Larry Wasserman, All of Statistics

Telmo Correa | edited by Patrick Toche

Revised: October 25, 2024

Abstract

Telmo Correa's Jupyter notebooks were downloaded from his Github repository, converted to LaTeX and adapted for personal use. Source: https://github.com/telmo-correa/all-of-statistics Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics).

1 2. Probability

1.1 2.2 Sample Spaces and Events

The **sample space** Ω is the set of possible outcomes of an experiment. Points ω in Ω are called **sample outcomes** or **realizations**. **Events** are subsets of Ω.

Given an event A, let $A^c = \{\omega \in \Omega : \text{not } (\omega \in A)\}$ denote the complement of A. The complement of Ω is the empty set $\varnothing$. The union of events A and B is defined as $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$. If $A_1, A_2, \ldots$ is a sequence of sets, then

$$\cup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for some } i\}$$

The intersection of A and B is $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. If $A_1, A_2, \ldots$ is a sequence of sets then

$$\cap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for all } i\}$$

Let $A - B = \{\omega \in \Omega : \omega \in A \text{ and not } (\omega \in B)\}$. If every element of A is contained in B we write $A \subset B$ or $B \supset A$. If A is a finite set, let $|A|$ denote the number of elements in A.

notation	meaning
Ω	sample space
ω	outcome
A	event (subset of Ω)
$\|A\|$	number of elements in A (if finite)
A^c	complement of A (not A)
$A \cup B$	union (A or B)
$A \cap B$ or AB	intersection (A and B)
$A - B$	set difference (points in A but not in B)
$A \subset B$	set inclusion (A is a subset of or equal to B)
$\varnothing$	null event (always false)
Ω	true event (always true)

We say that $A_1, A_2, \ldots$ are **disjoint** or **mutually exclusive** if $A_i \cap A_j = \varnothing$ whenever $i \neq j$. A **partition** of Ω is a sequence of disjoint sets $A_1, A_2, \ldots$ such that $\cup_{i=1}^{\infty} A_i = \Omega$. Given an event A, define the **indicator function of** A by

$$I_A(\omega) = I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

A sequence of sets $A_1, A_2, \ldots$ is **monotone increasing** if $A_1 \subset A_2 \subset \ldots$, and we define $\lim_{n\to\infty} A_n = \cup_{i=1}^{\infty} A_i$. A sequence of sets $A_1, A_2, \ldots$ is **monotone decreasing** if $A_1 \supset A_2 \supset \ldots$ and then we define $\lim_{n\to\infty} A_n = \cap_{i=1}^{n} A_i$. In either case, we will write $A_n \to A$.

1.2 2.3 Probability

A function $\mathbb{P}$ that assign a real number $\mathbb{P}(A)$ to each event A is a **probability distribution** or a **probability measure** if it satisfies the following three axioms:

- **Axiom 1**: $\mathbb{P}(A) \geq 0$ for every A
- **Axiom 2**: $\mathbb{P}(\Omega) = 1$
- **Axiom 3**: If $A_1, A_2, \ldots$ are disjoint then

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

A few properties that can be derived from the axioms:

- $\mathbb{P}(\varnothing) = 0$
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}\left(A^c\right) = 1 - \mathbb{P}(A)$
- $A \cap B = \varnothing \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Lemma 2.6. For any events A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + (B) - \mathbb{P}(AB)$.

Proof.

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}\left((AB^c) \cup (AB) \cup (A^c B)\right) && (1)\\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) && (2)\\
&= \mathbb{P}(AB^c) + \mathbb{P}(AB) + \mathbb{P}(A^c B) + \mathbb{P}(AB) - \mathbb{P}(AB) && (3)\\
&= \mathbb{P}((AB^c) \cup (AB)) + \mathbb{P}((A^c B) \cup (AB)) - \mathbb{P}(AB) && (4)\\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB) && (5)
\end{aligned}$$

Theorem 2.8 (Continuity of Probabilities). If $A_n \to A$ then $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.

Proof. Suppose that A_n is monotone increasing, $A_1 \subset A_2 \subset \ldots$. Let $B_1 = A_1$, and $B_{n+1} = A_{n+1} - A_n$ for $n > 1$. The B_i's are disjoint by construction, and $A_n = \cup_{i=1}^{n} A_i = \cup_{i=1}^{n} B_i$ for all n. From axiom 3,

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

and so

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

1.3 2.4 Probability on Finite Sample Spaces

If Ω is finite and each outcome is equally likely, then

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

which is called the **uniform probability distribution**.

We will need a few facts from counting theory later.

- Given n objects, the number of way or ordering these objects is $n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1$. We define $0! = 1$.
- We define

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

  read "n choose k", which is the number of different ways of choosing k objects from n.
- Note that choosing a subset k objects can be mapped to choosing the complement set of $n-k$ objects, so

  $$\binom{n}{k} = \binom{n}{n-k}$$

  and that there is only one way of choosing the empty set, so

  $$\binom{n}{0} = \binom{n}{n} = 1$$

1.4 2.5 Independent Events

- Two events A and B are **independent** if

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

  and we write $A \quad B$. A set of events $\{A_i : i \in I\}$ is independent if

  $$\mathbb{P}\left(\cap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

  for every finite subset J of I.
- Independence is sometimes assumed and sometimes derived.
- Disjoint events with positive probability are not independent.

1.5 2.6 Conditional Probability

- If $\mathbb{P}(B) > 0$ then the **conditional probability** of A given B is

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

- $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability, for fixed B. In general, $\mathbb{P}(A|\cdot)$ does **not** satisfies the axioms of probability for fixed A.
- In general, $\mathbb{P}(B|A) \neq \mathbb{P}(A|B)$.
- A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

1.6 2.7 Bayes' Theorem

Theorem 2.15 (The Law of Total Probability). Let $A_1, \ldots, A_k$ be a partition of Ω. Then, for any event B,

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

Proof. Let $C_j = BA_j$. Note that the C_j's are disjoint and that $B = \cup_{i=1}^{k} C_j$. Hence

$$\mathbb{P}(B) = \sum_j \mathbb{P}(C_j) = \sum_j \mathbb{P}(BA_j) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)$$

Theorem 2.16 (Bayes' Theorem). Let $A_1, \ldots, A_k$ be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for each i. If $\mathbb{P}(B) > 0$, then, for each $i = 1, \ldots, k$,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

We call $\mathbb{P}(A_i)$ the **prior probability** of A_i and $\mathbb{P}(A_i|B)$ the **posterior probability** of A_i.

Proof. We apply the definition of conditional probability twice, followed by the law of total probability:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_iB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

1.7 2.9 Technical Appendix

Generally, it is not feasible to assign probabilities to all the subsets of a sample space Ω. Instead, one restricts attention to a set of events called a **σ-algebra** or a **σ-field**, which is a class $\mathcal{A}$ that satisfies:

- $\varnothing \in \mathcal{A}$

- If $A_1, A_2, \cdots \in \mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$
- $A \in \mathcal{A}$ implies that $A^c \in \mathcal{A}$

The sets in $\mathcal{A}$ are said to be **measurable**. We call $(\Omega, \mathcal{A})$ a **measurable space**. If $\mathbb{P}$ is a probability measure defined in $\mathcal{A}$ then $(\Omega, \mathcal{A}, \mathbb{P})$ is a **probability space**. When Ω is the real line, we take $\mathcal{A}$ to be the smallest σ-field that contains all off the open sets, which is called the **Borel σ-field**.

1.8 2.10 Exercises

Exercise 2.10.1. Fill in the details in the proof of Theorem 2.8. Also, prove the monotone decreasing case.

If $A_n \to A$ then $\mathbb{P}(A_n) \to \mathbb{P}(A)$ as $n \to \infty$.

Solution.

Suppose that A_n is monotone increasing, $A_1 \subset A_2 \subset \ldots$. Let $B_1 = A_1$, and $B_{i+1} = A_{i+1} - A_i$ for $i > 1$.

The B_i's are disjoint by construction: assuming without loss of generality $i < j$, $\omega \in B_i \cap B_j$ implies that ω is in A_j, A_i, but not in A_{j-1}, A_{i-1}, where $A_0 = \varnothing$. In particular, this means that $\omega \in A_i$ but not $\omega \in A_{j-1}$. Since $A_i \subset A_{j-1}$, this implies that no such ω can satisfy those properties, and so B_i and B_j are disjoint.

Note that $A_n = \cup_{i=1}^n A_i = \cup_{i=1}^n B_i$ for all n:

$$\cup_{i=1}^n B_i = \cup_{i=1}^n (A_i - A_{i-1}) \subset \cup_{i=1}^n A_i = A_n$$

Also note that $A_n \subset \cup_{i=1}^n B_i$, since, if $f(\omega) = \min\{k : \omega \in A_k\}$, then $\omega \in B_{f(\omega)}$, so all elements of A_n are in some B_k.

The proof follows as given; from axiom 3,

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

and so

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

The monotone decreasing case can be obtained by looking at the complementary series $A_1^c, A_2^c, \ldots$,which is monotone increasing. We get

$$\lim_{n\to\infty} \mathbb{P}(A_n^c) = \mathbb{P}(A^c) \tag{6}$$

$$\lim_{n\to\infty} 1 - \mathbb{P}(A_n^c) = 1 - \mathbb{P}(A^c) \tag{7}$$

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A) \tag{8}$$

Exercise 2.10.2. Prove the statements in equation (2.1).

- $\mathbb{P}(\varnothing) = 0$
- $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- $0 \leq \mathbb{P}(A) \leq 1$
- $\mathbb{P}\left(A^c\right) = 1 - \mathbb{P}(A)$
- $A \cap B = \varnothing \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Solution.

- By partitioning the event space Ω into disjoint partitions $(\Omega, \varnothing)$ we get

$$\mathbb{P}(\Omega) + \mathbb{P}(\varnothing) = \mathbb{P}(\Omega) \Rightarrow \mathbb{P}(\varnothing) = 0$$

- Assuming $A \subset B$ and partitioning B as $(A, B - A)$, we get

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

- $\mathbb{P}(A) \geq 0$ from axiom 1. By partitioning Ω as (A, A^c), we get

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1 \Rightarrow \mathbb{P}(A) \leq 1$$

- By partitioning Ω as (A, A^c), we get

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

- Assuming A, B are disjoint, we partition $A \cup B$ in (A, B) and get:

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

Exercise 2.10.3. Let Ω be a sample space and let $A_1, A_2, \ldots$ be events. Define $B_n = \cup_{i=n}^{\infty} A_i$ and $C_n = \cap_{i=n}^{\infty} A_i$

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

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

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

Solution.

(a) By construction, $B_{n+1} = A_n \cup B_n$ and so $B_{n+1} \supset B_n$. Similarly, $C_{n+1} = A_n \cap C_n$ and so $C_{n+1} \subset C_n$.

(b)

- Assume ω belongs to an infinite number of the events, $\omega \in A_j$ for $j \in J(\omega)$. Then, for every n, there is a $m \geq n$ such that $m \in J(\omega)$, and so $\omega \in B_n$ for every n. This implies that $\omega \in \cap_{n=1}^{\infty} B_n$.

- Assume that $\omega \in \cap_{n=1}^{\infty} B_n$. Then, for every n, $\omega \in B_n$, so for every n there is a $m \geq n$ such that $\omega \in A_m$. This implies there is an infinite number of such events A_m.

(c)

Let's prove the contrapositive.

- Assume that ω does not belong to an infinite number of events A_i. Then, for every n, there is a $m \geq$ such that $\omega \in A_m^c$, and so ω is not in C_n. Since ω is not in none of the C_n's, it is not in the union of all C_n's either.

- Assume that ω is not in the union of all C_n. This implies that ω is is not in any event C_n. This implies that, for every n, there is a $m \geq n$ such that ω is not in A_m. This implies that there is an infinite number of such events A_m.

Exercise 2.10.4. Let $\{A_i : i \in I\}$ be a collection of events where I is an arbitrary index set. Show that

$$(\cup_{i\in I} A_i)^c = \cap_{i\in I} A_i^c \quad \text{and} \quad (\cap_{i\in I} A_i)^c = \cup_{i\in I} A_i^c$$

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

Solution.

We can prove the result directly by noting that every outcome ω belongs to or does not belong to both sides of each equality:

$$\begin{aligned}
&\omega \in (\cup_{i\in I} A_i)^c && (9)\\
\iff &\text{not } (\omega \in \cup_{i\in I} A_i) && (10)\\
\iff &\forall i \in I, \text{not } (\omega \in A_i) && (11)\\
\iff &\forall i \in I, \omega \in A_i^c && (12)\\
\iff &\omega \in \cap_{i\in I} A_i^c && (13)
\end{aligned}$$

and

$$\begin{aligned}
&\omega \in (\cap_{i\in I} A_i)^c && (14)\\
\iff &\text{not } (\omega \in \cap_{i\in I} A_i) && (15)\\
\iff &\text{not } (\forall i \in I, \omega \in A_i) && (16)\\
\iff &\exists i \in I, \text{not } \omega \in A_i && (17)\\
\iff &\exists i \in I, \omega \in A_i^c && (18)\\
\iff &\omega \in \cup_{i\in I} A_i^c && (19)
\end{aligned}$$

Exercise 2.10.5. 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?

Solution. The sample space is a set of coin toss results sequences containing two heads, and ending in heads:

$$S = \left\{ (r_1, \ldots, r_k) : r_i \in \{\text{head}, \text{tails}\}, \left| \{r_j = \text{head}\} \right| = 2, r_k = \text{head} \right\}$$

The probability of requiring exactly k tosses is 0 if $k < 2$, as there are no such sequences in the event space.

