Introduction to Probability for Electrical Engineering

Prapun Suksompong
School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853
ps92@cornell.edu

August 24, 2009

Contents

1 Mathematical Background 4
  1.1 Set Theory .................................................. 4
  1.2 Enumeration / Combinatorics / Counting ......................... 9
  1.3 Dirac Delta Function .......................................... 19

2 Classical Probability 20
  2.1 Examples ..................................................... 23

3 Probability Foundations 25
  3.1 Algebra and $\sigma$-algebra .................................. 25
  3.2 Kolmogorov’s Axioms for Probability .......................... 30
  3.3 Properties of Probability Measure ............................. 31
  3.4 Countable $\Omega$ ............................................. 34
  3.5 Independence .................................................. 35

4 Random Element 37
  4.1 Random Variable .............................................. 37
  4.2 Distribution Function ......................................... 38
  4.3 Discrete random variable ...................................... 41
  4.4 Continuous random variable .................................. 43
  4.5 Mixed/hybrid Distribution .................................... 45
  4.6 Independence .................................................. 46
  4.7 Misc .......................................................... 48

5 PMF Examples 48
  5.1 Random/Uniform ............................................... 49
  5.2 Bernoulli and Binary distributions .............................. 50
  5.3 Binomial: $B(n,p)$ ............................................. 51
  5.4 Geometric: $G(\beta)$ ........................................... 52
5.5 Poisson Distribution: $P(\lambda)$ ......................................................... 53
5.6 Compound Poisson ................................................................. 57
5.7 Hypergeometric ................................................................. 58
5.8 Negative Binomial Distribution (Pascal / Pólya distribution) ................ 59
5.9 Beta-binomial distribution ....................................................... 60
5.10 Zipf or zeta random variable .................................................. 60

6 PDF Examples
6.1 Uniform Distribution ............................................................. 60
6.2 Gaussian Distribution ........................................................... 62
6.3 Exponential Distribution ....................................................... 66
6.4 Pareto: Par($\alpha$)–heavy-tailed model/density ............................... 67
6.5 Laplacian: $L(\alpha)$ ................................................................. 68
6.6 Rayleigh ............................................................... 68
6.7 Cauchy ............................................................... 69
6.8 More PDFs ............................................................... 70

7 Expectation .............................................................. 71

8 Inequalities .............................................................. 78

9 Random Vectors
9.1 Random Sequence ............................................................. 88

10 Transform Methods
10.1 Probability Generating Function ................................................ 88
10.2 Moment Generating Function .................................................. 89
10.3 One-Sided Laplace Transform ................................................ 90
10.4 Characteristic Function ........................................................ 91

11 Functions of random variables
11.1 SISO case ............................................................... 93
11.2 MISO case ............................................................... 98
11.3 MIMO case ............................................................... 100
11.4 Order Statistics ............................................................. 103

12 Convergences
12.1 Summation of random variables ................................................ 116
12.2 Summation of independent random variables .................................. 116
12.3 Summation of i.i.d. random variable ........................................... 117
12.4 Central Limit Theorem (CLT) .................................................. 118

13 Conditional Probability and Expectation
13.1 Conditional Probability .......................................................... 119
13.2 Conditional Expectation ........................................................ 120
13.3 Conditional Independence ...................................................... 121
14 Real-valued Jointly Gaussian

15 Bayesian Detection and Estimation

A Math Review

A.1 Inequalities .................... 128
A.2 Summations .................... 131
A.3 Calculus ....................... 133
  A.3.1 Derivatives ................. 133
  A.3.2 Integration ............... 137
A.4 Gamma and Beta functions ..... 140
1 Mathematical Background

1.1 Set Theory

1.1. Basic Set Identities:

- Idempotence: \((A^c)^c = A\)
- Commutativity (symmetry):
  \[
  A \cup B = B \cup A, \quad A \cap B = B \cap A
  \]
- Associativity:
  \[
  \begin{align*}
  A \cap (B \cap C) &= (A \cap B) \cap C \\
  A \cup (B \cup C) &= (A \cup B) \cup C
  \end{align*}
  \]
- Distributivity
  \[
  \begin{align*}
  A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) \\
  A \cap (B \cup C) &= (A \cap B) \cup (A \cap C)
  \end{align*}
  \]
- de Morgan laws
  \[
  \begin{align*}
  (A \cup B)^c &= A^c \cap B^c \\
  (A \cap B)^c &= A^c \cup B^c
  \end{align*}
  \]

1.2. Basic Terminology:

- \(A \cap B\) is sometimes written simply as \(AB\).
- Sets \(A\) and \(B\) are said to be disjoint \((A \perp B)\) if and only if \(A \cap B = \emptyset\)
- A collection of sets \((A_i : i \in I)\) is said to be pair-wise disjoint or mutually exclusive \([9, p. 9]\) if and only if \(A_i \cap A_j = \emptyset\) when \(i \neq j\).
- A collection \(\Pi = (A_\alpha : \alpha \in I)\) of subsets of \(\Omega\) (in this case, indexed or labeled by \(\alpha\) taking values in an index or label set \(I\)) is said to be a partition of \(\Omega\) if
  \[
  \begin{align*}
  (a) & \quad \Omega = \bigcup_{\alpha \in I} A_\alpha \\
  (b) & \quad \text{For all } i \neq j, \ A_i \perp A_j \text{ (pairwise disjoint)}.
  \end{align*}
  \]
  In which case, the collection \((B \cap A_\alpha : \alpha \in I)\) is a partition of \(B\). In other words, any set \(B\) can be expressed as \(B = \bigcup_{\alpha} (B \cap A_\alpha)\) where the union is a disjoint union.
- The cardinality (or size) of a collection or set \(A\), denoted \(|A|\), is the number of elements of the collection. This number may be finite or infinite.
<table>
<thead>
<tr>
<th>name</th>
<th>rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commutative laws</td>
<td>$A \cap B = B \cap A$ $A \cup B = B \cup A$</td>
</tr>
<tr>
<td>Associative laws</td>
<td>$A \cap (B \cap C) = (A \cap B) \cap C$</td>
</tr>
<tr>
<td></td>
<td>$A \cup (B \cup C) = (A \cup B) \cup C$</td>
</tr>
<tr>
<td>Distributive laws</td>
<td>$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$</td>
</tr>
<tr>
<td></td>
<td>$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$</td>
</tr>
<tr>
<td>DeMorgan’s laws</td>
<td>$\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\overline{A \cup B} = \overline{A} \cap \overline{B}$</td>
</tr>
<tr>
<td>Complement laws</td>
<td>$A \cap \overline{A} = \emptyset$ $A \cup \overline{A} = \Omega$</td>
</tr>
<tr>
<td>Double complement law</td>
<td>$\overline{\overline{A}} = A$</td>
</tr>
<tr>
<td>Idempotent laws</td>
<td>$A \cap A = A$ $A \cup A = A$</td>
</tr>
<tr>
<td>Absorption laws</td>
<td>$A \cap (A \cup B) = A$ $A \cup (A \cap B) = A$</td>
</tr>
<tr>
<td>Dominance laws</td>
<td>$A \cap \emptyset = \emptyset$ $A \cup \Omega = \Omega$</td>
</tr>
<tr>
<td>Identity laws</td>
<td>$A \cup \emptyset = A$ $A \cap \Omega = A$</td>
</tr>
</tbody>
</table>

Figure 1: Set Identities

- **Inclusion-Exclusion Principle:**
  $$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{\phi \not\subset \{1, \ldots, n\}} (-1)^{|I|+1} \left| \bigcap_{i \in I} A_i \right|.$$  

- $$\left| \bigcap_{i=1}^{n} A_i^c \right| = |\Omega| + \sum_{\phi \not\subset \{1, \ldots, n\}} (-1)^{|I|} \left| \bigcap_{i \in I} A_i \right|.$$  

- If $\forall i, A_i \subset B$ (or equivalently, $\bigcup_{i=1}^{n} A_i \subset B$), then  
  $$\left| \bigcap_{i=1}^{n} (B \setminus A_i) \right| = |B| + \sum_{\phi \not\subset \{1, \ldots, n\}} (-1)^{|I|} \left| \bigcap_{i \in I} A_i \right|.$$  

- An infinite set $A$ is said to be **countable** if the elements of $A$ can be enumerated or listed in a sequence: $a_1, a_2, \ldots$. Empty set and finite sets are also said to be countable. By a countably infinite set, we mean a countable set that is not finite.  

