{\beta} \int_{(k+1)\beta}^{\infty} \exp(-x/\beta)\ dx \\
&= -\exp(-x/\beta) \Big|_{(k+1)\beta}^{\infty} \\
&= \exp(-(k+1))
\end{aligned}$$

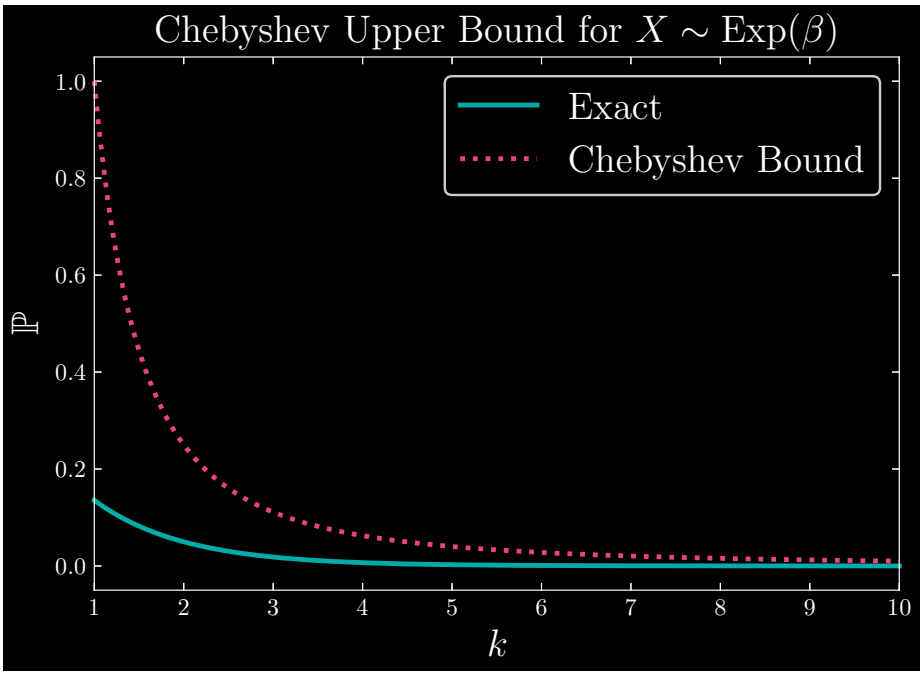
with the first equality following from the fact that since $x > 0$, $\mu = \sigma_X = \beta$, no x can be $k\sigma_X = k\beta$ to the left of the mean.

By Chebyshev's inequality,

$$\mathbb{P}(|X - \beta| \geq k\sigma_X) \leq \frac{\sigma_X^2}{k^2\sigma_X^2} = \frac{1}{k^2}$$

where $\exp(-(k+1)) < 1/k^2$ for $k > 1$.

Here we graph the exact probability and the Chebyshev bound as functions of k:



Question: 4.5.2

Let $X \sim \text{Poisson}(\lambda)$. Use Chebyshev's inequality to show that $\mathbb{P}(X \geq 2\lambda) \leq 1/\lambda$.

Proof. For $X \sim \text{Poisson}(\lambda)$, we have $\mu_X = \sigma_X^2 = \lambda$. In the case where $X - \lambda \geq 0$ and $t = \lambda$, we have

$$\begin{aligned}\mathbb{P}(|X - \lambda| \geq \lambda) &= \mathbb{P}(X - \lambda \geq \lambda) \\ &= \mathbb{P}(X \geq 2\lambda)\end{aligned}$$

Note that when $X - \lambda < 0$ with $t = \lambda$, we have

$$\begin{aligned}\mathbb{P}(|X - \lambda| \geq t) &= \mathbb{P}(-(X - \lambda) \geq \lambda) \\ &= \mathbb{P}(X \leq \lambda - \lambda) \\ &= \mathbb{P}(X \leq 0) \\ &= 0\end{aligned}$$

so we need only consider the case where $X - \lambda \geq 0$. Applying Chebyshev's inequality, we can conclude

$$\mathbb{P}(X \geq 2\lambda) \leq \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

□

Question: 4.5.3

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Bound $\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right)$ using Chebyshev's inequality and using Hoeffding's inequality. Show that, when n is large, the bound from Hoeffding's inequality is smaller than the bound from Chebyshev's inequality.

PROOF. For $\overline{X}_n$, we have $\mathbb{E}\left(\overline{X}_n\right) = p$ and $\mathbb{V}\left(\overline{X}_n\right) = p(1-p)/n$. The upper bounds from Chebyshev and Hoeffding are

$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2} \qquad \text{(Chebyshev)}$$
$$\mathbb{P}\left(|\overline{X}_n - p| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \qquad \text{(Hoeffding)}$$

Taking the ratio of the Hoeffding to Chebyshev limits, we aim to calculate

$$\lim_{n\to\infty} \frac{2e^{-2n\epsilon^2}}{p(1-p)/n\epsilon^2} = \frac{1}{p(1-p)} \lim_{n\to\infty} \frac{2n\epsilon^2}{e^{2n\epsilon^2}}$$

Using l'Hôpital's rule to evaluate the limit since $\lim_{n\to\infty} 2n\epsilon^2 = +\infty$ and $\lim_{n\to\infty} e^{2n\epsilon^2} = +\infty$, we have

$$\frac{1}{p(1-p)} \lim_{n\to\infty} \frac{2n\epsilon^2}{e^{2n\epsilon^2}} = \frac{1}{p(1-p)} \lim_{n\to\infty} \frac{2\epsilon^2}{2\epsilon^2 e^{2n\epsilon^2}} = \frac{1}{p(1-p)} \lim_{n\to\infty} \frac{1}{e^{2n\epsilon^2}} = 0$$

ascertaining that Hoeffding is a tighter bound than Chebyshev. □

Question: 4.5.4

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$.

(a) Let $\alpha > 0$ be fixed and define

$$\epsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

Let $\widehat{p}_n = n^{-1} \sum_{i=1}^n X_i$. Define $C_n = (\widehat{p}_n - \epsilon_n, \widehat{p}_n + \epsilon_n)$. Use Hoeffding's inequality to show that

$$\mathbb{P}(C_n \text{ contains } p) \geq 1 - \alpha.$$

In practice, we truncate the interval so it does not go below 0 or above 1.

(b) (Computer Experiment.) Let's examine the properties of this confidence interval. Let $\alpha = 0.05$ and $p = 0.4$. Conduct a simulation study to see how often the interval contains p (called the coverage). Do this for various values of n between 1 and 10000. Plot the coverage versus n.

(c) Plot the length of the interval versus n. Suppose we want the length of the interval to be no more than .05. How large should n be?

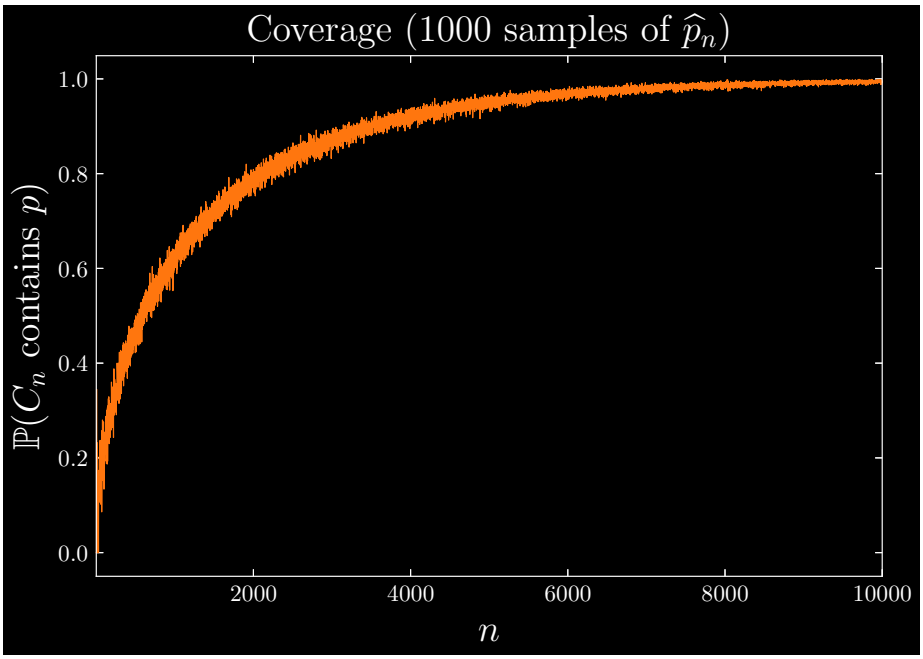
PROOF. (a) Since $\overline{X}_n = \widehat{p}_n$ and $\mathbb{P}\left(|\overline{X}_n - p| > \epsilon_n\right) \leq \alpha$, by complementary events (either $|\overline{X}_n - p| > \epsilon_n$ or $|\overline{X}_n - p| \leq \epsilon_n$), we have

$$\mathbb{P}\left(|\overline{X}_n - p| \leq \epsilon_n\right) \geq 1 - \alpha$$

□

(b) Here, we sample numerous values of $\widehat{p}_n$, construct C_n from each value, and determine if $p \in C_n$.