The probability of stopping after k tosses is the probability of obtaining exactly 1 head in the first $k-1$ tosses, in a procedure that would not stop after any number of tosses, followed by the probability of getting a head in the k-th toss. This value is

$$\left((k-1)\left(\frac{1}{2}\right)^{k-1}\right)\left(\frac{1}{2}\right) = \frac{k-1}{2^k}$$

Note that, besides this combinatorial argument, we can verify that these probabilities do indeed add up to 1:

$$\frac{1}{1-x} = \sum_{k=0}^{\infty} x^k \tag{20}$$

$$\frac{d}{dx}\frac{1}{1-x} = \sum_{k=0}^{\infty} \frac{d}{dx} x^k \tag{21}$$

$$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} k x^{k-1} \tag{22}$$

$$\frac{x}{(1-x)^2} = \sum_{k=0}^{\infty} k x^k \tag{23}$$

so, for $x = 1/2$, $\sum_{k=0}^{\infty} 2^{-k} k = 2$, and so

$$\sum_{k=0}^{\infty} \frac{k}{2^{k+1}} = 1$$

Exercise 2.10.6. Let $\Omega = \{1, 2, \dots\}$. 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.

Solution. Assume that such a distribution exists, and let $\mathbb{P}(\{1\}) = p$. Since the distribution is uniform, the probability associated with any set of size 1 is p, and the probability associated with any set of size n is np.

- If $p > 0$, then a finite set A of size $|A| = \lceil 2/p \rceil$ would have probability value $\mathbb{P}(A) = \lceil 2/p \rceil p \geq (2/p)p = 2$, which is greater than 1 – a contradiction.
- If $p = 0$, then any finite set A must have $\mathbb{P}(A) = 0$. But then $\mathbb{P}(\Omega) = \sum_i \mathbb{P}(\{i\}) = \sum_i 0 = 0$, instead of 1 – a contradiction.

Exercise 2.10.7. Let $A_1, A_2, \dots$ be events. Show that

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

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

Solution. Following the hint, let $B_n = A_n - \cup_{i=1}^{n-1} A_i$.

• Note that, for $i < j$, B_i and B_j are disjoint, since all elements of B_i must be elements of A_i, and all elements of A_i are explicitly excluded on the definition of B_j.

• Also note that $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} B_n$: $A_n = \cup_{i=1}^{n} B_i$ by construction, so $\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} \cup_{i=1}^{n} B_i = \cup_{n=1}^{\infty} B_n$, since $B_i \cup B_i = B_i$ and we can include each B_i only once in the expression.

Now, we have:

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

since $B_n \cup \left(\cup_{i=1}^{n-1} A_i\right) = A_n$ and so $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ for every n.

Exercise 2.10.8. Suppose that $\mathbb{P}(A_i) = 1$ for each i. Prove that

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

Solution. Using the result from exercise 4,

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

Using the result from exercise 7,

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i^c\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i^c) = \sum_{i=1}^{\infty} \left(1 - \mathbb{P}\left(A_i\right)\right) = \sum_{i=1}^{\infty} 0 = 0$$

so the equality holds, since a probability is non-negative. Therefore,

$$\mathbb{P}\left(\cap_{i=1}^{\infty} A_i\right) = 1 - \mathbb{P}\left(\cup_{i=1}^{\infty} A_i^c\right) = 1 - 0 = 1$$

Exercise 2.10.9. For fixed B such that $\mathbb{P}(B) > 0$, show that $\mathbb{P}(\cdot|B)$ satisfies the axioms of probability.

Solution.

- Axiom 1: $\mathbb{P}(\cdot|B) = \frac{\mathbb{P}(\cdot B)}{\mathbb{P}(B)} \geq 0$, since $\mathbb{P}(\cdot B) > 0$.
- Axiom 2: $\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$.
- Axiom 3: Assuming $A_1, A_2, \ldots$ are disjoint,

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i|B\right) = \frac{\mathbb{P}\left(B\left(\cup_{i=1}^{\infty} A_i\right)\right)}{\mathbb{P}(B)} = \frac{\mathbb{P}\left(\cup_{i=1}^{\infty}\left(A_i B\right)\right)}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

Exercise 2.10.10. 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 between 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 to 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 that 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.

Solution. Following the provided notation, the event space is

$$\Omega = \{(1,2), (1,3), (2,3), (3,2)\}$$

$\mathbb{P}[\omega_2]$ = probability of opening an empty door. The probability and the reward associated with switching for each outcome are:

ω	$\mathbb{P}$	R
$(1,2)$	$\frac{1}{3}\frac{1}{2}$	0
$(1,3)$	$\frac{1}{3}\frac{1}{2}$	0
$(2,3)$	$\frac{1}{3}1$	1
$(3,2)$	$\frac{1}{3}1$	1

Therefore,

$$\mathbb{P}[R|\omega_2 = 2] = \frac{\mathbb{P}(\{(3,2)\})}{\mathbb{P}(\{(3,2),(1,2)\})} = \frac{\frac{1}{3}1}{\frac{1}{3}1 + \frac{1}{3}\frac{1}{2}} = \frac{2}{3}$$

and, similarly, $\mathbb{P}[R|\omega_3 = 3]$, and so $\mathbb{P}[R] = \frac{2}{3}$.

Exercise 2.10.11. Suppose that A and B are independent events. Show that A^c and B^c are independent events.

Solution.

$$\begin{aligned}
\mathbb{P}(A^cB^c) &= \mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B) && (24)\\
&= 1 - (\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)) && (25)\\
&= 1 - \mathbb{P}(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) && (26)\\
&= 1 - (1 - \mathbb{P}(A^c)) - (1 - \mathbb{P}(B^c)) + (1 - \mathbb{P}(A^c))(1 - \mathbb{P}(B^c)) && (27)\\
&= \mathbb{P}(A^c)\mathbb{P}(B^c) && (28)
\end{aligned}$$

Exercise 2.10.12. There are three cards. The first card 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.

Solution. There are 6 potential card sides to be chosen, all with equal probability, of which only 3 are green – one belongs to the red / green card, and two belong to the green / green card. The probability that the other side is also green is the probability that the a side on the green / green card was chosen, which is 2 / 3.

Exercise 2.10.13. Suppose a fair coin is tossed repeatedly until both a head and a tail have appeared at least once.

(a) Describe the sample space Ω.

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

Solution.

(a). The sample space consists of the sequence of k identical coin toss results and a coin toss result with the opposite value,

$$\Omega = \{(r_1, \ldots, r_k, r_{k+1}) : r_i \in \{\text{head}, \text{tails}\}, r_1 = \cdots = r_k \neq r_{k+1}\}$$

(b) Exactly 3 tosses will be required if the first 3 results are (h, h, t) or (t, t, h).

If we map all infinite coin toss sequences to Ω by truncating it whenever the stop condition occurs, the probability of a (single-outcome) event in Ω is the same as the probability of all outcomes mapped into it. In particular, the probability of a sequence with its first 3 symbols being a specific sequence is 1/8, and so the probability of the desired outcome is 1/8 + 1/8 = 1/4.

Exercise 2.10.14. 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.

Solution.

If $\mathbb{P}(A) = 0$, then $\mathbb{P}(AB) = \mathbb{P}(A) - \mathbb{P}(A - B) = 0 - \mathbb{P}(A - B) \leq 0$, and since probabilities are non-negative we must have $\mathbb{P}(AB) = 0$. Therefore $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B) = 0$ for all events B, and A is independent of every other event.

If $\mathbb{P}(A) = 1$, then $\mathbb{P}(A^c) = 0$, and so A^c and B are independent for every other event B. Then, from the result in exercise 10, A is also independent from every other event B^c – which covers all potential events, since every event has a complement.

If A is independent of itself, $\mathbb{P}(AA) = \mathbb{P}(A)\mathbb{P}(A)$, so $\mathbb{P}(A) = \mathbb{P}(A)^2$ or $\mathbb{P}(A)(\mathbb{P}(A) - 1) = 0$. Therefore $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

Exercise 2.10.15. The probability that a child has blue eyes is 1/4. Assume independence between children. Consider a family with 5 children.

(a) If it is known that at least one child has blue eyes, what is the probability that at least 3 children have blue eyes?

(b) If it is known that the youngest child has blue eyes, what is the probability that at least 3 children have blue eyes?

Solution.

(a) Represent the sample space as

$$\Omega = \{(x_1, x_2, x_3, x_4, x_5) : x_i \in \{0, 1\}\}$$

where $x_i = 1$ if the $i-$th child (youngest to oldest) has blue eyes.

- "At least one child has blue eyes" is the event $A = \Omega - \{(0, 0, 0, 0, 0)\}$.

- "At least 3 children have blue eyes" is the event B with 3 children with blue eyes, 4 children with blue eyes, or 5 children with blue eyes.
- The intersection of these events is $BA = B$.

Let $p = 1/4$ be the probability at a given child will have blue eyes. The desired probability is then:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(BA)}{\mathbb{P}(A)} = \frac{\binom{5}{3}p^3(1-p)^2 + \binom{5}{4}p^4(1-p) + \binom{5}{5}p^5}{1-(1-p)^5} = \frac{106}{781} \approx 0.1357$$

(b)

- "The youngest child has blue eyes" is the event $C = \{\omega = (1, x_2, x_3, x_4, x_5) : \omega \in \Omega\}$.
- The intersection of events B and C is BC, the set of outcomes starting with 1 and having the other 4 dimensions having 2, 3, or 4 values 1; $BC = \{\omega = (1, x_2, x_3, x_4, x_5) : \omega \in \Omega, x_2 + x_3 + x_4 + x_5 \geq 2\}$.

The desired probability is then

$$\mathbb{P}(B|C) = \frac{\mathbb{P}(BC)}{\mathbb{P}(C)} = \frac{p\left(\binom{4}{2}p^2(1-p)^2 + \binom{4}{3}p^3(1-p) + \binom{4}{4}p^4\right)}{p} = \frac{67}{256} \approx 0.2617$$

Exercise 2.10.16. Show that

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

Solution.

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

Exercise 2.10.17. Suppose k events for a partition of the sample space Ω, i.e. they are disjoint and $\cup_{i=1}^k A_i = \Omega$. Assume that $\mathbb{P}(B > 0)$. Prove that if $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$ then $\mathbb{P}(A_i|B) > \mathbb{P}(A_i)$ for some $i = 2, \ldots, k$.

Solution.

We have

$$B = B\Omega = B\left(\cup_{i=1}^k A_i\right) = \cup_{i=1}^k A_i B$$

and so

$$\mathbb{P}\left(\cup_{i=1}^k A_i B\right) = \mathbb{P}(B) \iff \sum_{i=1}^k \frac{\mathbb{P}(A_i B)}{\mathbb{P}(B)} = 1 \iff \sum_{i=1}^k \mathbb{P}(A_i|B) = 1 \iff \sum_{i=1}^k \mathbb{P}(A_i|B) = \sum_{i=1}^k \mathbb{P}(A_i)$$

If we assume that $\mathbb{P}(A_i|B) \leq \mathbb{P}(A_i)$ for all i and $\mathbb{P}(A_1|B) < \mathbb{P}(A_1)$, then we must have $\sum_{i=1}^k \mathbb{P}(A_i|B) < \sum_{i=1}^k \mathbb{P}(A_i)$, a contradiction. Therefore the desired statement must hold.

Exercise 2.10.18. Suppose that 30% of computer users use a Macintosh, 50% use Windows and 20% use Linux. Suppose that 65% of the Mac users have succumbed to a computer virus, 82% of the Windows users get the virus and 50% 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?

Solution. The event space can be described as:

outcome	probability
Mac, no virus	30% * 35%
Mac, virus	30% * 65%
Windows, no virus	50% * 18%
Windows, virus	50% * 82%
Linux, no virus	20% * 50%
Linux, virus	20% * 50%

The desired conditional probability is

$$\mathbb{P}(\text{Windows}|\text{virus}) = \frac{\mathbb{P}(\text{Windows}, \text{virus})}{\mathbb{P}(\text{virus})} = \frac{0.50 \cdot 0.82}{0.30 \cdot 0.65 + 0.50 \cdot 0.82 + 0.20 \cdot 0.50} \approx 0.5816$$

Exercise 2.10.19. 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

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

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 $\mathbb{P}(C_i|H)$ for $i = 1, \ldots, 5$.

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

(c) Find $\mathbb{P}(C_i|B_4)$ where B_4 means "first head is obtained on toss 4".

Solution.