- A **singleton** is a set with exactly one element.  

- $\mathbb{N} = \{1, 2, 3, \ldots\}$  
  $\mathbb{R} = (-\infty, \infty).$  

- For a set of sets, to avoid the repeated use of the word “set”, we will call it a collection/class/family of sets.  

**Definition 1.3.** Monotone sequence of sets
The sequence of events \( (A_1, A_2, A_3, \ldots) \) is monotone-increasing sequence of events if and only if
\[
A_1 \subset A_2 \subset A_3 \subset \ldots
\]
In which case,
\[
\bigcup_{i=1}^{n} A_i = A_n
\]
\[
\lim_{n \to \infty} A_n = \bigcup_{i=1}^{\infty} A_i.
\]
Put \( A = \lim_{n \to \infty} A_n \). We then write \( A_n \nearrow A \); that is \( A_n \nearrow A \) if and only if \( \forall n \)
\( A_n \subset A_{n+1} \) and \( \bigcup_{n=1}^{\infty} A_n = A \).

The sequence of events \( (B_1, B_2, B_3, \ldots) \) is monotone-decreasing sequence of events if and only if
\[
B_1 \supset B_2 \supset B_3 \supset \ldots
\]
In which case,
\[
\bigcap_{i=1}^{n} B_i = B_n
\]
\[
\lim_{n \to \infty} B_n = \bigcap_{i=1}^{\infty} B_i
\]
Put \( B = \lim_{n \to \infty} B_n \). We then write \( B_n \searrow B \); that is \( B_n \searrow B \) if and only if \( \forall n \)
\( B_{n+1} \subset B_n \) and \( \bigcap_{n=1}^{\infty} B_n = B \).

Note that \( A_n \searrow A \Leftrightarrow A_n^c \nearrow A^c \).

1.4. An (Event-)indicator function \( I_A : \Omega \to \{0, 1\} \) is defined by
\[
I_A(\omega) = \begin{cases} 
1, & \text{if } \omega \in A \\
0, & \text{otherwise.}
\end{cases}
\]

- Alternative notation: \( 1_A \).
- \( A = \{ \omega : I_A(\omega) = 1 \} \)
- \( A = B \) if and only if \( I_A = I_B \)
- \( I_{A^c}(\omega) = 1 - I_A(\omega) \)
- \( A \subset B \Leftrightarrow \{ \forall \omega, I_A(\omega) \leq I_B(\omega) \} \Leftrightarrow \{ \forall \omega, I_A(\omega) = 1 \Rightarrow I_B(\omega) = 1 \} \)
  \[
  I_{A \cap B}(\omega) = \min(I_A(\omega), I_B(\omega)) = I_A(\omega) \cdot I_B(\omega)
  \]
  \[
  I_{A \cup B}(\omega) = \max(I_A(\omega), I_B(\omega)) = I_A(\omega) + I_B(\omega) - I_A(\omega) \cdot I_B(\omega)
  \]
1.5. Suppose $\bigcup_{i=1}^{N} A_i = \bigcup_{i=1}^{N} B_i$ for all finite $N \in \mathbb{N}$. Then, $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$.

Proof. To show “$\subset$”, suppose $x \in \bigcup_{i=1}^{\infty} A_i$. Then, $\exists N_0$ such that $x \in A_{N_0}$. Now, $A_{N_0} \subset \bigcup_{i=1}^{N_0} A_i = \bigcup_{i=1}^{N_0} B_i \subset \bigcup_{i=1}^{\infty} B_i$. Therefore, $x \in \bigcup_{i=1}^{\infty} B_i$. To show “$\supset$”, use symmetry. \qed

1.6. Let $A_1, A_2, \ldots$ be a sequence of disjoint sets. Define $B_n = \bigcup_{i>n} A_i = \bigcup_{i>n} A_i$. Then, (1) $B_{n+1} \subset B_n$ and (2) $\cap_{n=1}^{\infty} B_n = \emptyset$.

Proof. $B_n = \bigcup_{i>n} A_i = (\bigcup_{i>n} A_i) \cup A_{n+1} = B_{n+1} \cup A_{n+1}$ So, (1) is true. For (2), consider two cases. (2.1) For element $x \notin \bigcup_i A_i$, we know that $x \notin B_n$ and hence $x \notin \cap_{n=1}^{\infty} B_n$. (2.2) For $x \in \bigcup_i A_i$, we know that $x \in A_{i_0}$. Note that $x$ can’t be in other $A_i$ because the $A_i$’s are disjoint. So, $x \notin B_{i_0+1}$ and therefore $x \notin \cap_{n=1}^{\infty} B_n$. \qed

1.7. Any countable union can be written as a union of pairwise disjoint sets: Given any sequence of sets $F_i$, define a new sequence by $A_1 = F_1$, and

$$A_n = F_n \cap F_{n-1}^c \cap \ldots \cap F_1^c = F_n \cap \bigcap_{i=[n-1]} \left( F_i^c \right) \bigcup \left( \bigcup_{i=[n-1]} F_i \right),$$

for $n \geq 2$. Then, $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} A_i$ where the union on the RHS is a disjoint union.

Proof. Note that the $A_n$ are pairwise disjoint. To see this, consider $A_{n_1}, A_{n_2}$ where $n_1 \neq n_2$. WLOG, assume $n_1 < n_2$. This implies that $F_{n_1}^c$ get intersected in the definition of $A_{n_2}$. So,

$$A_{n_1} \cap A_{n_2} = (F_{n_1} \cap F_{n_1}^c) \bigcap \left( \bigcap_{i=1}^{n_1-1} F_i^c \cap F_{n_2} \cap \bigcap_{i=1}^{n_2-1} F_i^c \right) = \emptyset$$

Also, for finite $N \geq 1$, we have $\bigcup_{n \in [N]} F_n = \bigcup_{n \in [N]} A_n$ (*). To see this, note that (*) is true for $N = 1$. Suppose (*) is true for $N = m$ and let $B = \bigcup_{n \in [m]} A_n = \bigcup_{n \in [m]} F_n$. Now, for $N = m + 1$, by definition, we have

$$\bigcup_{n \in [m+1]} A_n = \bigcup_{n \in [m]} A_n \cup A_{m+1} = (B \cup F_{m+1}) \cap \Omega$$

$$= B \cup F_{m+1} = \bigcup_{i \in [m+1]} F_i.$$

So, (*) is true for $N = m + 1$. By induction, (*) is true for all finite $N$. The extension to $\infty$ is done via (1.5). \qed

For finite union, we can modify the above statement by setting $F_n = \emptyset$ for $n \geq N$. Then, $A_n = \emptyset$ for $n \geq N$.

By construction, $\{A_n : n \in \mathbb{N}\} \subset \sigma (\{F_n : n \in \mathbb{N}\})$. However, in general, it is not true that $\{F_n : n \in \mathbb{N}\} \subset \sigma (\{A_n : n \in \mathbb{N}\})$. For example, for finite union with $N = 2$, we can’t get $F_2$ back from set operations on $A_1, A_2$ because we lost information about $F_1 \cap F_2$. To create a disjoint union which preserve the information about the overlapping parts of the
we can define the $A$’s by $\cap_{n \in \mathbb{N}} B_n$ where $B_n$ is $F_n$ or $F_n^c$. This is done in (1.8). However, this leads to uncountably many $A_\alpha$, which is why we used the index $\alpha$ above instead of $n$. The uncountability problem does not occur if we start with a finite union. This is shown in the next result.