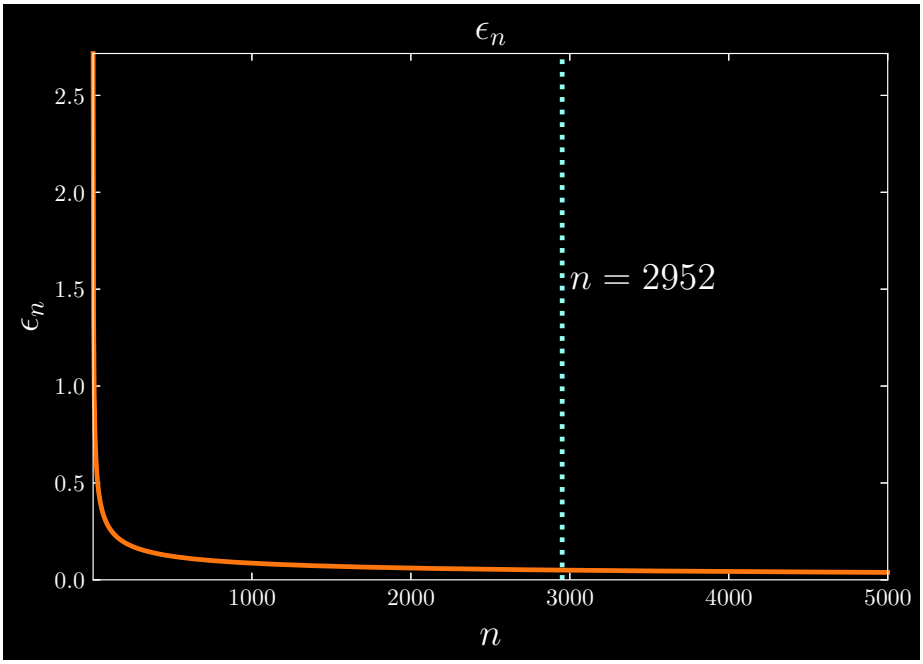
```
k, n = 1000, 10000
alpha = 0.05
eps = np.sqrt((1/(2*n))*np.log(2/alpha))
p = 0.4

def confinttest(k,n):
    s = np.random.binomial(1,p,size=(k,n)).mean(axis=1)
    return (p > s - eps) & (p < s + eps);

def coverage():
    coverage = []
    for i in range(1,n+1):
        count = np.count_nonzero(confinttest(k,i) == True)
        coverage.append(count / k)
    return coverage;
```



As $n \to +\infty$, we see that the coverage asymptotically approaches 1. Intuitively this is sensible. Since $X_i \sim \text{Bernoulli}\,(p)$, for large n, $\widehat{p}_n \to p$ (this is the weak law of large numbers, covered in the next chapter).

(c) The length of C_n is $2\epsilon_n$, and for the length to be no more than 0.05, have $n \geq 2952$.

```
def intlength(n):
    epsilons = np.empty(n)
    for i in range(1,n+1):
        epsilon = np.sqrt((1/(2*i))*np.log(2/alpha))
        epsilons[i-1] = 2*epsilon
    return epsilons;

n = 5000
nsolution = np.argwhere(intlength(n) < 0.05).min() + 1
```



Question: 4.5.5

Prove Mill's inequality, Theorem 4.7. Hint. Note that $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$. Now write out what $\mathbb{P}(Z > t)$ means and note that $x/t > 1$ whenever $x > t$.

Theorem (Wasserman 4.7) (Mill's Inequality)
Let $Z \sim N(0,1)$. Then

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}.$$

Proof. By symmetry of the standard normal distribution,

$$\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$$

Then

$$\begin{aligned}\mathbb{P}(|Z| > t) &= 2\mathbb{P}(Z > t) \\ &= 2\int_t^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz\end{aligned}$$

Now, multiply both sides by t:

$$\begin{aligned}\implies t\mathbb{P}(|Z|>t) &= \sqrt{\frac{2}{\pi}}t\int_t^\infty \exp\left(-\frac{z^2}{2}\right)\,dz \\ &\leq \sqrt{\frac{2}{\pi}}\int_t^\infty z\exp\left(-\frac{z^2}{2}\right)\,dz \\ &= \sqrt{\frac{2}{\pi}}e^{-t^2/2}\end{aligned}$$

which lastly implies that $\mathbb{P}(|Z|>t) \leq \sqrt{\frac{2}{\pi}}e^{-t^2/2}/t$. □

Question: 4.5.6

Let $Z \sim N(0,1)$. Find $\mathbb{P}(|Z|>t)$ and plot this as a function of t. From Markov's inequality, we have the bound $\mathbb{P}(|Z|>t) \leq \frac{\mathbb{E}|Z|^k}{t^k}$ for any $k>0$. Plot these bounds for $k=1,2,3,4,5$ and compare them to the true value of $\mathbb{P}(|Z|>t)$. Also, plot the bound from Mill's inequality.

Some initial observations to bear in mind:

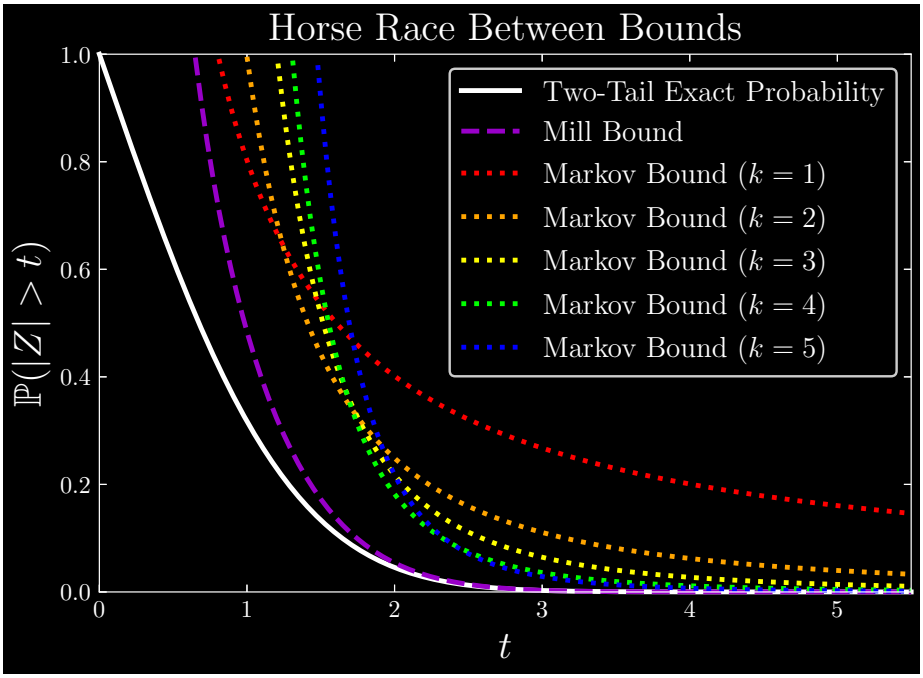
- Observe that $\mathbb{P}(|Z|>t)$ is the two-tailed probability that we observe $Z>t$, when $Z\geq 0$, or $Z<-t$, when $Z<0$.
- Using Markov's inequality, $\mathbb{P}(|Z|>t) = \mathbb{P}\left(|Z|^k > t^k\right) \leq \mathbb{E}|Z|^k/t^k$.
- The expectation term $\mathbb{E}(|Z|)$ is defined as

$$\mathbb{E}|Z| = \int_0^{+\infty} z f_Z(z)\,dz + \int_{-\infty}^0 z f_Z(z)\,dz = 2\int_0^{+\infty} z f_Z(z)\,dz = \sqrt{\frac{2}{\pi}}$$

Running a horse race between the exact two-tail probability and the Markov and Mill bounds, we find:

```
def normtest(t):
    normal_dist = norm(0,1)
    domain = np.arange(0,t+0.01,0.01)
    probs = np.empty(len(domain))
    for i, j in zip(domain, range(0, len(domain))):
        probs[j] = (2*(1-normal_dist.cdf(i)))
    return probs;

def markov(k, t):
    domain = np.arange(0.01,t+0.01,0.01)
    markovs = np.empty(len(domain))
    exptabszk = np.power(
                np.absolute(np.random.normal(0,1,1000)), k).mean()
    for i, j in zip(domain, range(0, len(domain))):
        markovs[j] = np.divide(exptabszk, np.power(i,k))
    return markovs;
```

```
def mill(t):
    domain = np.arange(0.01,t+0.01,0.01)
    mills = np.empty(len(domain))
    for i, j in zip(domain, range(0, len(domain))):
        mills[j] = np.sqrt(2 / math.pi)*(np.exp(—np.square(i)/2)/i)
    return mills;

t = 5.5
```



Question: 4.5.7

Let $X_1, \ldots, X_n \sim N(0,1)$. Bound $\mathbb{P}\left(|\overline{X}_n| > t\right)$ using Mill's inequality, where $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Compare to the Chebyshev bound.

First note that $\mathbb{E}\left(\overline{X}_n\right) = 0$ and $\mathbb{V}\left(\overline{X}_n\right) = 1/n$. Thus $\overline{X}_n \sim N(0, 1/n)$. In order for Mill's inequality to be applicable, we transform $\overline{X}_n$ and t to be standard normal:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\overline{X}_n}{\sqrt{1/n}}\right| > \frac{t}{\sqrt{1/n}}\right) &= \mathbb{P}\left(\left|\frac{\overline{X}_n}{\sqrt{1/n}}\right| > \sqrt{n}t\right) \\ &\leq \sqrt{\frac{2}{\pi}}\frac{e^{-\left(\sqrt{n}t\right)^2/2}}{\sqrt{n}t} \\ &= \sqrt{\frac{2}{n\pi}}\frac{e^{-nt^2/2}}{t}\end{aligned}$$

And Chebyshev's inequality is given simply by

$$\mathbb{P}\left(|\overline{X}_n| \geq t\right) \leq \frac{1}{nt^2}$$

Taking the limit of the ratio of the Mill bound to the Chebyshev for large n gives us

$$\lim_{n\to\infty} \frac{\sqrt{\frac{2}{n\pi}}\frac{e^{-nt^2/2}}{t}}{1/nt^2} = \lim_{n\to\infty} \sqrt{\frac{2n}{\pi}} te^{-nt^2/2} = \boxed{0}$$

yielding the conclusion that Mill is a smaller bound than Chebyshev.

Chapter 5: Convergence of Random Variables

Question: 5.8.1

Let $X_1, \ldots, X_n$ be IID with finite mean $\mu = \mathbb{E}(X_1)$ and finite variance $\sigma^2 = \mathbb{V}(X_1)$. Let $\overline{X}_n$ be the sample mean and let S_n^2 be the sample variance.

(a) Show that $\mathbb{E}\left(S_n^2\right) = \sigma^2$.

(b) Show that $S_n^2 \xrightarrow{\text{P}} \sigma^2$. Hint: Show that $S_n^2 = c_n n^{-1} \sum_{i=1}^n X_i^2 - d_n \overline{X}_n^2$ where $c_n \to 1$ and $d_n \to 1$. Apply the law of large numbers to $n^{-1} \sum_{i=1}^n X_i^2$ and to $\overline{X}_n$. Then use part (d) of Theorem 5.5.