(a) We have:

$$\mathbb{P}(C_i|H) = \frac{\mathbb{P}(C_i H)}{\mathbb{P}(H)} = \frac{\mathbb{P}(C_i H)}{\sum_j \mathbb{P}(C_j H)} = \frac{\mathbb{P}(C_i)\mathbb{P}(H|C_i)}{\sum_j \mathbb{P}(C_j)\mathbb{P}(H|C_j)}$$

Assuming that the coin selection is uniformly random, $\mathbb{P}(C_i) = 1/5$ for $i = 1, \ldots, 5$, and the above simplifies to

$$\frac{\mathbb{P}(H|C_i)}{\sum_j \mathbb{P}(H|C_j)} = \frac{p_i}{\sum_j p_j}$$

Therefore,

$$\mathbb{P}(C_1|H) = 0 \quad \mathbb{P}(C_2|H) = 1/10 \quad \mathbb{P}(C_3|H) = 1/5 \quad \mathbb{P}(C_4|H) = 3/10 \quad \mathbb{P}(C_5|H) = 2/5$$

(b) We have:

$$\mathbb{P}(H_2|H_1) = \frac{\mathbb{P}(H_2H_1)}{\mathbb{P}(H_1)} = \frac{\sum_j \mathbb{P}(C_jH_1H_2)}{\sum_j \mathbb{P}(C_jH_1)} = \frac{\sum_j (1/5)p_j^2}{\sum_j (1/5)p_j} = \frac{3}{16}$$

(c) We have:

$$\mathbb{P}(C_i|B_4) = \frac{\mathbb{P}(C_iB_4)}{\mathbb{P}(B_4)} = \frac{\mathbb{P}(C_iB_4)}{\sum_j \mathbb{P}(C_jB_4)}$$

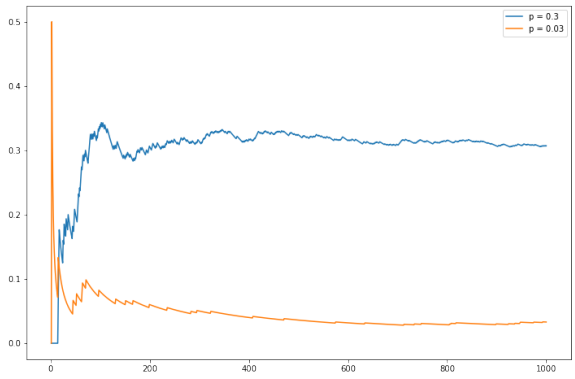
But $\mathbb{P}(C_iB_4) = (1/5)(1-p_i)^3 p_i$ – selecting coin i, then obtaining 3 tails followed by a head on that coin – so

$$\mathbb{P}(C_1|B_4) = 0 \quad \mathbb{P}(C_2|B_4) = \frac{27}{46} \quad \mathbb{P}(C_3|B_4) = \frac{8}{23} \quad \mathbb{P}(C_4|B_4) = \frac{3}{46} \quad \mathbb{P}(C_5|B_4) = 0$$

Exercise 2.10.20 (Computer Experiment). Suppose a coin has probability p of falling heads. 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 $n = 1000$ and simulate n coin flips. Plot the proportion of heads as a function of n. Repeat for $p = .03$.

<MINTED>

<MINTED>



Exercise 2.10.21 (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 averages of the X's to np. Try this for $p = .3$ and $n = 10, 100, 1000$.

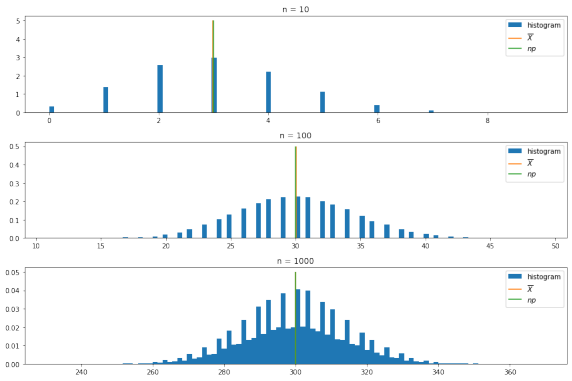
<MINTED>

<MINTED>

<MINTED>

X1 mean: 3.010 X1 np: 3.000
X2 mean: 30.013 X2 np: 30.000
X3 mean: 300.104 X3 np: 300.000

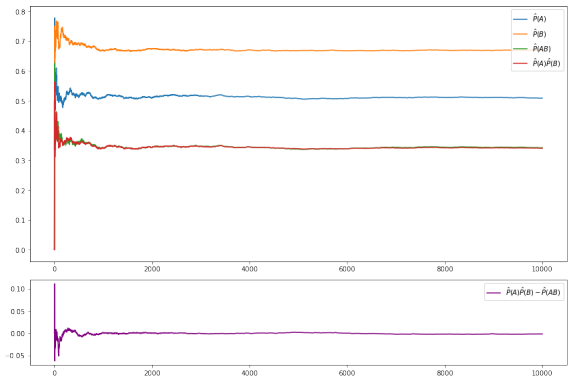
<MINTED>



Exercise 2.10.22 (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 $\hat{P}(AB) = \hat{P}(A)\hat{P}(B)$ where $\hat{P}$ is the proportion of times an event occurred in the simulation. Now find two events A and B that are not independent. Compute $\hat{P}(A)$, $\hat{P}(B)$ and $\hat{P}(AB)$. Compare the calculated values to their theoretical values. Report your results and interpret.

<MINTED>

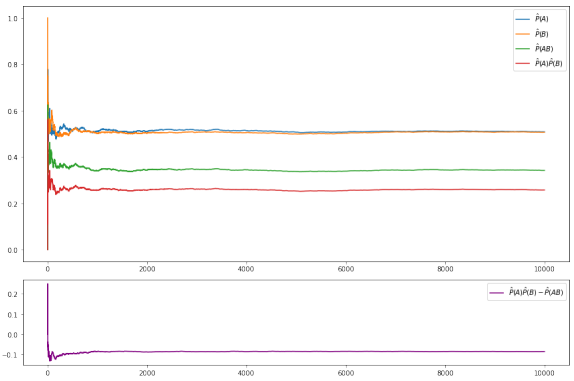
<MINTED>



For our own choice of non-independent events, let $A = \{2, 4, 6\}$ and $B = \{2, 4, 5\}$. Then $\mathbb{P}(A) = \mathbb{P}(B) = 1/2$ but $\mathbb{P}(AB) = 1/3$.

<MINTED>

<MINTED>



As noted, the estimates converges to the theoretical value – and the product of the estimates only converge to the estimate of the joint event in the scenario where the events are independent.

2 3. Random Variables

2.1 3.1 Introduction

A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

Technically, a random variable must be measurable. See the technical appendix for details.

Given a random variable X and a subset A of the real line, define $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$ and let

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega; X(\omega) \in A\}) \tag{29}$$
$$\mathbb{P}(X = x) = \mathbb{P}(X^{-1}(x)) = \mathbb{P}(\{\omega \in \Omega; X(\omega) = x\}) \tag{30}$$

X denotes a random variable and x denotes a possible value of X.

2.2 3.2 Distribution Functions and Probability Functions

The **cumulative distribution function**, CDF, $F_X : \mathbb{R} \to [0, 1]$ of a random variable X is defined by

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

The following result shows that the CDF completely determines the distribution of a random variable.

Theorem 3.7. Let X have CDF F and let Y have CDF G. If $F(x) = G(x)$ for all x then $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for all A.

Technically, we only have that $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ *for every measurable event* A.

Theorem 3.8. A function F mapping the real line to $[0,1]$ is a CDF for some probability measure $\mathbb{P}$ if and only if if satisfies the following three conditions:

- F is non-decreasing, i.e. $x_1 < x_2$ implies that $F(x_1) \leq F(x_2)$.
- F is normalized: $\lim_{n\to-\infty} F(x) = 0$ and $\lim_{n\to+\infty} F(x) = 1$.
- F is right-continuous, i.e. $F(x) = F(x^+)$ for all x, where

$$F(x^+) = \lim_{y\to x, y>x} F(y)$$

Proof. Suppose that F is a CDF. Let us show that F is right-continuous. Let x be a real number and let $y_1, y_2, \ldots$ be a sequence of real numbers such that $y_1 > y_2 > \ldots$ and $\lim_i y_i = x$. Let $A_i = (-\infty, y_i]$ and let $A = (-\infty, x]$. Note that $A = \cap_{i=1}^{\infty} A_i$ and also note that $A_1 \supset A_2 \supset \cdots$. Because the events are monotone, $\lim_i \mathbb{P}(A_i) = \mathbb{P}(\cap_i A_i)$. Thus,

$$F(x) = \mathbb{P}(A) = \mathbb{P}(\cap_i A_i) = \lim_i \mathbb{P}(A_i) = \lim_i F(y_i) = F(x^+)$$

Showing that F is non-decreasing and is normalized is similar. Proving the other direction, that is that F is non-decreasing, normalized, and right-continuous then it is a CDF for some random variable, uses some deep tools in analysis.

X is **discrete** if it takes countably many values

$$\{x_1, x_2, \ldots\}$$

We define the **probability density function** or **probability mass function** for X by

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

A set is countable if it is finite or if it can be put in a one-to-one correspondence with the integers.

Thus, $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and $\sum_i f_X(x_i) = 1$. The CDF of X is related to the PDF by

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

Sometimes we write f_X and F_X simply as f and F.

A random variable X is **continuous** if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and for every $a \leq b$,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

The function f_X is called the **probability density function**. We have that

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

and $f_X(x) = F'_X(x)$ at all points x at which F_X is differentiable.

Sometimes we shall write $\int f(x)dx$ or simply $\int f$ to mean $\int_{-\infty}^{\infty} f(x)dx$.

Warning: Note that if X is continuous then $\mathbb{P}(X = x) = 0$ for every x. We only have $f(x) = \mathbb{P}(X = x)$ for discrete random variables; we get probabilities from a PDF by integrating.

Lemma 3.15. Let F be the CDF for a random variable X. Then:

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

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

It is also useful to define the inverse CDF (or quantile function).

Let X be a random variable with CDF F. The **inverse CDF** or **quantile function** is defined by

$$F^{-1}(q) = \inf\{x : F(x) \geq q\}$$

for $q \in [0, 1]$. If F is strictly increasing and continuous then $F^{-1}(q)$ is the unique real number x such that $F(x) = q$.

We call $F^{-1}(1/4)$ the first quartile, $F^{-1}(1/2)$ the median (or second quartile), and $F^{-1}(3/4)$ the third quartile.

Two random variables X and Y are **equal in distribution** – written $X \stackrel{d}{=} Y$ – if $F_X(x) = F_Y(x)$ for all x. This does not mean that X and Y are equal. Rather, it means that probability statements about X and Y will be the same.

2.3 3.3 Some Important Discrete Random Variables

Warning about notation: It is traditional to write $X \sim F$ to indicate that X has distribution F. This is an unfortunate notation since the symbol $\sim$ is also used to denote an approximation. The notation is so pervasive we are stuck with it.

.. and we use $A \approx B$ to denote approximation. The LaTeX macros hint at this common usage: `\sim` *for $\sim$, and* `\approx` *for $\approx$.*

The Point Mass Distribution X has a point mass distribution at a, written $X \sim \delta_a$, if $\mathbb{P}(X = a) = 1$, in which case

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

The probability mass function is $f(x) = I(x = a)$, which takes value 1 if $x = a$ and 0 otherwise.

The Discrete Uniform Distribution Let $k > 1$ be a given integer. Suppose that X has probability mass function given by

$$f(x) = \begin{cases} 1/k & \text{for } x = 1, \ldots, k \\ 0 & \text{otherwise} \end{cases}$$

We say that X has a uniform distribution on $\{1, \ldots, k\}$.

The Bernoulli Distribution Let X represent a coin flip. Then $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ for some $p \in [0, 1]$. We say that X has a Bernoulli distribution, written $X \sim \text{Bernoulli}(p)$. The probability mass function is $f(x) = p^x(1 - p)^{1-x}$ for $x \in \{0, 1\}$.

The Binomial Distribution Suppose we have a coin which falls heads with probability p for some $0 \leq p \leq 1$. Flip the coin n times and let X be the number of heads. Assume that the tosses are independent. Let $f(x) = \mathbb{P}(X = x)$ be the mass function. It can be shown that

$$f(x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

A random variable with this mass function is called a Binomial random variable, and we write $X \sim \text{Binomial}(n, p)$. If $X_1 \sim \text{Binomial}(n, p_1)$ and $X_2 \sim \text{Binomial}(n, p_2)$ then $X_1 + X_2 \sim \text{Bimomial}(n, p_1 + p_2)$.

Note that X is a random variable, x denotes a particular value of the random variable, and n and p are **parameters**, that is, fixed real numbers.

The Geometric Distribution X has a geometric distribution with parameter $p \in (0, 1)$, written $X \sim \text{Geom}(p)$, if

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

We have that

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

Think of X as the number of flips needed until the first heads when flipping a coin.

The Poisson Distribution X has a Poisson distribution with parameter λ, written $X \sim \text{Poisson}(\lambda)$, if

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x \geq 0$$

Note that

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

The Poisson distribution is often used as a model for counts of rare events like radioactive decay and traffic accidents. If $X_1 \sim \text{Poisson}(n, \lambda_1)$ and $X_2 \sim \text{Poisson}(n, \lambda_2)$ then $X_1 + X_2 \sim \text{Poisson}(n, \lambda_1 + \lambda_2)$.

Warning: We defined random variables to be mappings from a sample space Ω to $\mathbb{R}$ but we did not mention sample space in any of the distributions above. Let's construct a sample space explicitly for a Bernoulli random variable. Let $\Omega = [0, 1]$, and define $\mathbb{P}$ to satisfy $\mathbb{P}([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$. Fix $p \in [0, 1]$ and define

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \leq p \\ 0 & \text{if } \omega > p \end{cases}$$

Then $\mathbb{P}(X = 1) = \mathbb{P}(\omega \leq p) = \mathbb{P}([0, p]) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Thus, $X \sim \text{Bernoulli}(p)$. We could do this for all of the distributions defined above. In practice, we think of a random variable like a random number but formally it is a mapping defined on some sample space.