1.8. Decomposition:

- Fix sets $A_1, A_2, \ldots, A_n$, not necessarily disjoint. Let $\Pi$ be a collection of all sets of the form
  \[ B = B_1 \cap B_2 \cap \cdots \cap B_n \]
  where each $B_i$ is either $A_j$ or its complement. There are $2^n$ of these, say $B^{(1)}, B^{(2)}, \ldots, B^{(2^n)}$.

Then,

(a) $\Pi$ is a partition of $\Omega$ and
(b) $\Pi \setminus \{ \cap_{j \in [n]} A_j \}$ is a partition of $\cup_{j \in [n]} A_j$.

Moreover, any $A_j$ can be expressed as $A_j = \bigcup_{i \in S_j} B^{(i)}$ for some $S_i \subset [2^n]$. More specifically, $A_j$ is the union of all $B^{(i)}$ which is constructed by $B_j = A_j$.

- Fix sets $A_1, A_2, \ldots$, not necessarily disjoint. Let $\Pi$ be a collection of all sets of the form
  \[ B = \bigcap_{n \in \mathbb{N}} B_n \]
  where each $B_n$ is either $A_n$ or its complement. There are uncountably of these hence we will index the $B$’s by $\alpha$; that is we write $B^{(\alpha)}$. Let $I$ be the set of all $\alpha$ after we eliminate all repeated $B^{(\alpha)}$; note that $I$ can still be uncountable. Then,

(a) $\Pi = \{ B^{(\alpha)} : \alpha \in I \}$ is a partition of $\Omega$ and
(b) $\Pi \setminus \{ \cap_{n \in \mathbb{N}} A_n^c \}$ is a partition of $\cup_{n \in \mathbb{N}} A_n$.

Moreover, any $A_j$ can be expressed as $A_j = \bigcup_{\alpha \in S_j} B^{(\alpha)}$ for some $S_i \subset I$. Because $I$ is uncountable, in general, $S_j$ can be uncountable. More specifically, $A_j$ is the (possibly uncountable) union of all $B^{(\alpha)}$ which is constructed by $B_j = A_j$. The uncountability of $S_j$ can be problematic because it implies that we need uncountable union to get $A_j$ back.

1.9. Let $\{ A_\alpha : \alpha \in I \}$ be a collection of disjoint sets where $I$ is a nonempty index set. For any set $S \subset I$, define a mapping $g$ on $2^I$ by $g(S) = \cup_{\alpha \in S} A_\alpha$. Then, $g$ is a 1:1 function if and only if none of the $A_\alpha$’s is empty.

1.10. Let

\[ A = \{(x, y) \in \mathbb{R}^2 : (x + a_1, x + b_1) \cap (y + a_2, y + b_2) \neq \emptyset \} \]

where $a_i < b_i$. Then,

\[ A = \{(x, y) \in \mathbb{R}^2 : x + (a_1 - b_2) < y < x + (b_1 - a_2) \} = \{(x, y) \in \mathbb{R}^2 : a_1 - b_2 < y - x < b_1 - a_2 \} . \]
1.2 Enumeration / Combinatorics / Counting

1.11. The four kinds of counting problems are:

(a) ordered sampling of $r$ out of $n$ items with replacement: $n^r$;

(b) ordered sampling of $r \leq n$ out of $n$ items without replacement: $(n)_r$;

(c) unordered sampling of $r \leq n$ out of $n$ items without replacement: $\binom{n}{r}$;

(d) unordered sampling of $r$ out of $n$ items with replacement: $\binom{n+r-1}{r}$.

1.12. Given a set of $n$ distinct items, select a distinct ordered sequence (word) of length $r$ drawn from this set.

- Sampling with replacement: $\mu_{n,r} = n^r$
  - Ordered sampling of $r$ out of $n$ items with replacement.
    - $\mu_{n,1} = n$
    - $\mu_{1,r} = 1$
    - $\mu_{n,r} = \mu_{n,r-1}$ for $r > 1$
    - Examples:
      * Suppose $A$ is a finite set, then the cardinality of its power set is $|2^A| = 2^{|A|}$.
      * There are $2^r$ binary strings/sequences of length $r$.  

Figure 2: The region $A = \{(x, y) \in \mathbb{R}^2 : (x+a_1, x+b_1) \cap (y+a_2, y+b_2) \neq \emptyset\}$. 
Sampling without replacement:

\[(n)_r = \prod_{i=0}^{r-1} (n - i) = \frac{n!}{(n - r)!}\]

\[= n \cdot (n - 1) \cdots (n - (r - 1)); \quad r \leq n\]

- Ordered sampling of \(r \leq n\) out of \(n\) items without replacement.
- For integers \(r, n\) such that \(r > n\), we have \((n)_r = 0\).
- The definition in product form \((n)_r = \prod_{i=0}^{r-1} (n - i) = n \cdot (n - 1) \cdots (n - (r - 1))\) can be extended to any real number \(n\) and a non-negative integer \(r\). We define \((n)_0 = 1\). (This makes sense because we usually take the empty product to be 1.)
- \((n)_1 = n\)
- \((n)_r = (n - (r - 1))(n)_{r-1}\). For example, \((7)_5 = (7 - 4)(7)_4\).
- \((1)_r = \begin{cases} 1, & \text{if } r = 1 \\ 0, & \text{if } r > 1 \end{cases}\)

Ratio:

\[\frac{(n)_r}{n^r} = \frac{\prod_{i=0}^{r-1} (n - i)}{\prod_{i=0}^{r-1} (n)} = \prod_{i=0}^{r-1} \left(1 - \frac{i}{n}\right)\]

\[\approx \prod_{i=0}^{r-1} e^{-\frac{i}{n}} = e^{-\frac{1}{n} \sum_{i=0}^{r-1} i} = e^{-\frac{r(r-1)}{2n}}\]

\[\approx e^{-\frac{r^2}{2n}}\]

1.13. **Factorial and Permutation**: The number of arrangements (permutations) of \(n \geq 0\) distinct items is \((n)_n = n!\).

- \(0! = 1! = 1\)
- \(n! = n(n - 1)!\)
- \(n! = \int_0^\infty e^{-t} t^n \, dt\)
- Stirling’s Formula:

\[n! \approx \sqrt{2\pi nn} e^{-n} = \left(\sqrt{2\pi e}\right) e^{(n+\frac{1}{2}) \ln(n)}\].
1.14. Binomial coefficient:

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!}
\]

This gives the number of unordered sets of size \(r\) drawn from an alphabet of size \(n\) without replacement; this is unordered sampling of \(r \leq n\) out of \(n\) items without replacement. It is also the number of subsets of size \(r\) that can be formed from a set of \(n\) elements. Some properties are listed below:

(a) Use \texttt{nchoosek(n,r)} in MATLAB.

(b) Use \texttt{combin(n,r)} in Mathcad. However, to do symbolic manipulation, use the factorial definition directly.

(c) Reflection property: \(\binom{n}{r} = \binom{n}{n-r}\).

(d) \(\binom{n}{n} = \binom{n}{0} = 1\).

(e) \(\binom{n}{1} = \binom{n}{n-1} = n\).

(f) \(\binom{n}{r} = 0\) if \(n < r\) or \(r\) is a negative integer.

(g) \(\max_r \binom{n}{r} = \binom{n}{\lfloor \frac{n+1}{2} \rfloor}\).

(h) Pascal’s “triangle” rule: \(\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}\). This property divides the process of choosing \(k\) items into two steps where the first step is to decide whether to choose the first item or not.

![Pascal Triangle](image)

Figure 3: Pascal Triangle.

(i) \(\sum_{k \text{ even}} \binom{n}{k} = \sum_{k \text{ odd}} \binom{n}{k} = 2^{n-1}\)

There are many ways to show this identity.
(i) Consider the number of subsets of \( S = \{a_1, a_2, \ldots, a_n\} \) of \( n \) distinct elements. First choose subsets \( A \) of the first \( n - 1 \) elements of \( S \). There are \( 2^{n-1} \) distinct \( S \). Then, for each \( A \), to get set with even number of element, add element \( a_n \) to \( A \) if and only if \( |A| \) is odd.