Proof. (a) See exercise 3.8.8, which provides a proof of Theorem 3.17.

(b) Using the hint, we can write S_n^2 as

$$\begin{aligned} S_n^2 &= \frac{1}{n-1} \sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(X_i^2 - 2X_i \overline{X}_n + \overline{X}_n^2\right) \\ &= \frac{1}{n-1} \sum_{i=1}^n X_i^2 - \frac{n}{n-1} \overline{X}_n^2 \\ &= c_n \frac{1}{n} \sum_{i=1}^n X_i^2 - d_n \overline{X}_n^2 \end{aligned}$$

where $c_n = d_n = \frac{n}{n-1} \to 1$ as $n \to \infty$.

By WLLN, $\overline{X}_n \xrightarrow{\text{P}} \mu$. By Theorem 5.5(f), $\overline{X}_n^2 \xrightarrow{\text{P}} \mu^2$.

For the $\frac{1}{n} \sum_{i=1}^n X_n^2$ term, the task is to determine if $X_1^2, \ldots, X_n^2$ are IID. Recall from exercise 2.14.10 that if $X \sqcup Y$, and if g, h are functions, then $g(X) \sqcup h(Y)$. Therefore, for X_i, X_j such that $i \neq j$, we must have $X_i^2 \sqcup X_j^2$. Since the X_i's come from identical distributions, so must the X_i^2's.

Thus, since $\mathbb{E}\left(X_1^2\right) = \sigma^2 + \mu^2$, by WLLN, $\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow{\text{P}} \sigma^2 + \mu^2$. Ergo, applying Theorem 5.5(d) to the products $c_n \frac{1}{n} \sum_{i=1}^n X_i^2$ and $d_n \overline{X}_n^2$, we find

$$S_n^2 = c_n \frac{1}{n} \sum_{i=1}^n X_i^2 - d_n \overline{X}_n^2 \xrightarrow{\text{P}} \sigma^2 + \mu^2 - \mu^2 = \sigma^2$$

□

Question: 5.8.2

Let $X_1, X_2, \ldots$ be a sequence of random variables. Show that $X_n \xrightarrow{\text{qm}} b$ if and only if

$$\lim_{n\to\infty} \mathbb{E}(X_n) = b \quad \text{and} \quad \lim_{n\to\infty} \mathbb{V}(X_n) = 0.$$

PROOF. ($\Longrightarrow$) Let $X_n \xrightarrow{\text{qm}} b$. Then $\mathbb{E}(X_n - b)^2 \to 0$. By Theorem 5.17(b), we also have that X_n converges in L_1 to b:

$$\mathbb{E}|X_n - b| \to 0$$

If $X_n - b \geq 0$, then $\mathbb{E}|X_n - b| = \mathbb{E}(X_n - b)$, and

$$\lim_{n\to\infty} \mathbb{E}(X_n - b) = \lim_{n\to\infty} \mathbb{E}(X_n) - b = 0 \implies \lim_{n\to\infty} \mathbb{E}(X_n) = b$$

Otherwise, $\mathbb{E}|X_n - b| = \mathbb{E}(-(X_n - b))$, and

$$\lim_{n\to\infty} \mathbb{E}(-(X_n - b)) = 0 \implies \lim_{n\to\infty} \mathbb{E}(X_n - b) \implies \lim_{n\to\infty} \mathbb{E}(X_n) = b$$

as before.

Thus we must have $\lim_{n\to\infty} \mathbb{E}(X_n) = b$ in either case. Combining this fact and the premise, it must follow that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}(X_n - b)^2 &= \lim_{n\to\infty} \mathbb{E}\left(X_n^2 - 2bX_n + b^2\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(X_n^2\right) - 2b \lim_{n\to\infty} \mathbb{E}(X_n) + b^2 \\
&= \lim_{n\to\infty} \mathbb{E}\left(X_n^2\right) - 2 \lim_{n\to\infty} \mathbb{E}(X_n)^2 + \lim_{n\to\infty} \mathbb{E}(X_n)^2 \\
&= \lim_{n\to\infty} \left[\mathbb{E}\left(X_n^2\right) - \mathbb{E}(X_n)^2\right] \\
&= \lim_{n\to\infty} \mathbb{E}(X_n - \mathbb{E}(X_n))^2 \\
&= \lim_{n\to\infty} \mathbb{V}(X_n) \\
&= 0
\end{aligned}$$

($\Longleftarrow$) In the converse direction, let

$$\lim_{n\to\infty} \mathbb{E}(X_n) = b \qquad \lim_{n\to\infty} \mathbb{V}(X_n) = 0$$

Then

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{V}(X_n) &= \lim_{n\to\infty} \mathbb{E}(X_n - \mathbb{E}(X_n))^2 \\
&= \lim_{n\to\infty} \mathbb{E}\left(X_n^2 - 2X_n\mathbb{E}(X_n) + \mathbb{E}(X_n)^2\right) \\
&= \lim_{n\to\infty} \mathbb{E}\left(X_n^2\right) - 2 \lim_{n\to\infty} \mathbb{E}(X_n)^2 + \lim_{n\to\infty} \mathbb{E}(X_n)^2 \\
&= \lim_{n\to\infty} \mathbb{E}\left(X_n^2\right) - b^2 \\
&= \lim_{n\to\infty} \mathbb{E}(X_n - b)^2 \\
&= 0
\end{aligned}$$

implies $X_n \xrightarrow{\text{qm}} b$. □

Question: 5.8.3

Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_1)$. Suppose that the variance is finite. Show that $\overline{X}_n \xrightarrow{\text{qm}} \mu$.

PROOF. Since $\mathbb{E}\left(\overline{X}_n\right) = \mu$ and $\mathbb{V}\left(\overline{X}_n\right) = \sigma^2/n$, observe that

$$\lim_{n\to\infty} \mathbb{E}\left(\overline{X}_n\right) = \lim_{n\to\infty} \mu = \mu$$
$$\lim_{n\to\infty} \mathbb{V}\left(\overline{X}_n\right) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

By exercise 5.8.2, $\overline{X}_n \xrightarrow{\text{qm}} \mu$. □

Question: 5.8.4

Let $X_1, X_2, \ldots$ be a sequence of random variables such that

$$\mathbb{P}\left(X_n = \frac{1}{n}\right) = 1 - \frac{1}{n^2} \quad \text{and} \quad \mathbb{P}(X_n = n) = \frac{1}{n^2}.$$

Does X_n converge in probability? Does X_n converge in quadratic mean?

By the errata, note that $n \geq 2$.

Conjecture: $X_n \xrightarrow{\text{P}} 0$

PROOF. Observe that $\lim_{n\to\infty} \mathbb{P}(|X_n| > \epsilon) = \mathbb{P}\left(n > \epsilon \text{ or } \frac{1}{n} > \epsilon\right)$. Then by mutual exclusivity:

$$\begin{aligned}
&= \lim_{n\to\infty} \mathbb{P}(n > \epsilon)\, \mathbb{P}(X_n = n) + \lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n} > \epsilon\right) \mathbb{P}\left(X_n = \frac{1}{n}\right) \\
&= \lim_{n\to\infty} \mathbb{P}(n > \epsilon)\, \frac{1}{n^2} + \lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n} > \epsilon\right) \left(1 - \frac{1}{n^2}\right) \\
&= \lim_{n\to\infty} \mathbb{P}\left(\frac{1}{n} > \epsilon\right) \\
&= 0
\end{aligned}$$

Thus $X_n \xrightarrow{\text{P}} 0$. □

Conjecture: $X_n \xrightarrow{\text{qm}} X$

PROOF. Suppose X_n converges in quadratic mean to X. Then we have

$$\begin{aligned}\mathbb{E}\left(X_n - X\right)^2 &= \mathbb{E}\left(X_n^2 - 2X_nX + X^2\right) \\ &= \mathbb{E}\left(X_n^2\right) - 2\mathbb{E}\left(X_nX\right) + \mathbb{E}\left(X^2\right) \\ &= \frac{1}{n^2}\left(1 - \frac{1}{n^2}\right) + n^2\left(\frac{1}{n^2}\right) \\ &\quad - 2\left(\frac{1}{n}\mathbb{E}\left(X\right)\left(1 - \frac{1}{n^2}\right) + n\mathbb{E}\left(X\right)\left(\frac{1}{n^2}\right)\right) + \mathbb{E}\left(X^2\right)\end{aligned}$$

Now take the limit for large n to find

$$\lim_{n\to\infty} \mathbb{E}\left(X_n - X\right)^2 = 1 + \mathbb{E}\left(X^2\right)$$

Now, by our initial assumption, we must have

$$1 + \mathbb{E}\left(X^2\right) = 0$$

But this is impossible, as it would require $\mathbb{E}\left(X^2\right) = -1$, which can never be the case since $\mathbb{V}\left(X\right) \geq 0$. So $X_n \xrightarrow{\text{qm}} X$ for any X. □

Question: 5.8.5

Let $X_1, \ldots, X_n \sim \text{Bernoulli}\left(p\right)$. Prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{P}} p \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{qm}} p.$$

PROOF. First we prove that $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{qm}} p$. The expectation and mean are

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) \\ &= \frac{1}{n}np \\ &= p \\ \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}\left(X_i^2\right) \\ &= \frac{1}{n}\left[\mathbb{E}\left(X_i^4\right) - \mathbb{E}\left(X_i^2\right)^2\right] \\ &= \frac{1}{n}\left(p - p^2\right)\end{aligned}$$

So we ascertain that

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - p\right)^2 = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)$$

and that the remaining task is to prove the variance vanishes in the limit:

$$\lim_{n\to\infty} \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) = \lim_{n\to\infty} \frac{1}{n}\left(p - p^2\right) = 0$$

Thus $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{qm}} p$. By Theorem 5.4(a), it immediately follows that $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{P}} p$. □

Question: 5.8.6

Suppose that the height of men has mean 68 inches and standard deviation 2.6 inches. We draw 100 men at random. Find (approximately) the probability that the average height of men in our sample will be at least 72 inches.