2.4 3.4 Some Important Continuous Random Variables

The Uniform Distribution X has a uniform distribution with parameters a and b, written $X \sim \text{Uniform}(a, b)$, if

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

where $a < b$. The CDF is

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

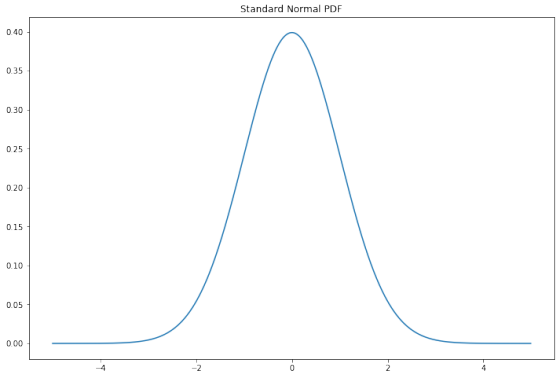
Normal (Gaussian) X has a Normal (or Gaussian) distribution with parameters μ and σ, denoted by $X \sim N(\mu, \sigma^2)$, if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. Later we shall see that μ is the "center" (or mean) of the distribution and that σ is the "spread" (or standard deviation) of the distribution. The Normal plays an important role in probability and statistics. Many phenomena have approximately Normal distributions. Later, we shall see that the distribution of a sum of random variables can be approximated by a Normal distribution (the central limit theorem).

We say that X has a **standard Normal distribution** if $\mu = 0$ and $\sigma = 1$. Tradition dictates that a standard Normal random variable is denoted by Z. The PDF and the CDF of a standard Normal are denoted by $\phi(z)$ and $\Phi(z)$. The PDF is plotted below.

<MINTED>



There is no closed-form expression for Φ. Here are some useful facts:

- If $X \sim N(\mu, \sigma^2)$ then $Z = (X - \mu)/\sigma \sim N(0, 1)$
- If $Z \sim N(0, 1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
- If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ are independent then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_u, \sum_{i=1}^{n} \sigma_i^2\right)$$

It follows from the first fact that if $X \sim N(\mu, \sigma^2)$ then

$$\mathbb{P}(a < X < b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Thus we can compute any probabilities we want as long as we can compute the CDF $\Phi(z)$ of the standard Normal. All statistical and computing packages will compute $\Phi(z)$ and $\Phi^{-1}(q)$.

Exponential Distribution X has an exponential distribution with parameter β, denoted by $X \sim \text{Exp}(\beta)$, if

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

where $\beta > 0$. The exponential distribution is used to model the lifetimes of electronic components and the waiting times between rare events.

Gamma Distribution For $\alpha > 0$, the **Gamma function** is defined by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$. X has a Gamma distribution with parameters α and β, denoted by $X \sim \text{Gamma}(\alpha, \beta)$, if

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

where $\alpha, \beta > 0$. The exponential distribution is just a Gamma$(1, \beta)$ distribution. If $X_i \sim \text{Gamma}(\alpha_i, \beta)$ are independent, then $\sum_{i=1}^n X_i \sim \text{Gamma}\left(\sum_{i=1}^n \alpha_i, \beta\right)$.

Beta Distribution X has Beta distribution with parameters $\alpha > 0$ and $\beta > 0$, denoted by $X \sim \text{Beta}(\alpha, \beta)$, if

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1$$

t and Cauchy Distribution X has a t distribution with ν degrees of freedom – written $X \sim t_\nu$ – if

$$f(x) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}$$

The t distribution is similar to a Normal but it has thicker tails. In fact, the Normal corresponds to a t with $\nu = \infty$. The Cauchy distribution is a special case of the t distribution corresponding to $\nu = 1$. The density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

The χ^2 Distribution X has a χ^2 distribution with p degrees of freedom – written $X \sim \chi^2_p$ – if

$$f(x) = \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{\frac{p}{2}}} x^{\frac{p}{2}-1}e^{-\frac{x}{2}}, \quad x > 0$$

If $Z_1, Z_2, \ldots$ are independent standard Normal random variables then $\sum_{i=1}^p Z_i^2 \sim \chi^2_p$.

2.5 3.5 Bivariate Distributions

Given a pair of discrete random variables X and Y, define the **joint mass function** by $f(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$. From now on, we write $\mathbb{P}(X = x \text{ and } Y = y)$ as $\mathbb{P}(X = x, Y = y)$. We write f as $f_{X,Y}$ when we want to be more explicit.

In the continuous case, we call a function $f(x, y)$ a PDF for the random variables (X, Y) if: - $f(x, y) \geq 0$ for all (x, y) - \$ $\int$

If (X, Y) have joint distribution with mass function $f_{X,Y}$, then the **marginal mass function** for X is defined by

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

and the **marginal mass function** for Y is defined by

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

For continuous random variables, the marginal densities are

$$f_X(x) = \int f(x, y)\, dy \quad \text{and} \quad f_Y(y) = \int f(x, y)\, dx$$

The corresponding marginal CDFs are denoted by F_X and F_Y.

2.6 3.7 Independent Random Variables

Two random variables X and Y are **independent** if, for every A and B,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

We write $X \quad Y$.

In principle, to check whether X and Y are independent we need to check the equation above for all subsets A and B. Fortunately, we have the following result which we state for continuous random variables though it is true for discrete random variables too.

Theorem 3.30. Let X and Y have joint PDF $f_{X,Y}$. Then $X \quad Y$ if and only if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all values x and y.

The statement is not rigorous because the density is defined only up to sets of measure 0.

The following result is helpful for verifying independence.

Theorem 3.33. 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.

2.7 3.8 Conditional Distributions

The **conditional probability mass function** is

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) > 0$.

For continuous random variables, the **conditional probability density function** is

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

assuming that $f_Y(y) > 0$. Then,

$$\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\, dx$$

We are treading in deep water here. When we compute $\mathbb{P}(X \in A|Y = y)$ in the continuous case we are conditioning on the event $\mathbb{P}(Y = y)$ which has probability 0. We avoid this problem by defining things in terms of the PDF. The fact that this leads to a well-defined theory is proved in more advanced courses. We simply take it as a definition.

2.8 3.9 Multivariate Distributions and IID Samples

Let $X = (X_1, \ldots, X_n)$ where the X_i's are random variables. We call X a **random vector**. Let $f(x_1, \ldots, x_n)$ denote the PDF. It is possible to define their marginals, conditionals, etc. much the same way as in the bivariate case. We say that $X_1, \ldots, X_n$ are independent if, for every $A_1, \ldots, A_n$,

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

It suffices to check that $f(x_1, \ldots, x_n) = \prod_{i=1}^n f_{X_i}(x_i)$. If $X_1, \ldots, X_n$ are independent and each has the same marginal distribution with density f, we say that $X_1, \ldots, X_n$ are IID (independent and identically distributed). We shall write this as $X_1, \ldots, X_n \sim f$ or, in terms of the CDF, $X_1, \ldots, X_n \sim F$. This means that $X_1, \ldots, X_n$ are independent draws from the same distribution. We also call $X_1, \ldots, X_n$ a **random sample** from F.

2.9 3.10 Two Important Multivariate Distributions

Multinomial Distribution The multivariate version of the Binomial is called a Multinomial. Consider drawing a ball from an urn which has balls of k different colors. Let $p = (p_1, \ldots, p_k)$ where $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$ and suppose that p_j is the probability of drawing a ball of color j. Draw n times (independent draws with replacement) and let $X = (X_1, \ldots, X_k)$, where X_j is the number of times that color j appears. Hence, $n = \sum_{j=1}^k X_j$. We say that X has a Multinomial distribution with parameters n and p, written $X \sim \text{Multinomial}(n, p)$. The probability function is

$$f(x) = \binom{n}{x_1 \cdots x_k} p_1^{x_1} \cdots p_k^{x_k}$$

where

$$\binom{n}{x_1 \cdots x_k} = \frac{n!}{x_1! \cdots x_k!}$$

Lemma 3.41. Suppose that $X \sim \text{Multinomial}(n, p)$ where $X = (X_1, \ldots, X_k)$ and $p = (p_1, \ldots, p_k)$. The marginal distribution of X_j is $\text{Binomial}(n, p_j)$.

Multivariate Normal The univariate Normal had two parameters, μ and σ. In the multivariate vector, μ is a vector and σ is replaced by a matrix Σ. To begin, let

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_k \end{pmatrix}$$

where $Z_1, \ldots, Z_k \sim N(0,1)$ are independent. The density of Z is

$$f(z) = \prod_{i=1}^{k} f(z_i) = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{k} z_j^2\right\} = \frac{1}{(2\pi)^{k/2}} \exp\left\{-\frac{1}{2} z^T z\right\}$$

We say that Z has a standard multivariate Normal distribution, written $Z \sim N(0, I)$ where it is understood that 0 represents a vector of k zeroes and I is the $k \times k$ identity matrix.

More generally, a vector X has multivariate Normal distribution, denoted by $X \sim N(\mu, \Sigma)$, if it has density

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

where $\det(\cdot)$ denotes the determinant of a matrix, μ is a vector of length k, and Σ is a $k \times k$ symmetric, positive definite matrix. Setting $\mu = 0$ and $\Sigma = I$ gives back the standard Normal.

A matrix Σ is positive definite if, for all non-zero vectors x, $x^T \Sigma x > 0$.

Since Σ is symmetric and positive definite, it can be shown that there exists a matrix $\Sigma^{1/2}$ – called the square root of Σ – with the following properties:

- $\Sigma^{1/2}$ is symmetric
- $\Sigma = \Sigma^{1/2}\Sigma^{1/2}$
- $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = I$, where $\Sigma^{1/2} = \left(\Sigma^{1/2}\right)^{-1}$

Theorem 3.42. If $Z \sim N(0, I)$ and $X = \mu + \Sigma^{1/2} Z$ then $X \sim N(\mu, \Sigma)$. Conversely, if $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0, I)$.

Suppose we partition a random Normal vector X as $X = (X_a, X_b)$, and we similarly partition

$$\mu = \begin{pmatrix} \mu_a & \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

Theorem 3.43. Let $X \sim N(\mu, \Sigma)$. Then:

1. The marginal distribution of X_a is $X_a \sim N(\mu_a, \Sigma_{aa})$.
2. The conditional distribution of X_b given $X_a = x_a$ is

$$X_b | X_a = x_a \sim N\left(\mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(x_a - \mu_a), \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)$$

3. If a is a vector then $a^T X \sim N(a^T \mu, a^T \Sigma a)$
4. $V = (X - \mu)^T \Sigma^{-1} (X - \mu) \sim \chi^2_k$

2.10 3.11 Transformations of Random Variables

Suppose that X is a random variable with PDF f_X and CDF F_X. Let $Y = r(X)$ be a function of X, such as $Y = X^2$ or $Y = e^X$. We call $Y = r(X)$ a transformation of X. How do we compute the PDF and the CDF of Y? In the discrete case, the answer is easy. The mass function of Y is given by

$$f_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}(r(X) = y) = \mathbb{P}(\{x; r(x) = y\}) = \mathbb{P}(X \in r^{-1}(y))$$

The continuous case is harder. There are 3 steps for finding f_Y:

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$.
2. Find the CDF

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(\{x; r(x) \leq y\}) = \int_{A_y} f_X(x)dx$$

3. The PDF is $f_Y(y) = F'_Y(y)$.

When r is strictly monotone increasing or strictly monotone decreasing then r has an inverse $s = r^{-1}$ and in this case one can show that

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

2.11 3.12 Transformation of Several Random Variables

In some cases we are interested in the transformation of several random variables. For example, if X and Y are given random variables, we might want to know the distribution of X/Y, $X + Y$, $\max\{X, Y\}$ or $\min\{X, Y\}$. Let $Z = r(X, Y)$. The steps for finding f_Z are the same as before:

1. For each z, find the set $A_z = \{(x, y) : r(x, y) \leq z\}$.
2. Find the CDF

$$F_Z(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(r(X, Y) \leq z) = \mathbb{P}(\{(x, y) : r(x, y) \leq z\}) = \int\int_{A_z} f_{X,Y}(x, y)\, dxdy$$

3. Find $f_Z(z) = F'_Z(z)$.

2.12 3.13 Technical Appendix

Recall that a probability measure $\mathbb{P}$ is defined on a σ-field $\mathcal{A}$ of a sample space Ω. A random variable X is **measurable** map $X : \Omega \to \mathbb{R}$. Measurable means that, for every x, \$\{ $\omega : X(\omega) \leq x\} \in \mathcal{A}$\$.

2.13 3.14 Exercises

Exercise 3.14.1. Show that

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

and

$$F(x_2) - F(x_1) = \mathbb{P}(X \leq x_2) - \mathbb{P}(X \leq x_1)$$

Solution.

By definition, $F(x_2) = \mathbb{P}(X \leq x_2)$ and $F(x_1) = \mathbb{P}(X \leq x_1)$, so the second equation is immediate.