(ii) Look at binomial expansion of \((x + y)^n\) with \( x = 1 \) and \( y = -1 \).

(iii) For odd \( n \), use the fact that \( \binom{n}{r} = \binom{n}{n-r} \).

(j) \[ \sum_{k=0}^{\min(n_1, n_2)} \binom{n_1}{k} \binom{n_2}{n_k} = \binom{n_1 + n_2}{n_1} = \binom{n_1 + n_2}{n_2}. \]

- This property divides the process of choosing \( k \) items from \( n_1 + n_2 \) into two steps where the first step is to choose from the first \( n_1 \) items.
- Can replace the \( \min(n_1, n_2) \) in the first sum by \( n_1 \) or \( n_2 \) if we define \( \binom{n}{k} = 0 \) for \( k > n \).
- \[ \sum_{r=1}^{n} \binom{n}{r}^2 = \left( \frac{2n}{n} \right). \]

(k) Parallel summation:
\[
\sum_{m=k}^{n} \binom{m}{k} = \binom{k}{k} + \binom{k+1}{k} + \cdots + \binom{n}{k} = \binom{n+1}{k+1}.
\]

To see this, suppose we try to choose \( k + 1 \) items from \( n + 1 \) items \( a_1, a_2, \ldots, a_{n+1} \). First, we choose whether to choose \( a_1 \). If so, then we need to choose the rest \( k \) items from the \( n \) items \( a_2, \ldots, a_{n+1} \). Hence, we have the \( \binom{n}{k} \) term. Now, suppose we didn’t choose \( a_1 \). Then, we still need to choose \( k + 1 \) items from \( n \) items \( a_2, \ldots, a_{n+1} \). We then repeat the same argument on \( a_2 \) in stead of \( a_1 \).

Equivalently,
\[
\sum_{k=0}^{r} \binom{n+k}{k} = \sum_{k=0}^{n} \binom{n+k}{n} = \binom{n+r+1}{n+1} + \binom{n+r+1}{r}.
\]

To prove the middle equality in (1), use induction on \( r \).

1.15. **Binomial theorem:**
\[
(x + y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r}
\]

(a) Let \( x = y = 1 \), then \[ \sum_{r=0}^{n} \binom{n}{r} = 2^n. \]

(b) Sum involving only the even terms (or only the odd terms):
\[
\sum_{r=0, r \text{ even}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2} ((x + y)^n + (y - x)^n), \text{ and}
\]
\[
\sum_{r=0, r \text{ odd}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2} ((x + y)^n - (y - x)^n).
\]
In particular, if $x + y = 1$, then

$$\sum_{r=0 \atop r \text{ even}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2} (1 + (1 - 2x)^n), \quad \text{and}$$

$$\sum_{r=0 \atop r \text{ odd}}^{n} \binom{n}{r} x^r y^{n-r} = \frac{1}{2} (1 - (1 - 2x)^n). \quad \text{(2a)}$$

(c) Approximation by the entropy function:

$$H(p) = - p \log_b (p) - (1-p) \log_b (1-p).$$

- Binary: $b = 2 \Rightarrow$

$$H_2(p) = - p \log_2 (p) - (1-p) \log_2 (1-p).$$

For $x + y = 1$, we have

$$\sum_{r=0}^{n} \binom{n}{r} x^r (1-x)^{n-r} = nx$$

and

$$\sum_{r=0}^{n} r^2 \binom{n}{r} x^r (1-x)^{n-r} = nx(nx + 1 - x).$$

All identities above can be verified easily via Mathcad.

1.16. **Multinomial Counting**: The multinomial coefficient $\binom{n}{n_1, n_2, \ldots, n_r}$ is defined as

$$\prod_{i=1}^{r} \left( \frac{n - \sum_{k=0}^{i-1} n_k}{n_i} \right)$$

$$= \left( \frac{n}{n_1} \right) \cdot \left( \frac{n-n_1}{n_2} \right) \cdot \left( \frac{n-n_1-n_2}{n_3} \right) \cdots \left( \frac{n-n_1-n_2-\cdots-n_{r-1}}{n_r} \right)$$

$$= \frac{n!}{\prod_{i=1}^{r} n!}$$

It is the number of ways that we can arrange $n = \sum_{i=1}^{r} n_i$ tokens when having $r$ types of symbols and $n_i$ indistinguishable copies/tokens of a type $i$ symbol.
<table>
<thead>
<tr>
<th>Identity Type</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factorial expansion</td>
<td>( \binom{n}{k} = \frac{n!}{k!(n-k)!} ), ( k = 0, 1, 2, \ldots, n )</td>
</tr>
<tr>
<td>Symmetry</td>
<td>( \binom{n}{k} = \binom{n}{n-k} ), ( k = 0, 1, 2, \ldots, n )</td>
</tr>
<tr>
<td>Monotonicity</td>
<td>( \binom{n}{0} &lt; \binom{n}{1} &lt; \cdots &lt; \binom{n}{\lfloor n/2 \rfloor} ), ( n \geq 0 )</td>
</tr>
<tr>
<td>Pascal’s identity</td>
<td>( \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} ), ( k = 0, 1, 2, \ldots, n )</td>
</tr>
<tr>
<td>Binomial theorem</td>
<td>((x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \ n \geq 0 )</td>
</tr>
<tr>
<td>Counting all subsets</td>
<td>( \sum_{k=0}^{n} \binom{n}{k} = 2^n, \ n \geq 0 )</td>
</tr>
<tr>
<td>Even and odd subsets</td>
<td>( \sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0, \ n \geq 0 )</td>
</tr>
<tr>
<td>Sum of squares</td>
<td>( \sum_{k=0}^{n} \binom{n}{k}^2 = \sum_{k=0}^{2n} \binom{2n}{k}, \ n \geq 0 )</td>
</tr>
<tr>
<td>Absorption/extraction</td>
<td>( \binom{n}{k} = \frac{n}{k} \binom{n-1}{k-1}, \ k \neq 0 )</td>
</tr>
<tr>
<td>Trinomial revision</td>
<td>( \binom{n}{m} \binom{m}{k} = \binom{n-k}{m-k} ), ( 0 \leq k \leq m \leq n )</td>
</tr>
<tr>
<td>Parallel summation</td>
<td>( \sum_{k=0}^{m} \binom{n+k}{k} = \binom{n+m+1}{m}, \ m,n \geq 0 )</td>
</tr>
<tr>
<td>Diagonal summation</td>
<td>( \sum_{k=0}^{n-m} \binom{m+k}{m} = \binom{n+1}{m+1}, \ n \geq m \geq 0 )</td>
</tr>
<tr>
<td>Vandermonde convolution</td>
<td>( \sum_{k=0}^{r} \binom{m}{r-k} \binom{n}{r} = \binom{m+n}{r}, \ m,n,r \geq 0 )</td>
</tr>
<tr>
<td>Diagonal sums in Pascal’s triangle</td>
<td>( \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n-k}{k} = F_{n+1} ) (Fibonacci numbers), ( n \geq 0 )</td>
</tr>
<tr>
<td>Other Common Identities</td>
<td>( \sum_{k=0}^{n} k \binom{n}{k} = n2^{n-1}, \ n \geq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{n} k^2 \binom{n}{k} = n(n+1)2^{n-2}, \ n \geq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{n} (-1)^k k \binom{n}{k} = 0, \ n \geq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{n} \binom{n}{k} \binom{k}{k+1} = \frac{2^{n+1} - 1}{n+1}, \ n \geq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{n} (-1)^k \frac{n}{k+1} = \frac{1}{n+1}, \ n \geq 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=1}^{n} (-1)^{k-1} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}, \ n &gt; 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{n-1} \binom{n}{k} \binom{n}{k+1} = \binom{2n}{n}, \ n &gt; 0 )</td>
</tr>
<tr>
<td></td>
<td>( \sum_{k=0}^{m} \binom{m}{k} \binom{n}{k+p} = \binom{m+n}{m+p}, \ m,n,p \geq 0, \ n \geq p + m )</td>
</tr>
</tbody>
</table>