Note that the errata is for 72 inches.
For $Z \sim N(0,1)$, by the Central Limit Theorem:

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n > 72\right) &= \mathbb{P}\left(\frac{\sqrt{100}\left(\overline{X}_n - 68\right)}{2.6} > \frac{\sqrt{100}\left(72 - 68\right)}{2.6}\right) \\
&\approx \mathbb{P}\left(10Z > 15.38\right) \\
&= \mathbb{P}\left(Z > 1.538\right) \\
&\approx \boxed{0.062}
\end{aligned}$$

Question: 5.8.7

Let $\lambda_n = 1/n$ for $n = 1, 2, \ldots$. Let $X_n \sim \text{Poisson}\left(\lambda_n\right)$.

(a) Show that $X_n \xrightarrow{\text{P}} 0$.

(b) Let $Y_n = nX_n$. Show that $Y_n \xrightarrow{\text{P}} 0$.

Proof. (a) Observe that

$$\lim_{n\to\infty} \mathbb{E}\left(X_n^2\right) = \lim_{n\to\infty} \mathbb{V}\left(X_n\right) + \mathbb{E}\left(X_n\right)^2 = \lim_{n\to\infty} \frac{1}{n} + \frac{1}{n^2} = 0$$

which implies $X_n \xrightarrow{\text{qm}} 0$, and by Theorem 5.4(a), we have $X_n \xrightarrow{\text{P}} 0$.

(b) By Theorem 5.5(f), if $Y_n = g\left(X_n\right) = nX_n$, then $X_n \xrightarrow{\text{P}} 0 \implies g\left(X_n\right) \xrightarrow{\text{P}} g\left(0\right) \implies Y_n \xrightarrow{\text{P}} 0$. □

Question: 5.8.8

Suppose we have a computer program consisting of $n = 100$ pages of code. Let X_i be the number of errors on the i^{th} page of code. Suppose that the X_i's are Poisson with mean 1 and that they are independent. Let $Y = \sum_{i=1}^{n} X_i$ be the total number of errors. Use the central limit theorem to approximate $\mathbb{P}(Y < 90)$.

By the Central Limit Theorem, $n\overline{X}_n \approx N\left(n\mu, n\sigma^2\right)$. Then calculate

$$\mathbb{P}(Y < 90) = \mathbb{P}\left(\frac{n\overline{X}_n - n\mu}{\sqrt{n\sigma^2}} < \frac{90 - (100)(1)}{\sqrt{100 \cdot 1}}\right) = \mathbb{P}(Z < -1) \approx \boxed{0.1587}$$

Question: 5.8.9

Suppose that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Define

$$X_n = \begin{cases} X & \text{with probability } 1 - \frac{1}{n} \\ e^n & \text{with probability } \frac{1}{n}. \end{cases}$$

Does X_n converge to X in probability? Does X_n converge to X in distribution? Does $\mathbb{E}(X - X_n)^2$ converge to 0?

Proof. First we determine if X_n converges to X in quadratic mean. Observe that

$$\begin{aligned} \mathbb{E}(X_n - X)^2 = \mathbb{E}(X - X_n)^2 &= \mathbb{E}\left(X^2 - 2XX_n + X_n^2\right) \\ &= \mathbb{E}\left(X^2\right) - 2\mathbb{E}(XX_n) + \mathbb{E}\left(X_n^2\right) \end{aligned}$$

Calculating each of the expectations yields

$$\begin{aligned}
\mathbb{E}\left(X^2\right) &= (1)^2\left(\frac{1}{2}\right) + (-1)^2 \\
\mathbb{E}\left(X_n^2\right) &= \mathbb{E}\left(X_n^2 \mid X=1\right)\mathbb{P}\left(X=1\right) + \mathbb{E}\left(X_n^2 \mid X=-1\right)\mathbb{P}\left(X=-1\right) \\
&= \left((1)^2\left(1-\frac{1}{n}\right) + e^{2n}\left(\frac{1}{n}\right)\right)\cdot\frac{1}{2} \\
&\quad + \left((-1)^2\left(1-\frac{1}{n}\right) + e^{2n}\left(\frac{1}{n}\right)\right)\cdot\frac{1}{2} \\
\mathbb{E}\left(XX_n\right) &= \mathbb{E}\left(XX_n \mid X=1\right)\mathbb{P}\left(X=1\right) + \mathbb{E}\left(XX_n \mid X=-1\right)\mathbb{P}\left(X=-1\right) \\
&= \left((1)^2\left(1-\frac{1}{n}\right) + (1)\,e^{n}\left(\frac{1}{n}\right)\right)\cdot\frac{1}{2} \\
&\quad + \left((-1)^2\left(1-\frac{1}{n}\right) + (-1)\,e^{n}\left(\frac{1}{n}\right)\right)\cdot\frac{1}{2} \\
&= 1-\frac{1}{n} \\
\mathbb{E}\left(X-X_n\right)^2 &= 1-2\left(1-\frac{1}{n}\right)+1+\frac{1}{n}\left(e^{2n}-1\right) \\
&= \frac{e^{2n}}{n}
\end{aligned}$$

Since $\lim_{n\to\infty}\mathbb{E}\left(X-X_n\right)^2 = \lim_{n\to\infty}\frac{e^{2n}}{n} \neq 0$, we have $X_n \xrightarrow{\text{qm}} X$.

Now we investigate if $X_n \xrightarrow{\text{P}} X$:

$$\begin{aligned}
\mathbb{P}\left(|X_n - X| > \epsilon\right) &= \mathbb{P}\left(0 > \epsilon \mid X_n = X\right)\mathbb{P}\left(X_n = X\right) \\
&\quad + \mathbb{P}\left(e^n - X > \epsilon \mid X_n = e^n\right)\mathbb{P}\left(X_n = e^n\right) \\
&= \mathbb{P}\left(e^n - X > \epsilon \mid X_n = e^n\right)\cdot\frac{1}{n}
\end{aligned}$$

Taking the limit, we get

$$\lim_{n\to\infty}\mathbb{P}\left(|X_n - X| > \epsilon\right) = \lim_{n\to\infty}\mathbb{P}\left(e^n - X > \epsilon \mid X_n = e^n\right)\cdot\frac{1}{n} = 0$$

proving that $X_n \xrightarrow{\text{P}} X$, and by Theorem 5.4(b), $X_n \rightsquigarrow X$. □

Question: 5.8.10

Let $Z \sim N(0,1)$. Let $t > 0$. Show that, for any $k > 0$,

$$\mathbb{P}(|Z| > t) \leq \frac{\mathbb{E}|Z|^k}{t^k}$$

Compare this to Mill's inequality in Chapter 4.

PROOF. By Markov's inequality,

$$\mathbb{P}(|Z| > t) = \mathbb{P}\left(|Z|^k > t^k\right) \leq \frac{\mathbb{E}|Z|^k}{t^k}$$

□

Question: 5.8.11

Suppose that $X_n \sim N(0, 1/n)$ and let X be a random variable with distribution $F(x) = 0$ if $x < 0$ and $F(x) = 1$ if $x \geq 0$. Does X_n converge to X in probability? (Prove or disprove). Does X_n converge to X in distribution? (Prove or disprove).

PROOF. We have $X \sim F$ where $F(x)$ is defined as:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

which is the Heaviside function. Meanwhile, X_n has PDF

$$f_n(x_n) = \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{x_n^2 n}{2}\right)$$

As $n \to 0$, $\lim_{n\to\infty} f_n(x_n) = 0$ when $x_n \neq 0$. At $x_n = 0$, $\lim_{n\to\infty} f_n(x_n)$ diverges to infinity. In other words, this is the Dirac delta function:

$$\lim_{n\to\infty} f_n(x_n) = \delta(x) = \begin{cases} +\infty & x = 0 \\ 0 & x \neq 0 \end{cases}$$

Then it follows that the distribution

$$F_n(t) = \sqrt{\frac{n}{2\pi}} \int_{-\infty}^{t} \exp\left(-\frac{x_n^2 n}{2}\right) \, dx_n$$

tends toward

$$\lim_{n\to\infty} F_n(t) = \int_{-\infty}^{t} \delta(x) \, dx = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases}$$

Thus we have

$$\lim_{n\to\infty} F_n(t) = F(t) \quad \Longrightarrow \quad X_n \rightsquigarrow X$$

Next, since $F(x)$ is the Heaviside function, X must also have a Dirac delta PDF (there are nuances here with respect to Lebesgue integrability but that is outside the scope of this textbook). We examine if X_n converges to X in L_1, assuming $X_n - X \geq 0$ (we will get the same result with $X_n - X < 0$ anyhow):

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}|X_n - X| &= \lim_{n\to\infty} \mathbb{E}(X_n - X) = \lim_{n\to\infty} (\mathbb{E}(X_n) - \mathbb{E}(X)) \\
&= \lim_{n\to\infty} \int_{-\infty}^{\infty} x_n \sqrt{\frac{n}{2\pi}} \exp\left(-\frac{x_n^2 n}{2}\right) \, dx_n - \int_{-\infty}^{\infty} x\delta(x) \, dx \\
&= \int_{-\infty}^{\infty} x'\delta(x') \, dx' - \int_{-\infty}^{\infty} x\delta(x) \, dx \\
&= 0
\end{aligned}$$

This proves that $X_n \xrightarrow{L_1} X$, which implies $X_n \xrightarrow{\mathrm{P}} X$. □

Question: 5.8.12

Let $X, X_1, X_2, X_3, \ldots$ be random variables that are positive and integer valued. Show that $X_n \rightsquigarrow X$ if and only if

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$$

for every integer k.