Also by definition,

$$F(x^+) = \lim_{y \to x, y > x} F(y) = \lim_{y \to x, y > x} \mathbb{P}(X \leq y) = \mathbb{P}(\exists y : X \leq y, y > x) = \mathbb{P}(X \leq x)$$

and

$$F(x^-) = \lim_{y \to x, y < x} F(y) = \lim_{y \to x, y > x} \mathbb{P}(X \leq y) = \mathbb{P}(\exists y : X \leq y, y < x) = \mathbb{P}(X < x)$$

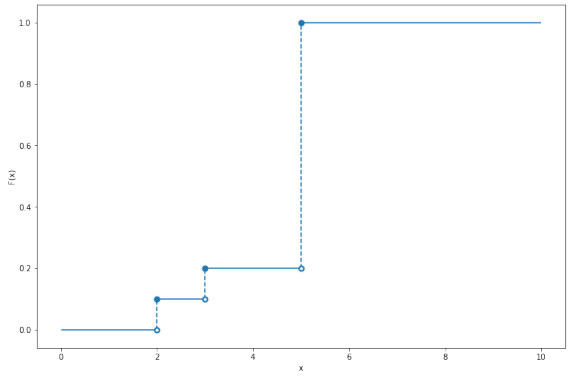
and so

$$F(x^+) - F(x^-) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = \mathbb{P}(X(\omega) \in \{x : X(\omega) \leq x\} - \{x : X(\omega) < x\}) = \mathbb{P}(X = x)$$

Exercise 3.14.2. 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 $\mathbb{P}(2 \leq X \leq 4.8)$.

Solution.

```
<MINTED>
```



$\mathbb{P}(2 < X \leq 4.8) = F(4.8) - F(2) = 0.2 - 0.1 = 0.1$ and $\mathbb{P}(2 \leq X \leq 4.8) = F(4.8) - F(2^-) = 0.2 - 0 = 0.2$

Exercise 3.14.3. Prove Lemma 3.15.

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

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

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

Solution.

Proof of first statement:

$$F(x) - F(x^-) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = \mathbb{P}(\{\omega : X(\omega) \leq x\} - \{\omega : X(\omega) < x\}) = \mathbb{P}(X = x)$$

Proof of second statement:

$$\mathbb{P}(x < X \leq y) = \mathbb{P}(\{\omega : X(\omega) \leq y\} - \{\omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq y) - \mathbb{P}(X \leq x) = F(y) - F(x)$$

Proof of third statement:

$$\mathbb{P}(X > x) = \mathbb{P}(\{\omega : X(\omega) > x\}) = \mathbb{P}(\{\omega : X(\omega) \leq x\}^c) = 1 - \mathbb{P}(X \leq x) = 1 - F(x)$$

Proof of fourth statement:

If F is continuous, then $\mathbb{P}(X = a) = \mathbb{P}(X = b) = 0$, and so

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(a < X < b) + \mathbb{P}(X = b) = \mathbb{P}(a < X < b) \tag{31}$$
$$\mathbb{P}(a \leq X < b) = \mathbb{P}(X = a) + \mathbb{P}(a < X < b) = \mathbb{P}(a < X < b) \tag{32}$$
$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X = a) + \mathbb{P}(a < X < b) + \mathbb{P}(X = b) = \mathbb{P}(a < X < b) \tag{33}$$

Exercise 3.14.4. Let X have probability density function

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

(a) Find the cumulative distribution function of X.

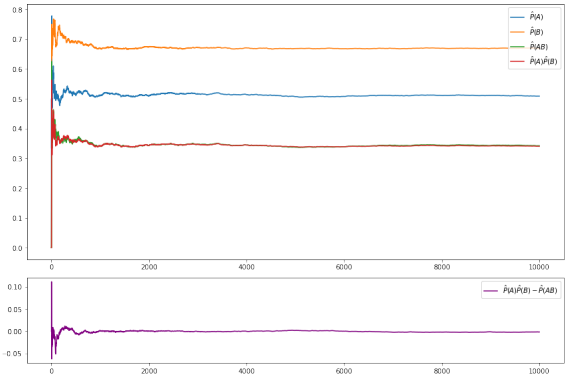
(b) Let $Y = 1/X$. Find the probability density function $f_Y(y)$ for Y. Hint: Consider three cases, $\frac{1}{5} \leq y \leq \frac{1}{3}$, $\frac{1}{3} \leq y \leq 1$, and $y \geq 1$.

Solution.

(a)

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

<MINTED>



(b)

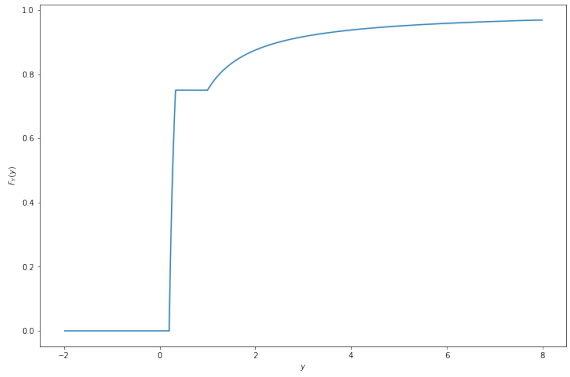
$\mathbb{P}(X > 0) = 1$, so $\mathbb{P}(Y > 0) = 1$. For $y > 0$, we have:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(\frac{1}{X} \leq y\right) = \mathbb{P}\left(X \geq \frac{1}{y}\right) = 1 - F_X\left(\frac{1}{y}\right)$$

so

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

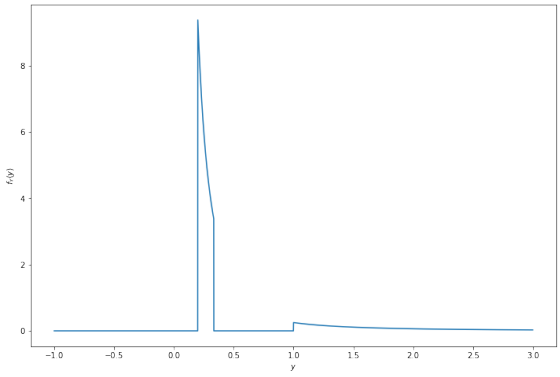
<MINTED>



The probability density function is $f_y(y) = F'_Y(y)$, so

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

<MINTED>



Exercise 3.14.5. 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.

Solution. If x or y take values that have probability mass 0, then trivially $f_{X,Y}(x,y) = 0$ and $f_X(x)f_Y(y) = 0$, so we only need to consider x and y with positive probability mass.

If X and Y are independent, then $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$ for all events A, B. In particular, this is true for $A = \{x\}$ and $B = \{y\}$ for all x, y, proving the implication in one direction.

In the other direction, we have

$$\mathbb{P}(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} f_{X,Y}(x,y)$$

and

$$\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x) \quad \text{and} \quad \mathbb{P}(Y \in B) = \sum_{y \in B} f_Y(y)$$

so $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y implies that for all A and B,

$$\mathbb{P}(X \in A)\mathbb{P}(Y \in B) = \left(\sum_{x \in A} f_X(x)\right)\left(\sum_{y \in B} f_Y(y)\right) = \sum_{x \in A}\sum_{y \in B} f_X(x)f_Y(y) = \sum_{x \in A}\sum_{y \in B} f_{X,Y}(x,y) = \mathbb{P}(X \in A, Y \in$$

so X, Y are independent, proving the implication in the other direction.

Exercise 3.14.6. 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:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

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

Solution. Note that I_A can only return values 0 or 1, so Y is a discrete variable with non-zero probability mass only at values 0 and 1.

But $\$P(Y = 1) = P(x \in A) = \int \_Af\_X(x)dx\$, sothePDFforY$ is:

$$f_Y(y) = \begin{cases} \int_A f_X(x)dx & \text{if } y = 1 \\ 1 - \int_A f_X(x)dx & \text{if } y = 0 \\ 0 & \text{otherwise} \end{cases}$$

and so the CDF of Y is:

$$F_Y(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \int_A f_X(x)dx & \text{if } 0 \leq y < 1 \\ 1 & \text{if } y \geq 1 \end{cases}$$

Exercise 3.14.7. 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 $\mathbb{P}(Z > z)$.

Solution.

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

But F_X and F_Y are both the CDF of the Uniform$(0, 1)$ distribution, so

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

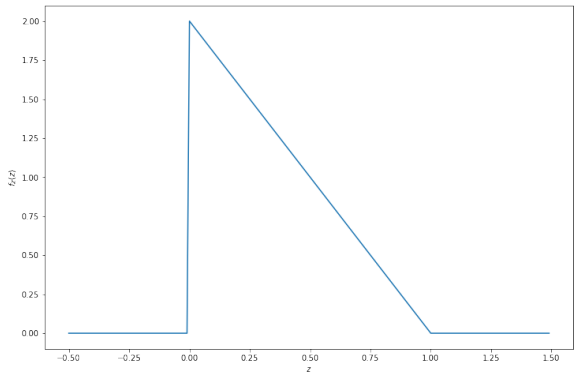
and so

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

and the PDF is $f_Z(z) = F'_Z(z)$:

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

<MINTED>



Exercise 3.14.8. Let X have CDF F. Find the CDF of $X^+ = \max\{0, X\}$.

Solution.

$$F_{X^+}(x) = \mathbb{P}(X^+ \leq x) = \mathbb{P}(0 < x, X < x) = I(0 < x)F_X(x)$$

Exercise 3.14.9. Let $X \sim \text{Exp}(\beta)$. Find $F(x)$ and $F^{-1}(q)$.

Solution. Let's start from the definition for the PDF of the exponential distribution:

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

We have:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \int_{-\infty}^{x} \frac{1}{\beta}e^{-t/\beta}dt = \frac{1}{\beta}\beta\left(1 - e^{-x/\beta}\right) = 1 - e^{-x/\beta}$$

and so

$$q = 1 - e^{-F^{-1}(q)/\beta} \implies F^{-1}(q) = -\beta\log(1-q)$$

Exercise 3.14.10. Let X and Y be independent. Show that $g(X)$ is independent of $h(Y)$ where g and h are functions.

Solution. Let $g^{-1}(A) = \{x : g(x) \in A\}$ and $h^{-1}(B) = \{y : h(y) \in B\}$. We have:

$$\mathbb{P}(g(X) \in A, h(Y) \in B) = \mathbb{P}(X \in g^{-1}(A), Y \in h^{-1}(B) = \mathbb{P}(X \in g^{-1}(A))\mathbb{P}(Y \in h^{-1}(B)) = \mathbb{P}(g(X) \in A)\mathbb{P}(h(Y)$$

Exercise 34.14.11. Suppose we toss a coin once and let p be the probability of heads. Let X denote the number of heads and Y denote the number of tails.

(a) Prove that X and Y are dependent.

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

Solution.

(a) The joint probability mass function is:

$f_{X,Y}$	$Y=0$	$Y=1$
$X=0$	0	$1-p$
$X=1$	p	0

In particular, $f_{X,Y}(1,0) = p$ and $f_X(1)f_Y(0) = p^2$, so the events are dependent unless $p \in \{0,1\}$.

(b) The joint probability mass function is:

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

where \$g(x) = e^{-\lambda}\frac{\lambda^x}{x!}p^x\$, \$h(y) = \frac{\lambda^y}{y!}(1-p)^y\$. The result then follows from theorem 3.33.

Exercise 3.14.12. Prove theorem 3.33.

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.

Solution.

Given that F is the joint probability density function of X and Y, then the marginal distributions for X and Y are

$$f_X(x) = \int f(x,y)dy = \int g(x)h(y)dy = g(x)\int h(y)dy = Hg(x) \quad \text{for some constant } H$$

and

$$f_Y(y) = \int f(x,y)dx = \int g(x)h(y)dx = h(y)\int g(x)dx = Gh(y) \quad \text{for some constant } G$$

Therefore,

$$f_X(x)f_Y(y) = HGg(x)h(y) = HGf(x,y)$$

Integrating over x, and fixing $y = y_0$ with non-zero probability density,

$$\int f_X(x)f_Y(y_0)dx = HG\int f(x,y_0)dx \implies f_Y(y_0) = HGf_Y(y_0) \implies HG = 1$$

and so

$$f_X(x)f_Y(y) = f(x,y)$$

therefore X and Y are independent.

Exercise 3.14.13. 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).

Solution.

(a)

Assuming $y > 0$,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq \log y) = \Phi(\log y)$$

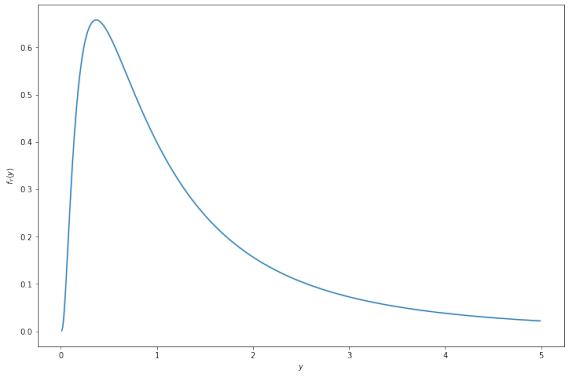
and so

$$f_Y(y) = \frac{d}{dy}\Phi(\log y) = \frac{d\Phi(\log y)}{d\log y}\frac{d\log y}{dy} = \frac{\phi(\log(y))}{y}$$

And, of course, Y can never take a negative value, so

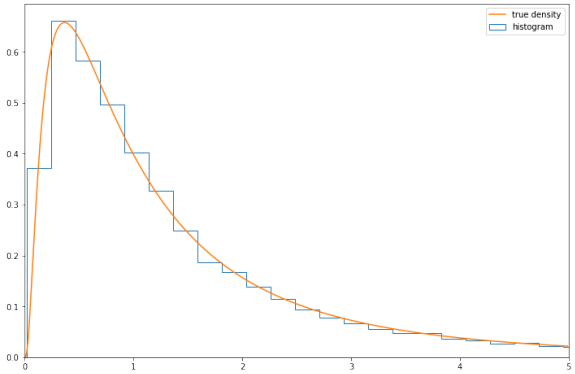
$$f_Y(y) = \begin{cases} \frac{\phi(\log(y))}{y} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

<MINTED>



(b)

<MINTED>



Exercise 3.14.14. Let (X, Y) be uniformly distributed on the unit disc $\{(x, y) : x^2 + y^2 \leq 1\}$. Let $R = \sqrt{X^2 + Y^2}$. Find the CDF and the PDF of R.