Figure 4: Binomial coefficient identities [18]
1.17. Multinomial Theorem

\[(x_1 + \ldots + x_r)^n = \sum_{i_1=0}^{n} \sum_{i_2=0}^{n-i_1} \cdots \sum_{i_{r-1}=0}^{n-i_{r-2}} \frac{n!}{(n - \sum_{k<i} i_k)! \prod_{k<i} i_k!} x_r^{\sum_{k<i} i_k} \]

- \(r\)-ary entropy function: Consider any vector \(p = (p_1, p_2, \ldots, p_r)\) such that \(p_i \geq 0\) and \(\sum_{i=1}^{r} p_i = 1\). We define

\[H(p) = -\sum_{i=1}^{r} p_i \log_b p_i.\]

As a special case, let \(p_i = \frac{n_i}{n}\), then

\[\begin{pmatrix} n \\ n_1 \ n_2 \ \cdots \ n_r \end{pmatrix} = \frac{n!}{\prod_{i=1}^{r} n_i!} \approx 2^{nH_2(p)} \approx 2^{\log_2(n!)}.\]

1.18. The number of solutions to \(x_1 + x_2 + \cdots + x_n = k\) for the \(x_i\)'s are nonnegative integers is \(\binom{k+n-1}{n-1}\).

(a) Suppose we further require that the \(x_i\) are strictly positive (\(x_i > 1\)), then there are \(\binom{k}{n-1}\) solutions.

(b) **Extra Lower-bound Requirement**: Suppose we further require that \(x_i \geq a_i\) where the \(a_i\) are some given nonnegative integers, then the number of solution is \(\binom{k-(a_1+a_2+\ldots+a_n)}{n-1}\). Note that here we work with equivalent problem: \(y_1 + y_2 + \cdots + y_n = k - \sum_{i=1}^{n} a_i\) where \(y_i \geq 0\).

(c) **Extra Upper-bound Requirement**: Suppose we further require that \(0 \leq x_i < b_i\). Let \(A_i\) be the set of solutions such that \(x_i \geq b_i\) and \(x_j \geq 0\) for \(j \neq i\). The number of solutions is \(\binom{k+n-1}{n-1} - |\bigcup_{i=1}^{n} A_i|\) where the second term can be found via the inclusion/exclusion principle

\[|\bigcup_{i\in[n]} A_i| = \sum_{I\subseteq[n], |I|\neq0} (-1)^{|I|+1} \left|\bigcap_{i\in I} A_i\right| \]

and the fact that for any index set \(I \subseteq [n]\), we have \(\left|\bigcap_{i\in I} A_i\right| = \binom{k-(\sum_{i\in I} b_i)}{n-1}\).

(d) **Extra Range Requirement**: Suppose we further require that \(a_i \leq x_i < b_i\) where \(0 \leq a_i < b_i\), then we work instead with \(y_i = x_i - a_i\). The number of solutions is

\[\left(\binom{k-n \sum_{i=1}^{n} a_i}{n-1} + n - 1\right) + \sum_{I\subseteq[n], |I|\neq0} (-1)^{|I|} \left(\binom{k-n \sum_{i=1}^{n} a_i}{n-1} - \left(\sum_{i\in I} (b_i-a_i)\right) + n - 1\right).\]
2.1 SUMMARY OF COUNTING PROBLEMS

• Consider the distribution of \( r = 10 \) indistinguishable balls into \( n = n \) distinguishable cells. Then, we only concern with the number of balls in each cell. Using \( n - 1 = 4 \) bars, we can divide \( r = 10 \) stars into \( n = 5 \) groups. For example, \( ****|***|**|*\) would mean \( (4,3,0,2,1) \). In general, there are \( \binom{n+r-1}{r} \) ways of arranging the bars and stars.

• There are \( \binom{n+r-1}{r} \) distinct vector \( x = x^n_r \) of nonnegative integers such that \( x_1 + x_2 + \cdots + x_n = r \). We use \( n - 1 \) bars to separate \( r \) 1’s.

• Suppose \( r \) letters are drawn with replacement from a set \( \{a_1, a_2, \ldots, a_n\} \). Given a drawn sequence, let \( x_i \) be the number of \( a_i \) in the drawn sequence. Then, there are \( \binom{n+r-1}{r} \) possible \( x = x^n_r \).

<table>
<thead>
<tr>
<th>objects</th>
<th>number of objects</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arranging objects in a row:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( n ) distinct objects</td>
<td>( n! = P(n,n) = n(n-1) \cdots 2 \cdot 1 )</td>
<td>§2.3.1</td>
</tr>
<tr>
<td>( k ) out of ( n ) distinct objects</td>
<td>( \binom{n}{k} = P(n,k) = n(n-1) \cdots (n-k+1) )</td>
<td>§2.3.1</td>
</tr>
<tr>
<td>some of the ( n ) objects are identical:</td>
<td>( \binom{n}{k_1 k_2 \ldots k_j} = \binom{n}{k_1} \binom{n-k_1}{k_2} \ldots \binom{n-k_1-k_2}{k_j} )</td>
<td>§2.3.2</td>
</tr>
<tr>
<td>( k ) of a first kind, ( k_2 ) of a second kind, \ldots, ( k_j ) of a ( j )th kind, and where ( k_1 + k_2 + \cdots + k_j = n )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>none of the ( n ) objects remains in its original place (derangements)</td>
<td>( D_n = n!(1 - \frac{1}{1!} + \cdots + (-1)^n \frac{1}{n!}) )</td>
<td>§2.4.2</td>
</tr>
</tbody>
</table>

| Arranging objects in a circle (where rotations, but not reflections, are equivalent): | | |
| \( n \) distinct objects | \( (n-1)! \) | §2.2.1 |
| \( k \) out of \( n \) distinct objects | \( \frac{P(n,k)}{k} \) | §2.2.1 |

| Choosing \( k \) objects from \( n \) distinct objects: | | |
| order matters, no repetitions | \( P(n,k) = \binom{n}{k} = \binom{n}{(n-k)!} = n^k \) | §2.3.1 |
| order matters, repetitions allowed | \( P^R(n,k) = n^k \) | §2.3.3 |
| order does not matter, no repetitions | \( C(n,k) = \binom{n}{k} = \binom{n}{k(n-k)!} \) | §2.3.2 |
| order does not matter, repetitions allowed | \( C^R(n,k) = \binom{k+n-1}{k} \) | §2.3.3 |

Figure 5: Counting problems and corresponding sections in [18].
<table>
<thead>
<tr>
<th>objects</th>
<th>number of objects</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subsets:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of size $k$ from a set of size $n$</td>
<td>$\binom{n}{k}$</td>
<td>§2.3.2</td>
</tr>
<tr>
<td>of all sizes from a set of size $n$</td>
<td>$2^n$</td>
<td>§2.3.4</td>
</tr>
<tr>
<td>of ${1,\ldots,n}$, without consecutive elements</td>
<td>$F_{n+2}$</td>
<td>§3.1.2</td>
</tr>
</tbody>
</table>

| Placing $n$ objects into $k$ cells: | | |
| distinct objects into distinct cells | $k^n$ | §2.2.1 |
| distinct objects into distinct cells, no cell empty | $\{ n \choose k \} k!$ | §2.5.2 |
| distinct objects into identical cells | $\{ n \choose 1 \} + \{ n \choose 2 \} + \cdots + \{ n \choose k \} = B_n$ | §2.5.2 |
| distinct objects into identical cells, no cell empty | $\{ n \choose 1 \}$ | §2.5.2 |
| distinct objects into distinct cells, with $k_i$ in cell $i$ ($i = 1,\ldots,n$), and where $k_1 + k_2 + \cdots + k_j = n$ | $\binom{n}{k_1 k_2 \ldots k_j}$ | §2.3.2 |
| identical objects into distinct cells | $\binom{n+k-1}{n}$ | §2.3.3 |
| identical objects into distinct cells, no cell empty | $\binom{n-1}{k-1}$ | §2.3.3 |
| identical objects into identical cells | $p_k(n)$ | §2.5.1 |
| identical objects into identical cells, no cell empty | $p_k(n) - p_{k-1}(n)$ | §2.5.1 |
| Placing $n$ distinct objects into $k$ nonempty cycles | $\binom{n}{k}$ | §2.5.2 |