Proof. ($\implies$) Let $X_n \rightsquigarrow X$. Then

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

where $F_n(k) = \mathbb{P}(X_n \leq k)$ and $F(t) = \mathbb{P}(X \leq k)$ for any positive integer k. Beginning from $k = 1$, we must have

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

Then for $k = 2$,

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq 2) = \mathbb{P}(X \leq 2)$$

which implies that $\lim_{n\to\infty} \mathbb{P}(X_n = 2) = \mathbb{P}(X = 2)$, as we ascertained that $\lim_{n\to\infty} \mathbb{P}(X_n = 1) = \mathbb{P}(X = 1)$. We can continue in this fashion to conclude

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$$

($\impliedby$) Let $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(X = k)$ for every positive integer k. If we observe that

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq k) = \lim_{n\to\infty} \sum_{i=1}^{k} \mathbb{P}(X_n = i)$$

$$\mathbb{P}(X \leq k) = \sum_{i=1}^{k} \mathbb{P}(X \leq i)$$

Then we can conclude

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq k) = \mathbb{P}(X \leq k)$$

which is equivalent to claiming

$$\lim_{n\to\infty} F_n(k) = F(k) \quad \Longrightarrow \quad X_n \rightsquigarrow X$$

□

Question: 5.8.13

Let $Z_1, Z_2, \ldots$ be IID random variables with density f. Suppose that $\mathbb{P}(Z_i > 0) = 1$ and that $\lambda = \lim_{x\downarrow 0} f(x) > 0$. Let

$$X_n = n \min\{Z_1, \ldots, Z_n\}.$$

Show that $X_n \rightsquigarrow Z$ where Z has an exponential distribution with mean $1/\lambda$.

Proof. First, recall that the CDF of $Z \sim \exp(1/\lambda)$ is

$$F(t) = \int_0^t \lambda \exp(-\lambda x)\ dx = 1 - \exp(-\lambda t)$$

Here's our intuition: does $\mathbb{P}(X_n \leq t)$ follow a pattern that allows us to derive an exponential function by way of a limit?

For $X_n \leq t$, we must have $\min\{Z_1, \ldots, Z_n\} \leq t/n$. To guarantee this is the case, we must guarantee that *all* the $Z_i \leq t/n$ for all i. Then since the random variables are IID, it follows that

$$\begin{aligned}
\mathbb{P}(X_n \leq t) &= \prod^n \mathbb{P}\left(Z_i \leq \frac{t}{n}\right) \\
&= 1 - \prod^n \mathbb{P}\left(Z_i \geq \frac{t}{n}\right) \\
&= 1 - \prod^n \left(\mathbb{P}(Z_i > 0) - \mathbb{P}\left(Z_i \leq \frac{t}{n}\right)\right) \\
&= 1 - \prod^n \left(1 - \mathbb{P}\left(Z_i \leq \frac{t}{n}\right)\right)
\end{aligned}$$

Now, for $n \to \infty$, for fixed t, $t/n \to 0$. By premise, $\lim_{x\downarrow 0} f(x) = \lambda$. For infinitesimal t/n, we can approximate $\mathbb{P}(Z_i \leq t/n) = \int_0^{t/n} f\ dz_i$ as follows:

$$\lim_{n\to\infty} \mathbb{P}\left(Z_i \leq \frac{t}{n}\right) = \lim_{n\to\infty} \int_0^{t/n} f\ dz_i = \frac{\lambda t}{n}$$

where we can think of $f \approx \lambda$ and $\Delta z_i \approx t/n$. Now we can finish with

$$\begin{aligned}
\lim_{n\to\infty} F_n(t) &= \lim_{n\to\infty} \mathbb{P}(X_n \leq t) \\
&= 1 - \lim_{n\to\infty} \prod^{n} \left(1 - \mathbb{P}\left(Z_i \leq \frac{t}{n}\right)\right) \\
&= 1 - \lim_{n\to\infty} \prod^{n} \left(1 - \frac{\lambda t}{n}\right) \\
&= 1 - \lim_{n\to\infty} \left(1 - \frac{\lambda t}{n}\right)^{n} \\
&= 1 - \exp(-\lambda t) \\
&= F(t)
\end{aligned}$$

□

Question: 5.8.14

Let $X_1, \ldots, X_n \sim \text{Uniform}(0,1)$. Let $Y_n = \overline{X}_n^2$. Find the limiting distribution of Y_n.

Recall that for $X_1 \sim \text{Uniform}(0,1)$, $\mathbb{E}(X_1 = 1/2)$ and $\mathbb{V}(X_1) = 1/12$. Then $\mathbb{E}(\overline{X}_n) = 1/2$ and $\mathbb{V}(\overline{X}_n) = 1/12n$. By Theorem 5.13 (the Delta Method), since $\overline{X}_n \approx N(1/2, 1/12n)$, we have

$$Y_n = \overline{X}_n^2 \approx N\left(\frac{1}{4}, \frac{1}{12n}\right)$$

Question: 5.8.15

Let

$$\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix}, \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix}, \ldots, \begin{pmatrix} X_{1n} \\ X_{2n} \end{pmatrix}$$

be IID random vectors with mean $\mu = (\mu_1, \mu_2)$ and variance Σ. Let

$$\overline{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}, \quad \overline{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$$

and define $Y_n = \overline{X}_1/\overline{X}_2$. Find the limiting distribution of Y_n.

Let $Y_n = g(\overline{X}_1, \overline{X}_2)$ where $g(s_1, s_2) = s_1/s_2$. The gradient is

$$\nabla g(s) = \begin{pmatrix} \frac{\partial g}{\partial s_1} \\ \frac{\partial g}{\partial s_2} \end{pmatrix} = \begin{pmatrix} \frac{1}{s_2} \\ -\frac{s_1}{s_2^2} \end{pmatrix}$$

Then by the multivariate Delta Method,

$$\begin{aligned}\nabla_\mu^T \Sigma \nabla_\mu &= \begin{pmatrix} 1/\mu_2 & -\mu_1/\mu_2^2 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \begin{pmatrix} 1/\mu_2 \\ -\mu_1/\mu_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{\sigma_{11}}{\mu_2} - \frac{\mu_1 \sigma_{12}}{\mu_2^2} & \frac{\sigma_{12}}{\mu_2} - \frac{\mu_1 \sigma_{22}}{\mu_2^2} \end{pmatrix} \begin{pmatrix} 1/\mu_2 \\ -\mu_1/\mu_2^2 \end{pmatrix} \\ &= \frac{\sigma_{11}}{\mu_2^2} - \frac{2\mu_1\sigma_{12}}{\mu_2^3} + \frac{\mu_1^2 \sigma_{22}}{\mu_2^4}\end{aligned}$$

Thus we can conclude

$$\sqrt{n}\left(\begin{pmatrix} \overline{X}_1 \\ \overline{X}_2 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right) \rightsquigarrow N\left(0, \frac{\sigma_{11}}{\mu_2^2} - \frac{2\mu_1\sigma_{12}}{\mu_2^3} + \frac{\mu_1^2\sigma_{22}}{\mu_2^4}\right)$$

Question: 5.8.16

Construct an example where $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ but $X_n + Y_n$ does not converge in distribution to $X + Y$.

Define $X, Y \sim N(0,1)$, $X_n = X$, $Y_n = -X$, $Z = X + Y$, and $Z_n = X_n + Y_n$. It follows that $Z \sim N(0,2)$. Moreover, we satisfy the premise that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$; it is trivial in the former case, and for the latter, simply observe that

$$f_X(-x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-x)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) = f_X(x)$$

Since $Z_n = X_n + Y_n = 0$ is a constant random variable, its distribution is

$$F_n(Z_n) = \begin{cases} 0 & Z_n < 0 \\ 1 & Z_n \geq 0 \end{cases}$$

demonstrating that $Z_n \not\rightsquigarrow Z$.

Chapter 6: Models, Statistical Inference and Learning

Question: 6.6.1

Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$ and let $\widehat{\lambda} = n^{-1}\sum_{i=1}^n X_i$. Find the bias, se, and MSE of this estimator.

The bias, se, and MSE are given by

$$\begin{aligned}
\text{bias}(\widehat{\lambda}) &= \mathbb{E}(\widehat{\lambda}) - \lambda \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) - \lambda \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) - \lambda \\
&= \frac{1}{n}(n\lambda) - \lambda \\
&= \boxed{0} \\
\mathbb{V}\left(\widehat{\lambda}\right) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) \\
&= \frac{1}{n^2}(n\lambda) \\
&= \frac{\lambda}{n} \\
\text{se}\left(\widehat{\lambda}\right) &= \boxed{\sqrt{\lambda/n}} \\
\text{MSE} &= \text{bias}^2\left(\widehat{\lambda}\right) + \mathbb{V}\left(\widehat{\lambda}\right) \\
&= 0 + \lambda/n \\
&= \boxed{\lambda/n}
\end{aligned}$$

Question: 6.6.2

Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = \max\{X_{1,}, \ldots, X_n\}$. Find the bias, se, and MSE of this estimator.

Note that $\widehat{\theta} \leq \alpha$ if and only if $X_i \leq \alpha$ for all $i = 1, \ldots, n$. First we derive the CDF:

$$F_{\widehat{\theta}}(\alpha) = \mathbb{P}\left(\widehat{\theta} \leq \alpha\right) = \prod_{i=1}^n \int_0^\alpha \frac{1}{\theta}\, dx = \left(\frac{\alpha}{\theta}\right)^n$$

Differentiating gives us the PDF:

$$f_{\widehat{\theta}}(\alpha) = \frac{n}{\theta^n}\alpha^{n-1}$$

Now we derive the bias:

$$\begin{aligned}
\mathsf{bias}\left(\widehat{\theta}\right) &= \mathbb{E}\left(\widehat{\theta}\right) - \theta \\
&= \int_0^\theta \frac{n}{\theta^n}\alpha^n \, d\alpha - \theta \\
&= \boxed{\theta\left(\frac{n}{n+1} - 1\right)}
\end{aligned}$$

Observe that

$$\lim_{n\to\infty} \mathsf{bias}\left(\widehat{\theta}\right) = \lim_{n\to\infty} \theta\left(\frac{n}{n+1} - 1\right) = 0$$

The variance and se:

$$\begin{aligned}
\mathbb{V}\left(\widehat{\theta}\right) &= \int_0^\theta \frac{n}{\theta^n}\alpha^{n+1} \, d\alpha - \theta^2\left(\frac{n}{n+1}\right)^2 \\
&\quad \theta^2\left(\frac{n}{(n+2)(n+1)^2}\right) \\
\mathsf{se}\left(\widehat{\theta}\right) &= \boxed{\theta\left(\frac{n}{(n+2)(n+1)^2}\right)^{1/2}}
\end{aligned}$$

Again taking the limit, we find

$$\lim_{n\to\infty} \mathsf{se}\left(\widehat{\theta}\right) = \lim_{n\to\infty} \theta\left(\frac{n}{(n+2)(n+1)^2}\right)^{1/2} = 0$$

Lastly, the MSE:

$$\begin{aligned}
\mathsf{MSE} &= \mathsf{bias}^2\left(\widehat{\theta}\right) + \mathbb{V}\left(\widehat{\theta}\right) \\
&= \left(\theta\left(\frac{n}{n+1} - 1\right)\right)^2 + \theta\left(\frac{n}{(n+2)(n+1)^2}\right) \\
&= \boxed{\theta^2\left(\frac{2n+2}{(n+2)(n+1)^2}\right)}
\end{aligned}$$

with limit

$$\lim_{n\to\infty} \mathsf{MSE} = \lim_{n\to\infty} \theta^2\left(\frac{2n+2}{(n+2)(n+1)^2}\right) = 0$$

Question: 6.6.3

Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$ and let $\widehat{\theta} = 2\overline{X}_n$. Find the bias, se, and MSE of this estimator.