Solution.

Assuming $r > 0$,

$$F_R(r) = \mathbb{P}(R \leq r) = \mathbb{P}(X^2 + Y^2 \leq r^2) = \int\int_{x^2+y^2 \leq r^2} f(x, y)dxdy$$

is proportional to the area of a disc of radius r. Since $F_R(1) = 1$, we have that the CDF is

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

and the PDF is $f_R(r) = F_R'(r)$:

$$f_R(r) = \begin{cases} 0 & \text{if } r \leq 0 \\ 2r & \text{if } 0 < r \leq 1 \\ 0 & \text{if } r > 1 \end{cases}$$

Exercise 3.14.15 (A universal random number generator). Let X have a continuous, strictly increasing CDF. Let $Y = F(X)$. Find the density of Y. This is called the probability integral transformation. 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 a $\text{Exp}(\beta)$ distribution.

Solution.

For $0 \leq y \leq 1$, the CDF of Y is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F(X) \leq y) = \mathbb{P}(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y$$

and so $Y \sim \text{Uniform}(0, 1)$.

For $0 \leq q \leq 1$,

$$F_X(q) = \mathbb{P}(X \leq q) = \mathbb{P}(F(X) \leq F(q)) = \mathbb{P}(U \leq F(q)) = F(q)$$

and so $X \sim F$.

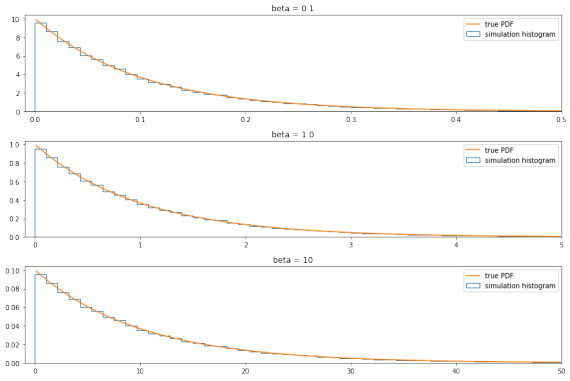
To create a generator for $\text{Exp}(\beta)$, let F be the CDF for this distribution,

$$F(x) = 1 - e^{-x/\beta} \quad F^{-1}(q) = -\beta \log(1 - q)$$

<MINTED>

<MINTED>

<MINTED>



Exercise 3.14.16. 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 $\text{Binomial}(n, \pi)$ 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 $\{X = x, X + Y = n\} = \{X = x, Y = n - x\}$

Solution.

We have:

$$\mathbb{P}(X = x | X + Y = n) = \frac{\mathbb{P}(X = x, X + Y = n)}{\mathbb{P}(X + Y = n)} \tag{34}$$

$$= \frac{\mathbb{P}(X = x, Y = n - x)}{\mathbb{P}(X + Y = n)} \tag{35}$$

$$= \frac{\mathbb{P}(X = x)\mathbb{P}(Y = n - x)}{\mathbb{P}(X + Y = n)} \tag{36}$$

$$= \frac{\frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{n-x} e^{-\mu}}{(n-x)!}}{\frac{(\lambda+\mu)^n e^{-(\lambda+\mu)}}{n!}} \tag{37}$$

$$= \frac{n!}{x!(n-x)!} \frac{\lambda^x \mu^{n-x}}{(\lambda + \mu)^n} \tag{38}$$

$$= \binom{n}{x} \left(\frac{\lambda}{\lambda + \mu}\right)^x \left(\frac{\mu}{\lambda + \mu}\right)^{n-x} \tag{39}$$

$$= \binom{n}{x} \pi^x (1 - \pi)^{n-x} \tag{40}$$

and so the result follows.

Exercise 3.14.17. Let

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

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

Solution.

The conditional density $f_{X|Y}(x|y)$ is:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int f_{X,Y}(t,y)dt} = \frac{c(x+y^2)}{\int_0^1 c(t+y^2)dt} = \frac{x+y^2}{\int_0^1 t+y^2 dt} = \frac{x+y^2}{\frac{1}{2}+y^2}$$

In particular, when $y = 1/2$,

$$f_{X|Y}(x|1/2) = \frac{4x+1}{3}$$

and so

$$\mathbb{P}\left(X < x \,\middle|\, Y = \frac{1}{2}\right) = \int_0^x \frac{4t+1}{3} dt = \frac{2x^2+x}{3}$$

so

$$\mathbb{P}\left(X < \frac{1}{2} \,\middle|\, Y = \frac{1}{2}\right) = \frac{1}{3}$$

Exercie 3.14.18. 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 $\mathbb{P}(|X| > |x|) = .05$.

Solution.

Rather than using tables, we will express these in terms of a standard Normal, and use "computer packages" to compute the expression based both on the original Normal and the standard Normal.

(a) $\mathbb{P}(X < 7) = \mathbb{P}\left(\frac{X-3}{4} < \frac{7-3}{4}\right) = \mathbb{P}\left(Z < 1\right) = \Phi\left(1\right)$

```
<MINTED>
```

```
<MINTED>                                    0.8413 0.8413
```

(b) $P(X > -2) = P\left(\frac{X-3}{4} > \frac{-2-3}{4}\right) = \mathbb{P}\left(Z < -\frac{5}{4}\right) = 1 - \Phi\left(-\frac{5}{4}\right)$ $

```
<MINTED>
```

<MINTED> 0.8944 0.8944

(c) $\mathbb{P}(X > x) = .05 \iff 1 - F_X(x) = .05 \iff F_X(x) = .95 \iff x = F_X^{-1}(.95)$

or: $\mathbb{P}(X > x) = .05 \iff \mathbb{P}\left(\frac{X-3}{4} > \frac{x-3}{4}\right) = .05 \iff 1 - \Phi\left(\frac{x-3}{4}\right) = .05 \iff \frac{x-3}{4} = \Phi^{-1}(.95) \iff x = 4\Phi^{-1}(.95) + 3$

<MINTED>

<MINTED> 9.5794 9.5794

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

or $\mathbb{P}(0 \leq X < 4) = \mathbb{P}\left(\frac{0-3}{4} < Z < \frac{4-3}{4}\right) = \Phi\left(\frac{1}{4}\right) - \Phi\left(\frac{-3}{4}\right)$

<MINTED>

<MINTED> 0.3721 0.3721

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

For a constant $c \geq 0$, we have

$$\mathbb{P}(|X| > c) = 1 - \mathbb{P}(|X| < c) = 1 - \mathbb{P}(-c < X < -c) \tag{41}$$

$$= 1 - \mathbb{P}\left(-\frac{c-3}{4} < Z < \frac{c-3}{4}\right) \tag{42}$$

$$= 1 - \Phi\left(\frac{c-3}{4}\right) + \Phi\left(-\frac{c-3}{4}\right) \tag{43}$$

$$= 2 - 2\Phi\left(\frac{c-3}{4}\right) = 2 - 2F_X(c) \tag{44}$$

So we can solve the original equation:

$$2 - 2F_X(c) = .05 \iff c = F_X^{-1}(0.975)$$

or

$$2 - 2\Phi\left(\frac{c-3}{4}\right) = .05 \iff c = 4\Phi^{-1}(0.975) + 3$$

<MINTED>

<MINTED> 10.8399 10.8399

Exercise 3.14.19. Prove formula (3.11).

Given $Y = r(X)$, when r is strictly monotone increasing or strictly monotone decreasing then r has an inverse $s = r^{-1}$ and in this case one can show that

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

Solution.

Assume r is strictly monotone increasing. Then \$ d s(y)$\overline{dy>0\$,andtheCDFofY}$ is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \leq s(y)) = F_X(s(y))$$

Derivating over y,

$$\frac{d}{dy}F_y(y) = \frac{dF_X(s(y))}{ds(y)}\frac{ds(y)}{dy} \tag{45}$$

$$f_Y(y) = f_X(s(y))\frac{ds(y)}{dy} \tag{46}$$

For the case when r is strictly monotone decreasing, \$ d s(y)$\overline{dy<0\$,andtheCDFofY}$ is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(r(X) \leq y) = \mathbb{P}(X \geq s(y)) = 1 - F_X(s(y))$$

Derivating over y,

$$\frac{d}{dy}F_y(y) = \frac{d - F_X(s(y))}{ds(y)}\frac{ds(y)}{dy} \tag{47}$$

$$f_Y(y) = -f_X(s(y))\frac{ds(y)}{dy} \tag{48}$$

and the result follows.

Exercise 3.14.20. Let $X, Y \sim$ Uniform$(0, 1)$ be independent. Find the PDF for $X - Y$ and X/Y.

Solution.

The joint density of X, Y is

$$f(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $A = X - Y$. The CDF of A is calculated over the area $A_a = \{(x, y) : x - y \leq a\}$:

$$F_A(a) = \mathbb{P}(X + Y \leq a) = \int_{A_a} f(x, y)\, dxdy$$

- If $0 \leq a \leq 1$, the area is consists of all points in the unit square, othe than the triangle at points $(a, 0), (1, 0), (1, 1 - a)$. Then,

$$F_A(a) = 1 - \frac{(1 - a)^2}{2}$$

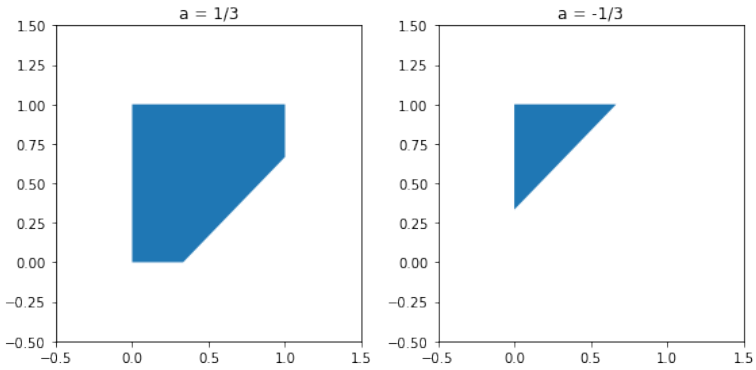
- If $-1 \leq a \leq 0$, the area consists of only the points in the triangle $(0,1),(1+a,1),(0,-a)$. Then,

$$F_A(a) = \frac{(1+a)^2}{2}$$

Therefore, the PDF is $f_A(a) = F'_A(a)$, or

$$f_A(a) = \begin{cases} 1+a & \text{if } -1 < a \leq 0 \\ 1-a & \text{if } 0 < a \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

<MINTED>



Let $B = X/Y$. The CDF of B is calculated over the area $B_b = \left\{(x,y) : \frac{x}{y} \leq b\right\}$.

$$F_B(a) = \mathbb{P}\left(\frac{X}{Y} \leq b\right) = \int_{B_b} f(x,y)\, dxdy$$

- If $0 < b \leq 1$, the relevant area is a triangle with points $(0,0),(1,0),(1,b)$. Then,

$$F_B(b) = \frac{b}{2}$$

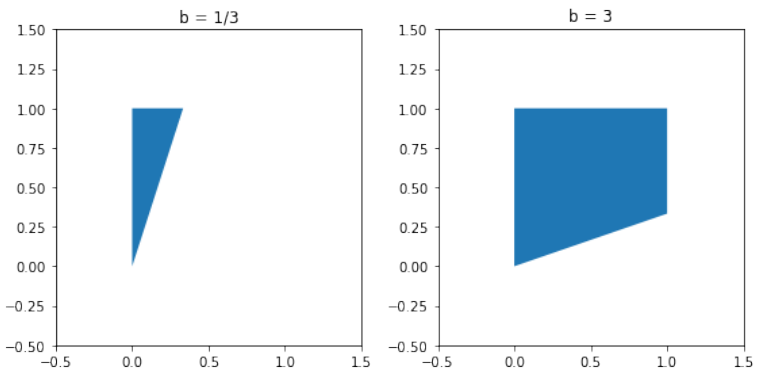
- If $b > 1$, the relevant area is the unit square minus the triangle with points $(0,0),(0,1),(1/b,1)$. Then,

$$F_B(b) = 1 - \frac{1}{2b}$$

Therefore, the PDF is $f_B(b) = F'_B(b)$, or

$$f_A(a) = \begin{cases} \frac{1}{2} & \text{if } 0 < b \leq 1 \\ \frac{1}{2b^2} & \text{if } b > 1 \\ 0 & \text{otherwise} \end{cases}$$

<MINTED>



Exercise 3.14.21. 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 $i = 1, \ldots, n$.

Solution. We have:

$$\mathbb{P}(Y \leq y) = \mathbb{P}(\forall i, X_i \leq y) = \prod_i \mathbb{P}(X_i \leq y) = \prod_i \left(1 - e^{-y/\beta}\right) = (1 - e^{-y/\beta})^n$$

and so the PDF is $f_Y(y) = F'_Y(y)$:

$$f_Y(y) = \frac{d(1 - e^{-y/\beta})^n}{dy} = \frac{n}{\beta} e^{-y/\beta}(1 - e^{-y/\beta})^{n-1}$$

Exercise 3.14.22. Let X and Y be random variables. Suppose that $\mathbb{E}(Y|X) = X$. Show that $\text{Cov}(X, Y) = \mathbb{V}(X)$.