Solutions to $x_1 + x_2 + \cdots + x_n = k$: |
| nonnegative integers | $\binom{k+n-1}{k}$ = $\binom{k+n-1}{n-1}$ | §2.3.3 |
| positive integers | $\binom{k-1}{n-1}$ | §2.3.3 |
| integers where $0 \leq a_i \leq x_i$ for all $i$ | $\binom{k-(a_1+\cdots+a_n)+n-1}{n-1}$ | §2.3.3 |
| integers where $0 \leq x_i \leq a_i$ for one or more $i$ | inclusion/exclusion principle | §2.4.2 |

Figure 6: Counting problems (con’t) and corresponding sections in [18].
Functions from a $k$-element set to an $n$-element set:

- All functions: $n^k$
- One-to-one functions ($n \geq k$): $n^k = \frac{n!}{(n-k)!} = P(n, k)$
- Onto functions ($n \leq k$): inclusion/exclusion
- Partial functions: $\binom{k}{0} + \binom{k}{1} n + \binom{k}{2} n^2 + \cdots + \binom{k}{n} n^k = (n + 1)^k$

Bit strings of length $n$:

- All strings: $2^n$
- With given entries in $k$ positions: $2^{n-k}$
- With exactly $k$ 0s: $\binom{n}{k}$
- With at least $k$ 0s: $\binom{n}{k} + \binom{n}{k+1} + \cdots + \binom{n}{n/2}$
- With equal numbers of 0s and 1s: $\binom{n/2}{n/2}$
- Palindromes: $2^\lceil n/2 \rceil$
- With an even number of 0s: $2^{n-1}$
- Without consecutive 0s: $F_{n+2}$

**Figure 7:** Counting problems (con’t) and corresponding sections in [18].

- $B(n)$ or $B_n$: Bell number
- $b(n,k)$: associated Stirling number of the second kind
- $C_n = \frac{1}{n+1} \binom{2n}{n}$: Catalan number
- $C(n,k) = \binom{n}{k} = \frac{n!}{k!(n-k)!}$: binomial coefficient
- $d(n,k)$: associated Stirling number of the first kind
- $E_n$: Euler number
- $\phi$: Euler phi-function
- $E(n,k)$: Eulerian number
- $F_n$: Fibonacci number

**Figure 8:** Notation from [18]
1.3 Dirac Delta Function

The (Dirac) delta function or (unit) impulse function is denoted by \( \delta(t) \). It is usually depicted as a vertical arrow at the origin. Note that \( \delta(t) \) is not a true function; it is undefined at \( t = 0 \). We define \( \delta(t) \) as a generalized function which satisfies the sampling property (or sifting property)

\[
\int \phi(t) \delta(t) dt = \phi(0)
\]

for any function \( \phi(t) \) which is continuous at \( t = 0 \). From this definition, it follows that

\[
(\delta * \phi)(t) = (\phi * \delta)(t) = \int \phi(\tau) \delta(t - \tau) d\tau = \phi(t)
\]

where we assume that \( \phi \) is continuous at \( t \). Intuitively, we may visualize \( \delta(t) \) as a infinitely tall, infinitely narrow rectangular pulse of unit area: \( \lim_{\varepsilon \to 0} 1/\varepsilon [ |t| \leq \varepsilon/2 ] \).

We list some interesting properties of \( \delta(t) \) here.

- \( \delta(t) = 0 \) when \( t \neq 0 \).
- \( \delta(t - T) = 0 \) for \( t \neq T \).
- \( \int_A \delta(t) dt = 1_A(0) \).
  - (a) \( \int \delta(t) dt = 1 \).
  - (b) \( \int_{\{0\}} \delta(t) dt = 1 \).
  - (c) \( \int_{-\infty}^{t} \delta(t) dt = 1_{[0, \infty)}(x) \). Hence, we may think of \( \delta(t) \) as the “derivative” of the unit step function \( U(t) = 1_{[0, \infty)}(x) \).
- \( \int \phi(t) \delta(t) dt = \phi(0) \) for \( \phi \) continuous at 0.
- \( \int \phi(t) \delta(t - T) dt = \phi(T) \) for \( \phi \) continuous at \( T \). In fact, for any \( \varepsilon > 0 \),

\[
\int_{T-\varepsilon}^{T+\varepsilon} \phi(t) \delta(t - T) dt = \phi(T).
\]

- \( \delta(at) = \frac{1}{|a|} \delta(t) \)
- \( \delta(t - t_1) \ast \delta(t - t_2) = \delta(t - (t_1 + t_2)) \).
- \( g(t) \ast \delta(t - t_0) = g(t - t_0) \).
- Fourier properties:
  - Fourier series: \( \delta(x - a) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{\infty} \cos(n(x - a)) \) on \( [-\pi, \pi] \).
  - Fourier transform: \( \delta(t) = \int 1 e^{j2\pi ft} df \)
• For a function $g$ whose real-values roots are $t_i$,

$$\delta(g(t)) = \sum_{k=1}^{n} \frac{\delta(t - t_i)}{|g'(t_i)|}$$  \hspace{1cm} (3)

Hence,

$$\int f(t)\delta(g(t))dt = \sum_{x: g(x) = 0} \frac{f(x)}{|g'(x)|}$$  \hspace{1cm} (4)

Note that the (Dirac) delta function is to be distinguished from the discrete time Kronecker delta function.

As a finite measure, $\delta$ is a unit mass at 0; that is for any set $A$, we have $\delta(A) = 1[0 \in A]$. In which case, we have again $\int g d\delta = \int f(x)\delta(dx) = g(0)$ for any measurable $g$.

For a function $g : D \to \mathbb{R}^n$ where $D \subset \mathbb{R}^n$,

$$\delta(g(x)) = \sum_{z: g(z) = 0} \frac{\delta(x - z)}{|\det dg(z)|}$$  \hspace{1cm} (5)

$\text{p 387}$.

2 Classical Probability

Classical probability, which is based upon the ratio of the number of outcomes favorable to the occurrence of the event of interest to the total number of possible outcomes, provided most of the probability models used prior to the 20th century. Classical probability remains of importance today and provides the most accessible introduction to the more general theory of probability.

Given a finite sample space $\Omega$, the classical probability of an event $A$ is

$$P(A) = \frac{\|A\|}{\|\Omega\|} = \frac{\text{the number of cases favorable to the outcome of the event}}{\text{the total number of possible cases}}.$$

• In this section, we are more apt to refer to equipossible cases as ones selected at random. Probabilities can be evaluated for events whose elements are chosen at random by enumerating the number of elements in the event.

• The bases for identifying equipossibility were often

  • physical symmetry (e.g. a well-balanced die, made of homogeneous material in a cubical shape) or

  • a balance of information or knowledge concerning the various possible outcomes.

• Equipossibility is meaningful only for finite sample space, and, in this case, the evaluation of probability is accomplished through the definition of classical probability.

2.1. Basic properties of classical probability:
• \(P(A) \geq 0\)
• \(P(\Omega) = 1\)
• \(P(\emptyset) = 0\)
• \(P(A^c) = 1 - P(A)\)
• \(P(A \cup B) = P(A) + P(B) - P(A \cap B)\) which comes directly from
  \[|A \cup B| = |A| + |B| - |A \cap B|\].