Calculate the bias, se, and MSE as:

$$\begin{aligned}
\text{bias}\left(\widehat{\theta}\right) &= \mathbb{E}\left(\widehat{\theta}\right) - \theta \\
&\mathbb{E}\left(2\overline{X}_n\right) - \theta \\
&= 2\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \theta \\
&= 2\left(\frac{\theta}{2}\right) - \theta \\
&= \boxed{0} \\
\mathbb{V}\left(\widehat{\theta}\right) &= \mathbb{V}\left(2\overline{X}_n\right) = 4\mathbb{V}\left(\overline{X}_n\right) \\
&= 4\mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{\theta^2}{3n} \\
\text{se}\left(\widehat{\theta}\right) &= \boxed{\theta/\sqrt{3n}} \\
\text{MSE} &= \text{bias}^2\left(\widehat{\theta}\right) + \mathbb{V}\left(\widehat{\theta}\right) \\
&= 0 + \theta^2/3n \\
&= \boxed{\theta^2/3n}
\end{aligned}$$

Chapter 7: Estimating the CDF and Statistical Functionals

Import Packages

Please see the associated GitHub repo for all code and comments to simulations. [LINK HERE]

```
import numpy as np
from numpy.random import choice
import random
import matplotlib.pyplot as plt
import scienceplots
```

Question: 7.4.1

Prove Theorem 7.3.

Theorem (Wasserman 7.3)

At any fixed value of x,

$$\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= F(x), \\
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \frac{F(x)(1-F(x))}{n}, \\
\mathsf{MSE} &= \frac{F(x)(1-F(x))}{n} \to 0, \\
\widehat{F}_n(x) &\xrightarrow{\mathrm{P}} F(x).
\end{aligned}$$

Proof. $\underline{\mathbb{E}\left(\widehat{F}_n(x)\right) = F(x)}$:

By the definition of $\widehat{F}_n(x)$, we have

$$\begin{aligned}
\mathbb{E}\left(\widehat{F}_n(x)\right) &= \mathbb{E}\left(\frac{\sum_{i=1}^n I(X_i \leq x)}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{P}(X_i \leq x) \\
&= \frac{1}{n}(nF(x)) \\
&= F(x)
\end{aligned}$$

$\underline{\mathbb{V}\left(\widehat{F}_n(x)\right) = F(x)(1-F(x))/n}$:

$$\begin{aligned}
\mathbb{V}\left(\widehat{F}_n(x)\right) &= \mathbb{E}\left(\widehat{F}_n^2(x)\right) - \mathbb{E}\left(\widehat{F}_n(x)\right)^2 \\
&= \mathbb{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} I(X_i \leq x)\right)^2\right) - F(X)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{j=1}^{n}\sum_{i=1}^{n} I(X_j \leq x) I(X_i \leq x)\right) - F(x)^2 \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{P}(X_i \leq x) + \sum_{j=1}^{n}\sum_{\substack{i=1 \\ i\neq j}}^{n}\mathbb{P}(X_j \leq x)\mathbb{P}(X_i \leq x)\right) - F(x)^2 \\
&= \frac{1}{n^2}\left(nF(x) + n(n-1)F(x)^2\right) - F(x)^2 \\
&= \frac{F(x) + (n-1)F(x)^2 - nF(x)^2}{n} \\
&= \frac{F(x)(1-F(x))}{n}
\end{aligned}$$

$\underline{\mathsf{MSE} = F(x)(1-F(x))/n \to 0:}$

$$\begin{aligned}
\mathsf{MSE} &= \mathsf{bias}^2\left(\widehat{F}_n(x)\right) + \mathbb{V}\left(\widehat{F}_n(x)\right) \\
&= \frac{F(x)(1-F(x))}{n} \to 0 \text{ as } n \to \infty
\end{aligned}$$

$\underline{\widehat{F}_n(x) \xrightarrow{\mathrm{P}} F(x):}$
Since

$$\mathbb{E}\left(\widehat{F}_n(x) - F(x)\right)^2 = \mathbb{V}\left(\widehat{F}_n(x)\right) \to 0 \text{ as } n \to \infty$$

It follows that $\widehat{F}_n(x) \xrightarrow{\mathrm{qm}} F(x)$, implying $\widehat{F}_n(x) \xrightarrow{\mathrm{P}} F(x)$. □

Question: 7.4.2

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $Y_1, \ldots, Y_m \sim \text{Bernoulli}(q)$. Find the plug-in estimator and estimated standard error for p. Find an approximate 90 percent confidence interval for p. Find the plug-in estimator and estimated standard error for $p - q$. Find an approximate 90 percent confidence interval for $p - q$.

By Theorem 7.9,

$$\widehat{p} = \int x d\widehat{F}_n(x) = \boxed{\overline{X}_n}$$

$$\widehat{\mathsf{se}} = \widehat{\sigma}/\sqrt{n} = \boxed{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}}$$

The 90 percent confidence interval for p has $z_{\alpha/2} = z_{.1/2} = 1.65$, so we have

$$\boxed{\overline{X}_n \pm 1.65\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}}$$

For $p - q$, we have

$$\widehat{p-q} = \int x d\widehat{F}_n(x) - \int y d\widehat{F}_n(y) = \boxed{\overline{X}_n - \overline{Y}_n}$$

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{\sigma_p^2}}{n} + \frac{\widehat{\sigma_q^2}}{m}} = \boxed{\sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n} + \frac{\widehat{q}\,(1-\widehat{q})}{m}}}$$

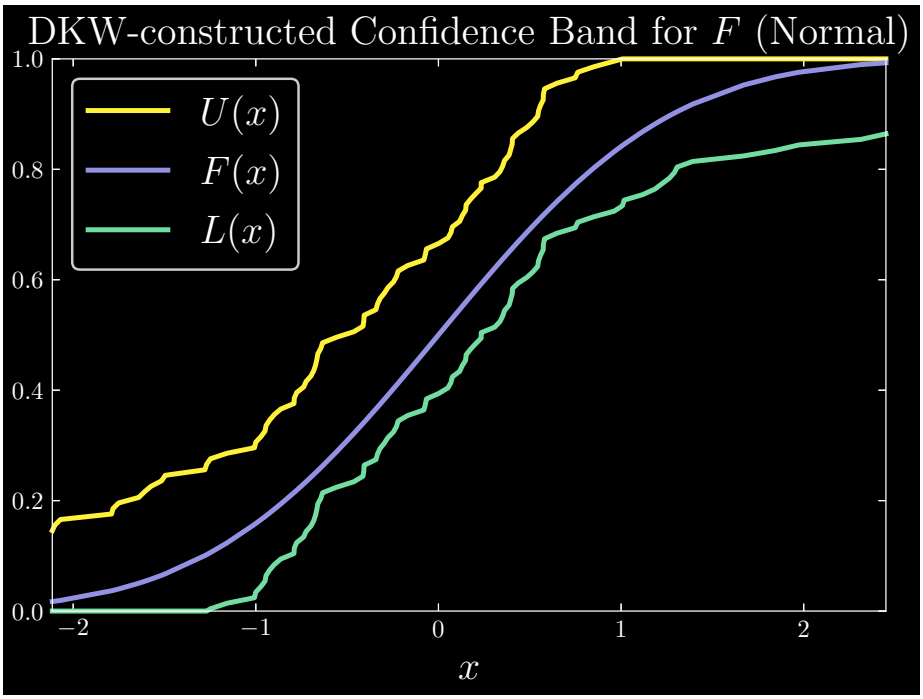
$$90 \text{ percent CI: } = \boxed{\overline{X}_n - \overline{Y}_n \pm 1.65\widehat{\mathsf{se}}}$$

Question: 7.4.3

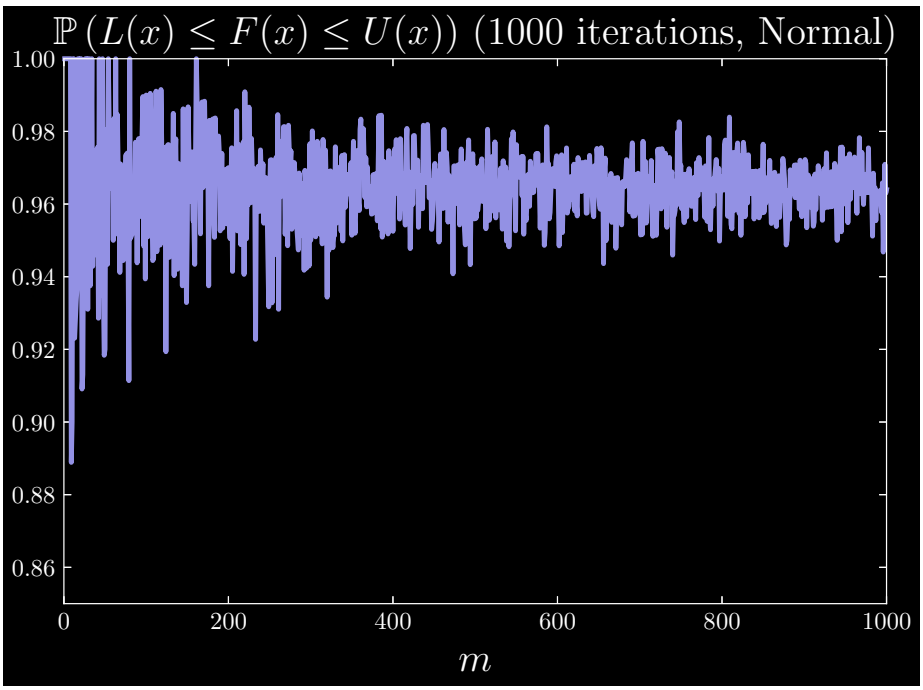
(Computer Experiment.) Generate 100 observations from a $N(0,1)$ distribution. Compute a 95 percent confidence band for the CDF F. Repeat this 1000 times and see how often the confidence band contains the true distribution function. Repeat using data from a Cauchy distribution.