Solution.

The covariance is:

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

and the variance is

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

We aim to prove that both of those expressions have the same value when $\mathbb{E}(Y|X) = X$.

Taking the expectaion on $\mathbb{E}(Y|X)$ we get:

$$\mathbb{E}(\mathbb{E}(Y|X)) = \int \mathbb{E}(Y = y|X) f_Y(y)\, dy = \int y f_Y(y) = \mathbb{E}(Y)$$

so taking the expectation on both sides of $\mathbb{E}(Y|X) = X$ gives us $\mathbb{E}(Y) = \mathbb{E}(X)$.

Now, we have

$$\mathbb{E}(XY) = \int \mathbb{E}(XY|X=x)f_X(x)dx \tag{49}$$
$$= \int \mathbb{E}(xY|X=x)f_X(x)dx \tag{50}$$
$$= \int x\mathbb{E}(Y|X=x)f_X(x)dx \tag{51}$$
$$= \int x^2 f_X(x)dx \tag{52}$$
$$= \mathbb{E}(X^2) \tag{53}$$

Using $\mathbb{E}(X^2) = \mathbb{E}(XY)$ and $\mathbb{E}(Y) = \mathbb{E}(X)$, we get the desired result.

Exercise 3.14.23. Let $X \sim \text{Uniform}(0,1)$. Let $0 < a < b < 1$. Let

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

and let

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

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

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

Solution.

(a) The joint probability mass function for Y and Z is:

	$Z=0$	$Z=1$
$Y=0$	0	$1-b$
$Y=1$	a	$b-a$

These are not independent; in particular,

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

since the equality would only hold if $b=0$ or $a=0$, and the inequality precludes both of these.

(b) We have:

$$\mathbb{E}(Y|Z=0) = \frac{\sum_y yf(y,0)}{f_Z(0)} = \frac{f(1,0)}{a} = 1$$

and

$$\mathbb{E}(Y|Z=1) = \frac{\sum_y yf(y,1)}{f_Z(1)} = \frac{f(1,1)}{1-a} = \frac{b-a}{1-a}$$

3 4. Expectation

3.1 4.1 Expectation of a Random Variable

The **expected value**, **mean** or **first moment** of X is defined to be

$$\mathbb{E}(X) = \int x \, dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int x f(x) \, dx & \text{if } X \text{ is continuous} \end{cases}$$

assuming that the sum (or integral) is well-defined. We use the following notation to denote the expected value of X:

$$\mathbb{E}(X) = \mathbb{E}X = \int x \, dF(x) = \mu = \mu_X$$

The expectation is a one-number summary of the distribution. Think of $\mathbb{E}(X)$ as the average value you'd obtain if you computed the numeric average $n^{-1}\sum_{i=1}^n X_i$ for a large number of IID draws $X_1, \ldots, X_n$. The fact that $\mathbb{E}(X) \approx n^{-1}\sum_{i=1}^n X_i$ is a theorem called the law of large numbers which we will discuss later. We use $\int x \, dF(x)$ as a convenient unifying notation between the discrete case $\sum_x x f(x)$ and the continuous case $\int x f(x) \, dx$ but you should be aware that $\int x \, dF(x)$ has a precise meaning discussed in real analysis courses.

To ensure that $\mathbb{E}(X)$ is well defined, we say that $\mathbb{E}(X)$ exists if $\int_x |x| \, dF_X(x) < \infty$. Otherwise we say that the expectation does not exist. From now on, wheneverwe discuss expectations, we implicitly assume they exist.

Theorem 4.6 (The rule of the lazy statician). Let $Y = r(X)$. Then

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) \, dF_X(x)$$

As a special case, let A be an event and let $r(x) = I_A(x)$, where $I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ otherwise. Then

$$\mathbb{E}(I_A(X)) = \int I_A(x) f_X(x) dx = \int_A f_X(x) dx = \mathbb{P}(X \in A)$$

In other words, probability is a special case of expectation.

Functions of several variables are handled in a similar way. If $Z = r(X, Y)$ then

$$\mathbb{E}(Z) = \mathbb{E}(r(X, Y)) = \int \int r(x, y) \, dF(x, y)$$

The k**-th moment** of X is defined to be $\mathbb{E}(X^k)$, assuming that $\mathbb{E}(|X|^k) < \infty$. We shall rarely make much use of moments beyond $k = 2$.

3.2 4.2 Properties of Expectations

Theorem 4.10. If $X_1, \ldots, X_n$ are random variables and $a_1, \ldots, a_n$ are constants, then

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i)$$

Theorem 4.12. Let $X_1, \ldots, X_n$ be independent random variables. Then,

$$\mathbb{E}\left(\prod_i X_i\right) = \prod_i \mathbb{E}(X_i)$$

Notice that the summation rule does not require independence but the product does.

3.3 4.3 Variance and Covariance

Let X be a random variable with mean μ. The **variance** of X – denoted by σ^2 or σ_X^2 or $\mathbb{V}(X)$ or $\mathbb{V}X$ – is defined by

$$\sigma^2 = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 \, dF(x)$$

assuming this expectation exists. The **standard deviation** is $\text{sd}(X) = \sqrt{\mathbb{V}(X)}$ and is also denoted by σ and σ_X.

Theorem 4.14. Assuming the variance is well defined, it has the following properties:

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$
2. If a and b are constants then $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$
3. If $X_1, \ldots, X_n$ are independent and $a_1, \ldots, a_n$ are constants then

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i)$$

If $X_1, \ldots, X_n$ are random variables then we define the **sample mean** to be

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

and the **sample variance** to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$$

Theorem 4.16. Let $X_1, \ldots, X_n$ be IID and let $\mu = \mathbb{E}(X_i)$, $\sigma^2 = \mathbb{V}(X_i)$. Then

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

If X and Y are random variables, then the covariance and correlation between X and Y measure how strong the linear relationship between X and Y is.

Let X and Y be random variables with means μ_X and μ_Y and standard deviation σ_X and σ_Y. Define the **covariance** between X and Y by

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

and the **correlation** by

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Theorem 4.18. The covariance satisfies:

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

The correlation satisfies:

$$-1 \leq \rho(X, Y) \leq 1$$

If $Y = a + bX$ for some constants a and b then $\rho(X, Y) = 1$ if $b > 0$ and $\rho(X, Y) = -1$ if $b < 0$. If X and Y are independent, then $\text{Cov}(X, Y) = \rho = 0$. The converse is not true in general.

Theorem 4.19.

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\text{Cov}(X, Y) \quad \text{and} \quad \mathbb{V}(X - Y) = \mathbb{V}(X) + \mathbb{V}(Y) - 2\text{Cov}(X, Y)$$

More generally, for random variables $X_1, \ldots, X_n$,

$$\mathbb{V}\left(\sum_i a_i X_i\right) = \sum_i a_i^2 \mathbb{V}(X_i) + 2\sum\sum_{i<j} a_i a_j \text{Cov}(X_i, X_j)$$

3.4 4.4 Expectation and Variance of Important Random Variables

Distribution	Mean	Variance
Point mass at p	a	0
Bernoulli(p)	p	$p(1-p)$
Binomial(n,p)	np	$np(1-p)$
Geometric(p)	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(β)	β	β^2
Gamma(α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta(α,β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$
t_ν	0 (if $\nu > 1$)	$\nu/(\nu-2)$ (if $\nu > 2$)
χ^2_p	p	$2p$
Multinomial(n,p)	np	see below
Multivariate Nornal(μ,Σ)	μ	Σ

The last two entries in the table are multivariate models which involve a random vector X of the form

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$$

The mean of a random vector X is defined by

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_k) \end{pmatrix}$$

The **variance-covariance matrix** Σ is defined to be

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_k) \\ \text{Cov}(X_2,X_1) & \mathbb{V}(X_2) & \cdots & \text{Cov}(X_2,X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_k,X_1) & \text{Cov}(X_k,X_2) & \cdots & \mathbb{V}(X_k) \end{pmatrix}$$

If $X \sim \text{Multinomial}(n,p)$ then

$$\mathbb{E}(X) = np = n(p_1,\ldots,p_k) \quad \text{and} \quad \mathbb{V}(X) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_2p_1 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_kp_1 & -np_kp_2 & \cdots & np_k(1-p_k) \end{pmatrix}$$

To see this:

- Note that the marginal distribution of any one component is $X_i \sim \text{Binomial}(n, p_i)$, so $\mathbb{E}(X_i) = np_i$ and $\mathbb{V}(X_i) = np_i(1-p_i)$.

- Note that, for $i \neq j$, $X_i + X_j \sim \text{Binomial}(n, p_i + p_j)$, so $\mathbb{V}(X_i + X_j) = n(p_i + p_j)(1 - (p_i + p_j))$.
- Using the formula for the covariance of a sum, for $i \neq j$,

$$\mathbb{V}(X_i + X_j) = \mathbb{V}(X_i) + \mathbb{V}(X_j) + 2\text{Cov}(X_i, X_j) = np_i(1-p_i) + np_j(1-p_j) + 2\text{Cov}(X_i, X_j)$$

Equating the last two formulas we get a formula for the covariance, $\text{Cov}(X_i, X_j) = -np_ip_j$.

Finally, here's a lemma that can be useful for finding means and variances of linear combinations of multivariate random vectors.

Lemma 4.20. If a is a vector and X is a random vector with mean μ and variance Σ then

$$\mathbb{E}(a^T X) = a^T \mu \quad \text{and} \quad \mathbb{V}(a^T X) = a^T \Sigma a$$

If A is a matrix then

$$\mathbb{E}(AX) = A\mu \quad \text{and} \quad \mathbb{V}(AX) = A\Sigma A^T$$

3.5 4.5 Conditional Expectation

The conditional expectation of X given $Y = y$ is

$$\mathbb{E}(X|Y=y) = \begin{cases} \sum x f_{X|Y}(x|y) & \text{discrete case} \\ \int x f_{X|Y}(x|y)dy & \text{continuous case} \end{cases}$$

If r is a function of x and y then

$$\mathbb{E}(r(X,Y)|Y=y) = \begin{cases} \sum r(x,y) f_{X|Y}(x|y) & \text{discrete case} \\ \int r(x,y) f_{X|Y}(x|y)dy & \text{continuous case} \end{cases}$$

While $\mathbb{E}(X)$ is a number, $\mathbb{E}(X|Y=y)$ is a function of y. Before we observe Y, we don't know the value of $\mathbb{E}(X|Y=y)$ so it is a random variable which we denote $\mathbb{E}(X|Y)$. In other words, $\mathbb{E}(X|Y)$ is the random variable whose value is $\mathbb{E}(X|Y=y)$ when Y is observed as y. Similarly, $\mathbb{E}(r(X,Y)|Y)$ is the random variable whose value is $\mathbb{E}(r(X,Y)|Y=y)$ when Y is observed as y.

Theorem 4.23 (The rule of iterated expectations). For random variables X and Y, assuming the expectations exist, we have that

$$\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y) \quad \text{and} \quad \mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X)$$

More generally, for any function $r(x,y)$ we have

$$\mathbb{E}[\mathbb{E}(r(X,Y)|X)] = \mathbb{E}(r(X,Y)) \quad \text{and} \quad \mathbb{E}[\mathbb{E}(r(X,Y)|Y)] = \mathbb{E}(r(X,Y))$$

Proof. We will prove the first equation.

$$\mathbb{E}[\mathbb{E}(Y|X)] = \int \mathbb{E}(Y|X=x) f_X(x) dx = \int\int y f(y|x) dy f(x) dx \tag{54}$$

$$= \int\int y f(y|x) f(x) dx dy = \int\int y f(x,y) dx dy = \mathbb{E}(Y) \tag{55}$$

The **conditional variance** is defined as

$$\mathbb{V}(Y|X=x) = \int (y - \mu(x))^2 f(y|x) dx$$

where $\mu(x) = \mathbb{E}(Y|X=x)$.

Theorem 4.26. For random variables X and Y,

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

3.6 4.6 Technical Appendix

4.6.1 Expectation as an Integral The integral of a measurable function $r(x)$ is defined as follows. First suppose that r is simple, meaning that it takes finitely many values $a_1, \ldots, a_k$ over a partition $A_1, \ldots, A_k$. Then $\int r(x) dF(x) = \sum_{i=1}^k a_i \mathbb{P}(r(X) \in A_i)$. The integral of a positive measurable function r is defined by $\int r(x) dF(x) = \lim_i \int r_i(x) dF(x)$, where r_i is a sequence of simple functions such that $r_i(x) \leq r(x)$ and $r_i(x) \to r(x)$ as $i \to \infty$. This does not depend on the particular sequence. The integral of a measurable function r is defined to be $\int r(x) dF(x) = \int r^+(x) dF(x) - \int r^-(x) dF(x)$ assuming both integrals are finite, where $r^+(x) = \max\{r(x), 0\}$ and $r^-(x) = \min\{r(x), 0\}$.

4.6.2 Moment Generating Functions The **moment generating function (mgf)** or **Laplace transform** of X is defined by

$$\psi_X(t) = \mathbb{E}(e^{tX}) = \int e^{tx} dF(x)$$

where t varies over the real numbers.