• \(A \perp B\) is equivalent to \(P(A \cap B) = 0\).
• \(A \perp B \Rightarrow P(A \cup B) = P(A) + P(B)\)
• Suppose \(\Omega = \{\omega_1, \ldots, \omega_n\}\) and \(P(\omega_i) = \frac{1}{n}\). Then \(P(A) = \sum_{\omega \in A} P(\omega)\).
  
  ○ The probability of an event is equal to the sum of the probabilities of its component outcomes because outcomes are mutually exclusive

\[2.2. \textbf{Classical Conditional Probability:}\] The conditional classical probability \(P(A|B)\) of event \(A\), given that event \(B \neq \emptyset\) occurred, is given by

\[
P(A|B) = \frac{|A \cap B|}{|B|} = \frac{P(A \cap B)}{P(B)}. \tag{6}
\]

• It is the updated probability of the event \(A\) given that we now know that \(B\) occurred.
• Read “conditional probability of \(A\) given \(B\)’”.
• \(P(A|B) = P(A \cap B|B) \geq 0\)
• For any \(A\) such that \(B \subset A\), we have \(P(A|B) = 1\). This implies
  \[P(\Omega|B) = P(B|B) = 1.\]

• If \(A \perp C\), \(P(A \cup C|B) = P(A|B) + P(C|B)\)
• \(P(A \cap B) = P(B)P(A|B)\)
• \(P(A \cap B) \leq P(A|B)\)
• \(P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)\)
• \(P(A \cap B) = P(A) \times P(B|A)\)
• \(P(A \cap B \cap C) = P(A \cap B) \times P(C|A \cap B)\)
• \(P(A, B|C) = P(A|C) P(B|A, C) = P(B|C) P(A|B, C)\)
2.3. **Total Probability and Bayes Theorem** If \( \{B_i, \ldots, B_n\} \) is a partition of \( \Omega \), then for any set \( A \),

- **Total Probability Theorem**: \( P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \).

- **Bayes Theorem**: Suppose \( P(A) > 0 \), we have \( P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)} \).

2.4. **Independence Events**: \( A \) and \( B \) are independent (\( A \perp B \)) if and only if
\[
P(A \cap B) = P(A)P(B) \tag{7}
\]

In classical probability, this is equivalent to
\[
|A \cap B|/|\Omega| = |A||B|.
\]

- Sometimes the definition for independence above does not agree with the everyday-language use of the word “independence”. Hence, many authors use the term “statistically independence” for the definition above to distinguish it from other definitions.

2.5. Having three pairwise independent events does **not** imply that the three events are jointly independent. In other words,
\[
A \perp B, B \perp C, A \perp C \not\implies A \perp B \perp C.
\]

Example: Experiment of flipping a fair coin twice. \( \Omega = \{HH, HT, TH, TT\} \). Define event \( A \) to be the event that the first flip gives a H; that is \( A = \{HH, HT\} \). Event \( B \) is the event that the second flip gives a H; that is \( B = \{HH, TH\} \). \( C = \{HH, TT\} \). Note also that even though the events \( A \) and \( B \) are not disjoint, they are independent.

2.6. Consider \( \Omega \) of size \( 2^n \). We are given a set \( A \subset \Omega \) of size \( n \). Then, \( P(A) = \frac{1}{2} \). We want to find all sets \( B \subset \Omega \) such that \( A \perp B \). (Note that without the required independence, there are \( 2^{2n} \) possible \( B \).) For independence, we need
\[
P(A \cap B) = P(A)P(B).
\]

Let \( r = |A \cap B| \). Then, \( r \) can be any integer from 0 to \( n \). Also, let \( k = |B \setminus A| \). Then, the condition for independence becomes
\[
\frac{r}{n} = \frac{1}{2} \frac{r+k}{n}
\]
which is equivalent to \( r = k \). So, the construction of the set \( B \) is given by choosing \( r \) elements from set \( A \), then choose \( r = k \) elements from set \( \Omega \setminus A \). There are \( \binom{n}{r} \) choices for the first part and \( \binom{n}{k} = \binom{n}{r} \) choice for the second part. Therefore, the total number of possible \( B \) such that \( A \perp B \) is
\[
\sum_{r=1}^{n} \binom{n}{r}^2 = \binom{2n}{n}.
\]
2.1 Examples

2.7. Background

(a) Historically, dice is the plural of die, but in modern standard English dice is used as both the singular and the plural. [Excerpted from Compact Oxford English Dictionary.]

2.8. Chevalier de Mere’s Scandal of Arithmetic:

Which is more likely, obtaining at least one six in 4 tosses of a fair die (event $A$), or obtaining at least one double six in 24 tosses of a pair of dice (event $B$).

We have

$$P(A) = 1 - \left(\frac{5}{6}\right)^4 = 0.518$$

and

$$P(B) = 1 - \left(\frac{35}{36}\right)^{24} = 0.491.$$ 

Therefore, the first case is more probable.

Remark: Probability theory was originally inspired by gambling problems. In 1654, Chevalier de Mere invented a gambling system which bet even money on the second case above. However, when the he began losing money, he asked his mathematician friend Blaise Pascal to analyze his gambling system. Pascal discovered that the Chevalier’s system would lose about 51 percent of the time. Pascal became so interested in probability and together with another famous mathematician, Pierre de Fermat, they laid the foundation of probability theory.

2.9. A random sample of size $r$ with replacement is taken from a population of $n$ elements. The probability of the event that in the sample no element appears twice (that is, no repetition in our sample) is

$$\frac{(n)_r}{n^r}.$$ 

The probability that at least one element appears twice is

$$p_u (n, r) = 1 - \prod_{i=1}^{r-1} \left(1 - \frac{i}{n}\right) \approx 1 - e^{-\frac{r(r-1)}{2n}}.$$ 

In fact, when $r - 1 < \frac{n}{2}$, (A.3) gives

$$e^{\frac{1}{2} \cdot \frac{r(r-1)}{n} \frac{3n+2r-1}{3n}} \leq \prod_{i=1}^{r-1} \left(1 - \frac{i}{n}\right) \leq e^{\frac{1}{2} \cdot \frac{r(r-1)}{n}}.$$ 

• From the approximation, to have $p_u (n, r) = p$, we need

$$r \approx \frac{1}{2} + \frac{1}{2} \sqrt{1 - 8n \ln(1-p)}.$$ 

23
• Probability of coincidence birthday: Probability that there is at least two people who have the same birthday in your class of \( n \) students

\[
  = \begin{cases} 
    1, & \text{if } r \geq 365, \\
    1 - \left( \frac{365}{365} \cdot \frac{364}{365} \cdot \ldots \cdot \frac{365 - (r - 1)}{365} \right) , & \text{if } 0 \leq r \leq 365
  \end{cases}
\]

○ Birthday Paradox: In a group of 23 randomly selected people, the probability that at least two will share a birthday (assuming birthdays are equally likely to occur on any given day of the year) is about 0.5. See also (3).

![Figure 9: \( p_u(n, r) \)](image)

2.10. Monte Hall’s Game: Started with showing a contestant 3 closed doors behind of which was a prize. The contestant selected a door but before the door was opened, Monte Hall, who knew which door hid the prize, opened a remaining door. The contestant was then allowed to either stay with his original guess or change to the other closed door. Question: better to stay or to switch? Answer: Switch. Because after given that the contestant switched, then the probability that he won the prize is \( \frac{2}{3} \).

2.11. False Positives on Diagnostic Tests: Let \( D \) be the event that the testee has the disease. Let + be the event that the test returns a positive result. Denote the probability of having the disease by \( p_D \).

Now, assume that the test always returns positive result. This is equivalent to \( P(+|D) = 1 \) and \( P(+|\neg D) = 0 \). Also, suppose that even when the testee does not have a disease, the test will still return a positive result with probability \( p_+ \); that is \( P(+|\neg D) \).