Using the DKW inequality, we can construct the confidence band from a standard normal distribution in the following manner:

```
n = 100
alpha = 0.05
eps = np.sqrt((1/(2*n))*np.log(2/alpha))

def ecdfhat(sample):
    cdfhat = [ i/len(sample) for i in range(1, len(sample)+1)]
    return cdfhat;

def dkwintnormal():
    dkwtest, lowerbound, upperbound =
        np.empty(n), np.empty(n), np.empty(n)
    sample = np.sort(np.random.normal(0,1,n))
    ecdf = ecdfhat(sample)
    for i,j in zip(range(0, len(sample)), sample):
        lowerbound[i] = max(ecdf[i] - eps, 0)
        upperbound[i] = min(ecdf[i] + eps, 1)
        dkwtest[i] =
                (norm.cdf(j) >= lowerbound[i])
                & (norm.cdf(j) <= upperbound[i])
    return dkwtest, lowerbound, upperbound, ecdf, sample;
```

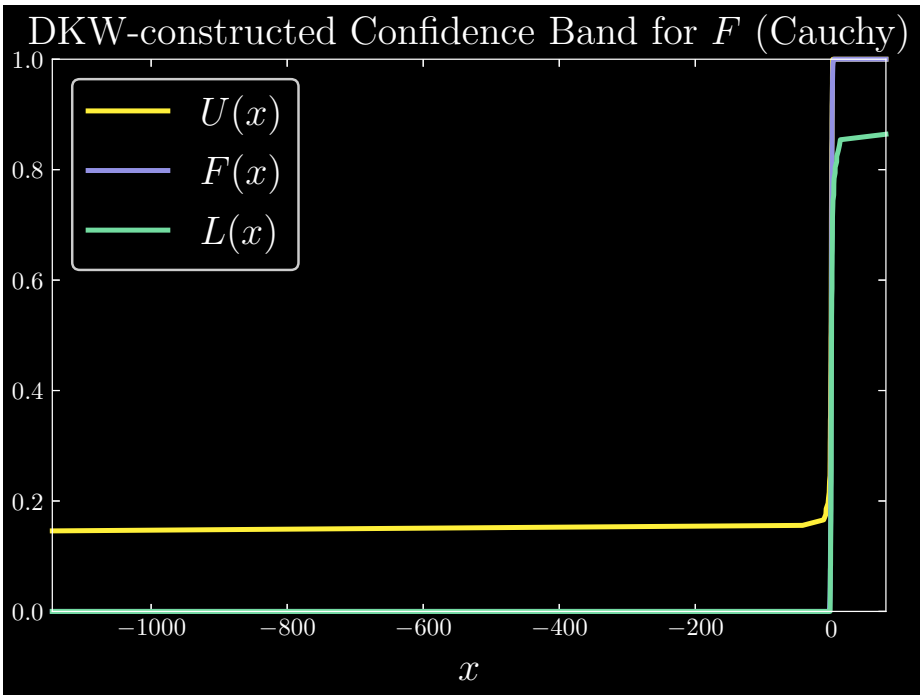


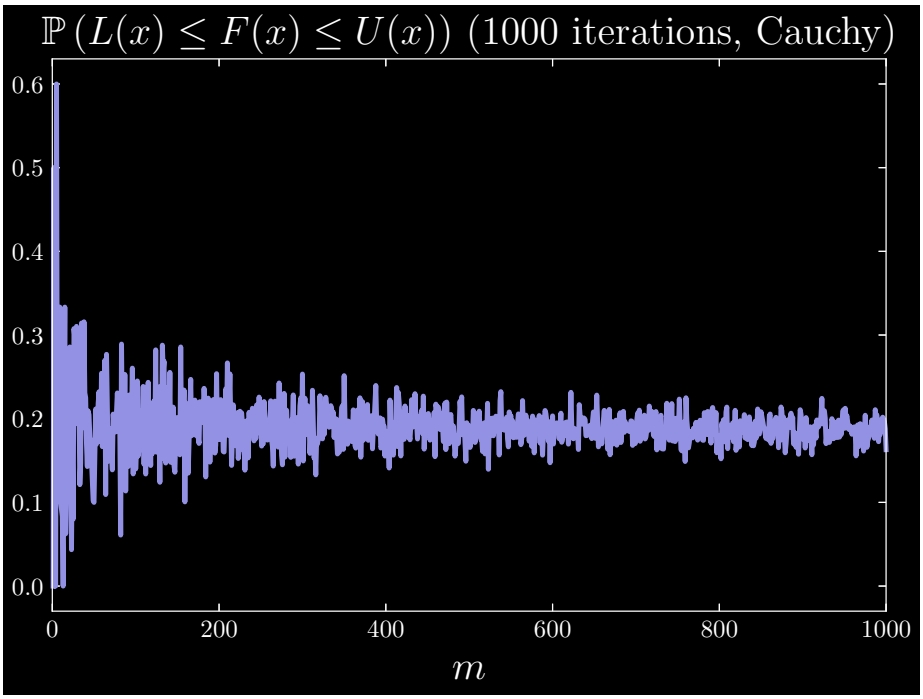
The probability that the CDF F is contained in the confidence band approaches 0.95 for large number of iterations:



Repeating the same for data from a standard Cauchy distribution, we get:

```
n = 100
alpha = 0.05
eps = np.sqrt((1/(2*n))*np.log(2/alpha))

def ecdfhat(sample):
    cdfhat = [ i/len(sample) for i in range(1, len(sample)+1)]
    return cdfhat;

def dkwintcauchy():
    dkwtest, lowerbound, upperbound =
        np.empty(n), np.empty(n), np.empty(n)
    sample = np.sort(np.random.standard_cauchy(n))
    ecdf = ecdfhat(sample)
    for i,j in zip(range(0, len(sample)), sample):
        lowerbound[i] = max(ecdf[i] — eps, 0)
        upperbound[i] = min(ecdf[i] + eps, 1)
        dkwtest[i] =
                (norm.cdf(j) >= lowerbound[i])
                & (norm.cdf(j) <= upperbound[i])
    return dkwtest, lowerbound, upperbound, ecdf, sample;
```





Question: 7.4.4

Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}_n(x)$ be the empirical distribution function. For a fixed x, use the central limit theorem to find the limiting distribution of $\widehat{F}_n(x)$.

Using Theorem 7.3 and the proof from exercise 7.4.1, by the Central Limit Theorem, we have

$$\widehat{F}_n(x) \approx N\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

Question: 7.4.5

Let x and y be two distinct points. Find $\mathsf{Cov}\left(\widehat{F}_n(x), \widehat{F}_n(y)\right)$.

By definition of covariance, we have $\mathsf{Cov}\left(\widehat{F}_n(x), \widehat{F}_n(y)\right)$ equal to:

$$
\begin{aligned}
&= \mathbb{E}\left(\left(\widehat{F}_n(x) - \mathbb{E}\left(\widehat{F}_n(x)\right)\right)\left(\widehat{F}_n(y) - \mathbb{E}\left(\widehat{F}_n(y)\right)\right)\right) \\
&= \mathbb{E}\left(\left(\widehat{F}_n(x) - F(x)\right)\left(\widehat{F}_n(y) - F(y)\right)\right) \\
&= \mathbb{E}\left(\widehat{F}_n(x)\,\widehat{F}_n(y)\right) - F(x)\,F(y) \\
&= \mathbb{E}\left(\frac{1}{n^2}\left(\sum_{i=1}^{n} I(X_i \leq x)\right)\left(\sum_{j=1}^{n} I(X_j \leq y)\right)\right) - F(x)\,F(y) \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n} I(X_i \leq x)\,I(X_i \leq y) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} I(X_i \leq x)\,I(X_j \leq y)\right) \\
&\quad - F(x)\,F(y)
\end{aligned}
$$

Now, for the first term to have non-zero terms in the sum, we must have $X_i \leq x, y$. Since x, y are distinct, either $x < y$ or $y < x$. Suppose without loss of generality that $x < y$. Then the first term resolves to

$$
\mathbb{E}\left(\sum_{i=1}^{n} I(X_i \leq x)\,I(X_i \leq y)\right) = n\mathbb{P}(X_i \leq y) = nF(y)
$$

For the second term, assuming independence, we have

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} I(X_i \leq x)\,I(X_j \leq y)\right) &= n(n-1)\,\mathbb{P}(X_i \leq x)\,\mathbb{P}(X_j <= y) \\
&= n(n-1)\,F(x)\,F(y)
\end{aligned}
$$

Combining the terms, we can conclude that the covariance is equal to

$$
\begin{aligned}
&= \frac{1}{n^2}\left(nF(y) + n(n-1)\,F(x)\,F(y)\right) - F(x)\,F(y) \\
&= \frac{F(y) + (n-1)\,F(x)\,F(y) - nF(x)\,F(y)}{n} \\
&= \boxed{\frac{F(y)\,(1 - F(x))}{n}}
\end{aligned}
$$

Question: 7.4.6

Let $X_1, \ldots, X_n \sim F$ and let $\widehat{F}$ be the empirical distribution function. Let $a < b$ be fixed numbers and define $\theta = T(F) = F(b) - F(a)$. Let $\widehat{\theta} = T\left(\widehat{F}_n\right) = \widehat{F}_n(b) - \widehat{F}_n(a)$. Find the estimated standard error of $\widehat{\theta}$. Find an expression for an approximate $1-\alpha$ confidence interval for θ.