In what follows, we assume the mgf is well defined for all t in small neighborhood of 0. A related function is the characteristic function, defined by $\mathbb{E}(e^{itX})$ where $i = \sqrt{-1}$. This function is always defined for all t. The mgf is useful for several reasons. First, it helps us compute the moments of a distribution. Second, it helps us find the distribution of sums of random variables. Third, it is used to prove the central limit theorem.

When the mgf is well defined, it can be shown that we can interchange the operations of differentiation and "taking expectation". This leads to

$$\psi'(0) = \left[\frac{d}{dt}\mathbb{E}e^{tX}\right]_{t=0} = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right]_{t=0} = \mathbb{E}[Xe^{tX}]_{t=0} = \mathbb{E}(X)$$

By taking further derivatives we conclude that $\psi^{(k)}(0) = \mathbb{E}(X^k)$. This gives us a method for computing the moments of a distribution.

Lemma 4.30. Properties of the mgf.

1. If $Y = aX + b$ then $\$\psi\_Y(t) = e^{\{bt\}}\psi\_X(at)\$if X_1, \ldots, X_n$ are independent and $Y = \sum_i X_i$ then $\psi_Y(t) = \prod_i \psi_i(t)$, where ψ_i is the mgf of X_i.

Theorem 4.32. Let X and Y be random variables. If $\psi_X(t) = \psi_Y(t)$ for all t in an open interval around 0, then $X \stackrel{d}{=} Y$.

Moment Generating Function for Some Common Distributions

Distribution	mgf
Bernoulli(p)	$pe^t + (1-p)$
Binomial(n, p)	$(pe^t + (1-p))^n$
Poisson(λ)	$e^{\lambda(e^t-1)}$
Normal(μ, σ^2)	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$
Gamma(α, β)	$\left(\frac{\beta}{\beta-t}\right)^\alpha$ for $t < \beta$

3.7 4.7 Exercises

Exercise 4.7.1. Suppose we play a game where we start with c dollars. On each play of the game you either double your money or half your money, with equal probability. What is your expected fortune after n trials?

Solution. Let the random variables X_i be the fortune after the i-th trial, $X_0 = c$ always taking the value c. Then:

$$\mathbb{E}[X_{i+1}|X_i = x] = 2x \cdot \frac{1}{2} + \frac{x}{2} \cdot \frac{1}{2} = \frac{5}{4}x$$

Taking the expectation on X_i on both sides (i.e. integrating over $F_{X_i}(x)$),

$$\mathbb{E}(\mathbb{E}[X_{i+1}|X_i = x]) = \frac{5}{4}\mathbb{E}(X_i) \Longrightarrow \mathbb{E}(X_{i+1}) = \frac{5}{4}\mathbb{E}(X_i)$$

Therefore, by induction,

$$\mathbb{E}(X_n) = \left(\frac{5}{4}\right)^n c$$

Note that this is **not** a martingale, as in the traditional double-or-nothing formulation – the expected value goes up at each iteration.

Exercise 4.7.2. Show that $\mathbb{V}(X) = 0$ if and only if there is a constant c such that $\mathbb{P}(X = c) = 1$.

Solution. We have $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$:

$$\mathbb{V}(X) = \int (x - \mu_X)^2 dF_X(x)$$

Since $(x - \mu_X)^2 \geq 0$, in order for the variance to be 0 we must have the integrand be zero with probability 1, i.e. $\mathbb{P}(X = \mu_X) = 1$.

Exercise 4.7.3. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$ and let $Y_n = \max\{X_1, \ldots, X_n\}$. Find $\mathbb{E}(Y_n)$.

Solution. The CDF of Y_n, for $0 \leq y \leq 1$, is:

$$F_{Y_n}(y) = \mathbb{P}(Y_n \leq y) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq y) = y^n$$

so its PDF is $f_{Y_n}(y) = F'_{Y_n}(y) = ny^{n-1}$ for $0 \leq y \leq 1$.

The expected value of Y_n then is

$$\mathbb{E}(Y_n) = \int_0^1 y f_{Y_n}(y) dy = \int_0^1 n y^n dy = \frac{n}{n+1}$$

Exercise 4.7.4. 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 move 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.)

Solution.

We can define $X_n = \sum_{i=1}^{n}(1 - 2Y_i)$, where $Y_i \sim \text{Bernoulli}(p)$ and the Y_i's are independent random variables representing the direction of each jump.

We then have:

$$\mathbb{E}(X_n) = \sum_{i=1}^{n} \mathbb{E}(1 - 2Y_i) = \sum_{i=1}^{n}(1 - 2p) = n(1 - 2p)$$

and

$$\mathbb{V}(X_n) = \sum_{i=1}^{n} \mathbb{V}(1 - 2Y_i) \sum_{i=1}^{n} 4\mathbb{V}(Y_i) = 4np(1 - p)$$

Exercise 4.7.5. A fair coin is tossed until a head is obtained. What is the expected number of tosses that will be required?

Solution. The number of tosses follows a geometric distribution, $X \sim \text{Geom}(p)$, where p is the probability of heads. Let's deduce its expected value, rather than use it as a known fact ($\mathbb{E}(X) = 1/p$). The PDF is

$$f_X(k) = p(1-p)^{k-1}, \quad k > 0$$

The expected value for X is

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=1}^{\infty} kp(1-p)^{k-1} && (56)\\
&= \sum_{k=1}^{\infty} p(1-p)^{k-1} + \sum_{k=2}^{\infty}(k-1)p(1-p)^{k-1} && (57)\\
&= p\left(1 + (1-p) + (1-p)^2 + \dots\right) + \sum_{k=1}^{\infty} kp(1-p)^k && (58)\\
&= p\left(\frac{1}{1-(1-p)}\right) + (1-p)\sum_{k=1}^{\infty} kp(1-p)^{k-1} && (59)\\
&= 1 + (1-p)\mathbb{E}(X) && (60)
\end{aligned}$$

from where we get $\mathbb{E}(X) = 1/p$.

Exercise 4.7.6. Prove Theorem 4.6 for discrete random variables.

Let $Y = r(X)$. Then

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x)\, dF_X(x)$$

Solution. The result is immediate from the definition of expectation:

$$Y(\omega) = r(X(\omega)) = r(x) \quad \forall \omega : X(\omega) = x$$

and so

$$\mathbb{E}(Y) = \int r(x) dF_x(x)$$

Exercise 4.7.7. 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 $\lim_{x\to+\infty} x|1 - F(x)| = 0$.

Solution. Let's prove the following, slightly more general, lemma.

Lemma: For every continuous random variable X,

$$\mathbb{E}(X) = \int_0^\infty (1 - F_X(y))dy - \int_{-\infty}^0 F_X(y)dy$$

Proof:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{61}$$

$$= \int_{-\infty}^{0} \int_{x}^{0} -f_X(x) dy dx + \int_{0}^{\infty} \int_{0}^{x} f_X(x) dy dx \tag{62}$$

$$= -\int_{-\infty}^{0} \int_{-\infty}^{y} f_X(x) dx dy + \int_{0}^{\infty} \int_{y}^{\infty} f_X(x) dx dy \tag{63}$$

$$= -\int_{\infty}^{0} \mathbb{P}(X \leq y) dy + \int_{0}^{\infty} \mathbb{P}(X \geq y) dy \tag{64}$$

$$= \int_{0}^{\infty} (1 - F_X(y)) dy - \int_{-\infty}^{0} F_X(y) dy \tag{65}$$

The result follows by imposing $\mathbb{P}(X > 0) = 1$, which implies $\int_{-\infty}^{0} F_X(y) dy = 0$.

Exercise 4.7.8. Prove Theorem 4.16.

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

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

Solution.

For the expected value of sample mean:

$$\mathbb{E}\left(\overline{X}_n\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i) = \frac{1}{n} n\mu = \mu$$

For the variance of sample mean:

$$\mathbb{V}\left(\overline{X}_n\right) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

For the expected value of sample variance:

$$\mathbb{E}(S_n^2) = \mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right) \tag{66}$$

$$= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right) \tag{67}$$

$$= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - 2X_i\overline{X}_n + \overline{X}_n^2\right) \tag{68}$$

$$= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}_n\sum_{i=1}^{n} X_i + n\overline{X}_n^2\right) \tag{69}$$

$$= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}_n \cdot n\overline{X}_n + n\overline{X}_n^2\right) \tag{70}$$

$$= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right) \tag{71}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\mathbb{E}\left(\overline{X}_n^2\right)\right) \tag{72}$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\mathbb{V}(X_i) + \mathbb{E}(X_i)^2\right) - n\left(\mathbb{V}\left(\overline{X}_n\right) + \mathbb{E}\left(\overline{X}_n\right)^2\right)\right) \tag{73}$$

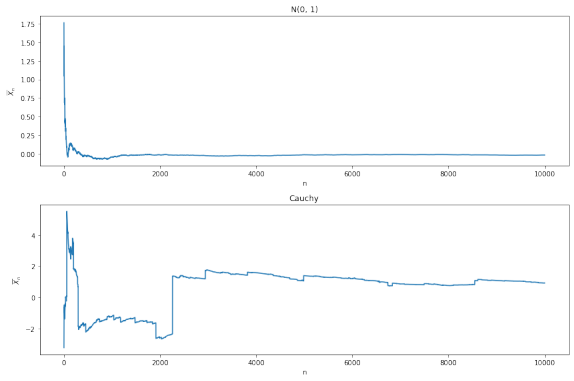
$$= \frac{1}{n-1}\left(n\left(\sigma^2 + \mu^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) \tag{74}$$

$$= \sigma^2 \tag{75}$$

Exercise 4.7.9 (Computer Experiment). Let $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_n \sim$ Cauchy. Explain why there is such a difference.

<MINTED>

<MINTED>



The mean on the Cauchy distribution is famously undefined: $\overline{X}_n$ is not going to converge.

Exercise 4.7.10. Let $X \sim N(0,1)$ and let $Y = e^X$. Find $\mathbb{E}(Y)$ and $\mathbb{V}(Y)$.

Solution.

The CDF of Y is, for $y > 0$:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq \log y) = \Phi(\log y)$$

and so the PDF is

$$f_Y(y) = F_Y'(y) = \frac{d}{dy}\Phi(\log y) = \frac{d\Phi(\log y)}{d\log y}\frac{d\log y}{dy} = \frac{\phi(\log y)}{y}$$

The expected value is

$$\mathbb{E}(Y) = \int y f_Y(y)dy = \int_0^\infty y\frac{\phi(\log y)}{y}dy = \int_0^\infty \phi(\log y)\ dy = \sqrt{e}$$

The expected value of Y^2 is

$$\mathbb{E}(Y^2) = \int y^2 f_Y(y)dy = \int_0^\infty y^2\frac{\phi(\log y)}{y}dy = \int_0^\infty y\phi(\log y)\ dy = e^2$$

and so the variance is

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = e(e-1)$$

Exercise 4.7.11 (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 runs look very different even though they were generated the same way. How do the calculations in (a) explain the second observation?

Solution.

(a) We have:

$$\mathbb{E}(X_n) = \mathbb{E}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \mathbb{E}(Y_i) = 0$$

and

$$\mathbb{E}(X_n^2) = \mathbb{E}\left(\left(\sum_{i=1}^n Y_i\right)^2\right) \tag{76}$$

$$= \mathbb{E}\left(\sum_{i=1}^n Y_i^2 + \sum_{i=1}^n \sum_{j=1, j\neq i}^n Y_i Y_j\right) \tag{77}$$

$$= \sum_{i=1}^n \mathbb{E}(Y_i^2) + \sum_{i=1}^n \sum_{j=1, j\neq i}^n \mathbb{E}(Y_i Y_j) \tag{78}$$

$$= \sum_{i=1}^n 1 + \sum_{i=1}^n \sum_{j=1, j\neq i}^n 0 \tag{79}$$

$$= n \tag{80}$$

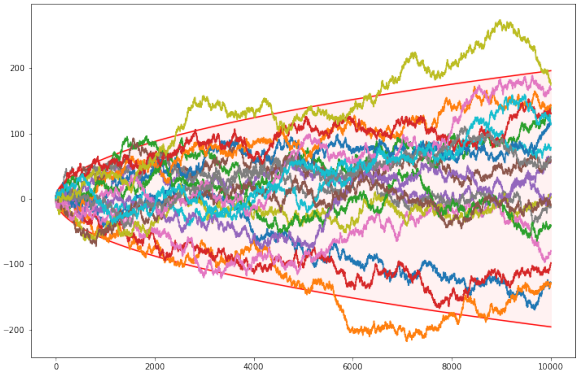
so

$$\mathbb{V}(X_n) = \mathbb{E}(X_n^2) - \mathbb{E}(X_n)^2 = n$$

(b)

<MINTED>

<MINTED>



The standard deviation is $\sqrt{n}$ – it scales up with the square root of the "time". The plot above draws $z_{\alpha/2}\sqrt{n}$ curves – confidence bands for $1 - \alpha = 95\%$ – that contain most of the randomly generated path.

Exercise 4.7.12. Prove the formulas given in the table at the beginning of Section 4.4 for the Bernoulli, Poisson, Uniform, Exponential, Gamma, and Beta. Here are some hints. For the mean of the Poisson, use the fact that $e^a = \sum_{x=0}^a a^x/x!$. 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 $\Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+\beta+1)$.

Solution.

We will do all expressions in the table instead (other than multinomial and multivariate normal, where proofs are already provided in the book).