Using exercise 7.4.5, we can derive

$$\begin{aligned}
\mathbb{V}\left(\widehat{\theta}\right) &= \mathbb{V}\left(\widehat{F}_n(b) - \widehat{F}_n(a)\right) \\
&= \mathbb{V}\left(\widehat{F}_n(b)\right) + \mathbb{V}\left(\widehat{F}_n(a)\right) + 2\,\mathsf{Cov}\left(\widehat{F}_n(b), \widehat{F}_n(a)\right) \\
&= \frac{1}{n}\left(F(b)(1-F(b)) + F(a)(1-F(a)) + 2F(b)(1-F(a))\right) \\
&= \frac{1}{n}\left(F(b) - F(a)\right)\left(1 - \left(F(b) - F(a)\right)\right) \\
\mathsf{se}\left(\widehat{\theta}\right) &= \sqrt{\mathbb{V}\left(\widehat{\theta}\right)} \\
\widehat{\mathsf{se}}\left(\widehat{\theta}\right) &= \sqrt{\frac{1}{n}\left(\widehat{F}(b) - \widehat{F}(a)\right)\left(1 - \left(\widehat{F}(b) - \widehat{F}(a)\right)\right)}
\end{aligned}$$

The $1-\alpha$ confidence interval is given by

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\mathsf{se}}\left(\widehat{\theta}\right)$$

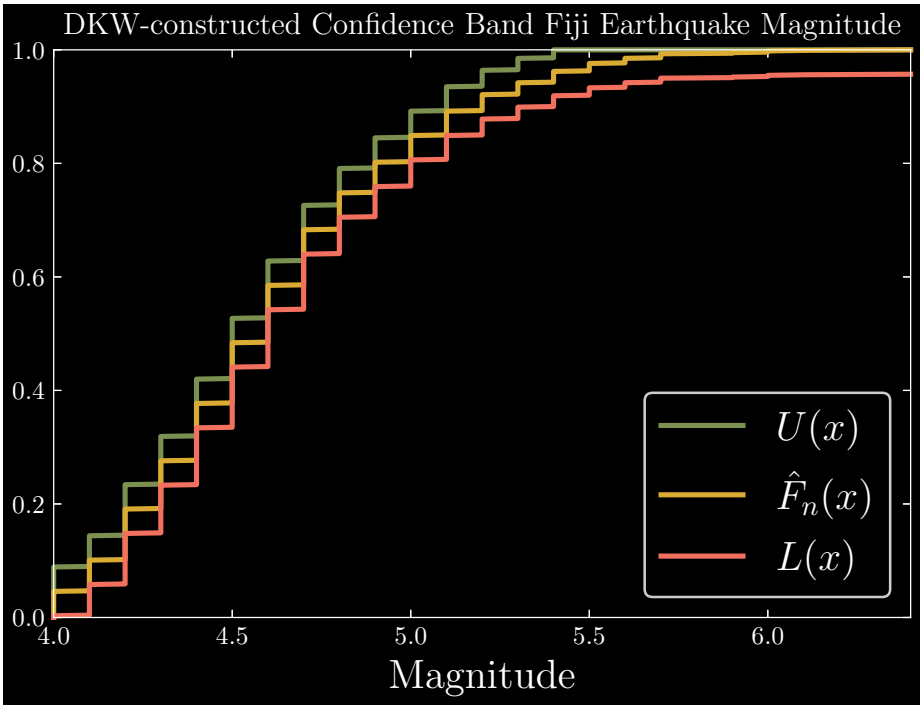
Question: 7.4.7

Data on the magnitudes of earthquakes near Fiji are available on the website for this book. Estimate the CDF $F(x)$. Compute and plot a 95 percent confidence envelope for F. Find an approximate 95 percent confidence interval for $F(4.9) - F(4.3)$.

The 95 percent confidence interval for $F(4.9) - F(4.3)$ is $(0.495, 0.557)$. The DKW-constructed confidence band for F is as follows:

```
fijidata = pd.read_csv('fijiquakes.dat', delim_whitespace=True)

magdata = np.sort(fijidata['mag'])

n = len(magdata)
alpha = 0.05
eps = np.sqrt((1/(2*n))*np.log(2/alpha))

def ecdfhat(sample):
    cdfhat = [ i/len(sample) for i in range(1, len(sample)+1)]
    return cdfhat;
```

```
def dkwbounds():
    lowerbound, upperbound =
        np.empty(len(magdata)),
        np.empty(len(magdata))
    ecdf = ecdfhat(magdata)
    for i in range(0, len(magdata)):
        lowerbound[i] = max(ecdf[i] — eps, 0)
        upperbound[i] = min(ecdf[i] + eps, 1)
    return lowerbound, upperbound;

# Use results of exercise 7.4.6 for confidence interval

theta = ecdf(4.9) — ecdf(4.3)
sehat = np.sqrt((1/len(magdata))*theta*(1—theta))
```



Question: 7.4.8

Get the data on eruption times and waiting times between eruptions of the Old Faithful geyser from the website. Estimate the mean waiting time and give a standard error for the estimate. Also, give a 90 percent confidence interval for the mean waiting time. Now estimate the median waiting time. In the next chapter we will see how to get the standard error for the median.

The estimated mean, standard error, 90 percent confidence interval, and median are 70.897, 0.824, $(69.537, 72.257)$, and 76.0, respectively.

```
faithful = pd.read_csv('faithful.dat', delim_whitespace=True)

waitingdata = faithful['waiting']

waitingmean = np.mean(waitingdata)
waitingse = np.std(waitingdata, ddof=1) / np.sqrt(np.size(waitingdata))

waitingci = (round(waitingmean - 1.65*waitingse, 3),
             round(waitingmean + 1.65*waitingse, 3))

waitingmedian = np.median(waitingdata)
```

Question: 7.4.9

100 people are given a standard antibiotic to treat an infection and another 100 are given a new antibiotic. In the first group, 90 people recover; in the second group, 85 people recover. Let p_1 be the probability of recovery under the standard treatment and let p_2 be the probability of recovery under the new treatment. We are interested in estimating $\theta = p_1 - p_2$. Provide an estimate, standard error, an 80 percent confidence interval, and a 95 percent confidence interval for θ.

We can model the odds of recovering as a Bernoulli variable. We estimate $\widehat{\theta}$ as

$$\widehat{\theta} = \widehat{p_1 - p_2} = \frac{90}{100} - \frac{85}{100} = \boxed{0.05}$$

The estimate of the standard error is

$$\begin{aligned}
\widehat{\mathsf{se}}\left(\widehat{\theta}\right) &= \sqrt{\frac{\widehat{p}_1\left(1-\widehat{p}_1\right)}{n} + \frac{\widehat{p}_2\left(1-\widehat{p}_2\right)}{n}} \\
&= \sqrt{\frac{90/100\left(1-90/100\right)}{100} + \frac{85/100\left(1-85/100\right)}{100}} \\
&\approx \boxed{0.0466}
\end{aligned}$$

80 percent CI: $z_{\alpha/2} = z_{.2/2} = z_{0.1} \approx 1.28$

$$0.05 \pm 0.0597$$

95 percent CI: $z_{\alpha/2} = z_{0.05/2} = z_{0.025} \approx 1.96$

$$0.05 \pm 0.0914$$

Question: 7.4.10

In 1975, an experiment was conducted to see if cloud seeding produced rainfall. 26 clouds were seeded with silver nitrate and 26 were not. The decision to seed or not was made at random.
Let θ be the difference in the mean precipitation from the two groups. Estimate θ. Estimate the standard error of the estimate and produce a 95 percent confidence interval.

The estimates for θ, standard error, and the 95 percent confidence interval are 277.396, 136.124, and $(10.593, 544.199)$, respectively.

```
clouds = pd.read_csv('clouds.dat', delim_whitespace=True)

unseeded = clouds['Unseeded_Clouds']
seeded = clouds['Seeded_Clouds']

diffmean = np.abs(np.mean(unseeded) - np.mean(seeded))
unseeded_se = np.std(unseeded) / np.sqrt(np.size(unseeded))
seeded_se = np.std(seeded) / np.sqrt(np.size(seeded))
diffse = np.sqrt(np.square(unseeded_se) + np.square(seeded_se))

diffci = (round(diffmean - 1.96*diffse, 3),
          round(diffmean + 1.96*diffse, 3))
```

Chapter 8: The Bootstrap

Question: 8.6.1

Consider the data in Example 8.6. Find the plug-in estimate of the correlation coefficient. Estimate the standard error using the bootstrap. Find a 95 percent confidence interval using the Normal, pivotal, and percentile methods.

Question: 8.6.4

Let $X_1, \ldots, X_n$ be distinct observations (no ties). Show that there are

$$\binom{2n-1}{n}$$

distinct bootstrap samples.
Hint: Imagine putting n balls into n buckets.

PROOF. First, let us establish some conceptual points to clarify the hint. Suppose we have three unique observations X_1, X_2, X_3. Then we gather our bootstrap samples, obtained *with* replacement. The set of all samples we can get is (using shorthand to denote that the first, second, and third digits in the sequence correspond to X_1^*, X_2^*, X_3^* equal to the first, second, or third observation):

111	112	123
222	113	
333	221	
	223	
	331	
	332	

Now, consider 3 (identical) balls with 3 bins. One way we could place the balls in the bins is to put all of them in bin 1. Or two in bin 1, and one in bin 2. One in each bin is also an option. In other words, the combinatorial problem of how to place n balls in n bins is precisely the equivalent conundrum to resolve. The three bootstrap samples are the balls, and the specific choice of observation (X_1, X_2, X_3) are analogous to the choice of bin.

Developing the combinatorial theory to complete the proof is beyond the scope here.[1] In short, the number of ways to fit k identical balls into n bins labeled $1, \ldots, n$ is

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

For $k = n$ balls, we have

$$\binom{2n-1}{n} = \frac{(2n-1)!}{n!\,(n-1)!}$$

□

[1] See Combinatorics: Discrete Mathematics and its Applications, Nicholas A. Loehr