If the test returns positive result, then the probability that the testee has the disease is

\[
P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{p_D}{p_D + p_+ (1 - p_D)} = \frac{1}{1 + \frac{p_D}{p_D} (1 - p_D)}
\]

\[
\approx \frac{p_D}{p_+}; \text{ for rare disease } (p_D \ll 1)
\]
3 Probability Foundations

To study formal definition of probability, we start with the probability space \((\Omega, \mathcal{A}, P)\). Let \(\Omega\) be an arbitrary space or set of points \(\omega\). Viewed probabilistically, a subset of \(\Omega\) is an event and an element \(\omega\) of \(\Omega\) is a sample point. Each event is a collection of outcomes which are elements of the sample space \(\Omega\).

The theory of probability focuses on collections of events, called event algebras and typically denoted \(\mathcal{A}\) (or \(\mathcal{F}\)) that contain all the events of interest (regarding the random experiment \(\mathcal{E}\)) to us, and are such that we have knowledge of their likelihood of occurrence. The probability \(P\) itself is defined as a number in the range \([0, 1]\) associated with each event in \(\mathcal{A}\).

3.1 Algebra and \(\sigma\)-algebra

The class \(2^\Omega\) of all subsets can be too large for us to define probability measures with consistency, across all member of the class. In this section, we present smaller classes which have several “nice” properties.

**Definition 3.1.** [7, Def 1.6.1 p38] An event algebra \(\mathcal{A}\) is a collection of subsets of the sample space \(\Omega\) such that it is

1. nonempty (this is equivalent to \(\Omega \in \mathcal{A}\));
2. closed under complementation (if \(A \in \mathcal{A}\) then \(A^c \in \mathcal{A}\));
3. and closed under finite unions (if \(A, B \in \mathcal{A}\) then \(A \cup B \in \mathcal{A}\)).

In other words, “A class is called an algebra on \(\Omega\) if it contains \(\Omega\) itself and is closed under the formation of complements and finite unions.”

3.2. Examples of algebras

- \(\Omega = \) any fixed interval of \(\mathbb{R}\). \(\mathcal{A} = \) \{finite unions of intervals contained in \(\Omega\}\)
- \(\Omega = (0, 1]\). \(\mathcal{B}_0 = \) the collection of all finite unions of intervals of the form \((a, b] \subset (0, 1]\)

\[^1\text{There is no problem when } \Omega \text{ is countable.}\]
Not a σ-field. Consider the set \( \bigcup_{i=1}^{\infty} \left( \frac{1}{2^{i+1}}, \frac{1}{2^{i}} \right) \)

### 3.3. Properties of an algebra \( \mathcal{A} \):

(a) Nonempty: \( \emptyset \in \mathcal{A}, X \in \mathcal{A} \)

(b) \( \mathcal{A} \subset 2^{\Omega} \)

(c) An algebra is closed under finite set-theoretic operations.

- \( A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A} \)
- \( A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}, A \cap B \in \mathcal{A}, A \setminus B \in \mathcal{A}, A \Delta B = (A \setminus B) \cup (B \setminus A) \in \mathcal{A} \)
- \( A_1, A_2, \ldots, A_n \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{F} \) and \( \bigcap_{i=1}^{n} A_i \in \mathcal{F} \)

(d) The collection of algebras in \( \Omega \) is closed under arbitrary intersection. In particular, let \( \mathcal{A}_1, \mathcal{A}_2 \) be algebras of \( \Omega \) and let \( \mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \) be the collection of sets common to both algebras. Then \( \mathcal{A} \) is an algebra.

(e) The smallest \( \mathcal{A} \) is \( \{\emptyset, \Omega\} \).

(f) The largest \( \mathcal{A} \) is the set of all subsets of \( \Omega \) known as the power set and denoted by \( 2^{\Omega} \).

(g) Cardinality of Algebras: An algebra of subsets of a finite set of \( n \) elements will always have a cardinality of the form \( 2^k \), \( k \leq n \). It is the intersection of all algebras which contain \( \mathcal{C} \).

### 3.4. There is a smallest (in the sense of inclusion) algebra containing any given family \( \mathcal{C} \) of subsets of \( \Omega \). Let \( \mathcal{C} \subset 2^\mathcal{X} \), the **algebra generated by** \( \mathcal{C} \) is

\[
\bigcap_{\mathcal{G} \in \mathcal{G}} \mathcal{G},
\]

i.e., the intersection of all algebra containing \( \mathcal{C} \). It is the **smallest algebra** containing \( \mathcal{C} \).

**Definition 3.5.** A **σ-algebra** \( \mathcal{A} \) is an event algebra that is also closed under countable unions,

\[
(\forall i \in \mathbb{N}) A_i \in \mathcal{A} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{A}.
\]

Remarks:

- A σ-algebra is also an algebra.
- A finite algebra is also a σ-algebra.

### 3.6. Because every σ-algebra is also an algebra, it has all the properties listed in (3.3).

Extra properties of σ-algebra \( \mathcal{A} \) are as followed:
(a) \( A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{j=1}^{\infty} A_j \in \mathcal{A} \) and \( \bigcap_{j=1}^{\infty} A_j \in \mathcal{A} \)

(b) A \( \sigma \)-field is closed under countable set-theoretic operations.

(c) The collection of \( \sigma \)-algebra in \( \Omega \) is closed under arbitrary intersection, i.e., let \( \mathcal{A}_\alpha \) be \( \sigma \)-algebra \( \forall \alpha \in A \) where \( A \) is some index set, potentially uncountable. Then, \( \bigcap_{\alpha \in A} \mathcal{A}_\alpha \) is still a \( \sigma \)-algebra.

(d) An infinite \( \sigma \)-algebra \( \mathcal{F} \) on \( X \) is uncountable i.e. \( \sigma \)-algebra is either finite or uncountable.

(e) If \( \mathcal{A} \) and \( \mathcal{B} \) are \( \sigma \)-algebra in \( X \), then, it is not necessary that \( \mathcal{A} \cup \mathcal{B} \) is also a \( \sigma \)-algebra. For example, let \( E_1, E_2 \subset X \) distinct, and not complement of one another. Let \( \mathcal{A}_i = \{ \emptyset, E_i, E_i^c, X \} \). Then, \( E_i \in \mathcal{A}_1 \cup \mathcal{A}_2 \) but \( E_1 \cup E_2 \notin \mathcal{A}_1 \cup \mathcal{A}_2 \).

Definition 3.7. Generation of \( \sigma \)-algebra Let \( \mathcal{C} \subset 2^\Omega \), the \( \sigma \)-algebra generated by \( \mathcal{C} \), \( \sigma (\mathcal{C}) \) is

\[
\bigcap_{G \subset \mathcal{G}} G,
\]

i.e., the intersection of all \( \sigma \)-field containing \( \mathcal{C} \). It is the smallest \( \sigma \)-field containing \( \mathcal{C} \).

- If the set \( \Omega \) is not implicit, we will explicitly write \( \sigma_X (\mathcal{C}) \)
- We will say that a set \( A \) can be generated by elements of \( \mathcal{C} \) if \( A \in \sigma (\mathcal{C}) \)

3.8. Properties of \( \sigma (\mathcal{C}) \):

(a) \( \sigma (\mathcal{C}) \) is a \( \sigma \)-field

(b) \( \sigma (\mathcal{C}) \) is the smallest \( \sigma \)-field containing \( \mathcal{C} \) in the sense that if \( \mathcal{H} \) is a \( \sigma \)-field and \( \mathcal{C} \subset \mathcal{H} \), then \( \sigma (\mathcal{C}) \subset \mathcal{H} \)

(c) \( \mathcal{C} \subset \sigma (\mathcal{C}) \)

(d) \( \sigma (\sigma (\mathcal{C})) = \sigma (\mathcal{C}) \)

(e) If \( \mathcal{H} \) is a \( \sigma \)-field and \( \mathcal{C} \subset \mathcal{H} \), then \( \sigma (\mathcal{C}) \subset \mathcal{H} \)

(f) \( \sigma (\{ \emptyset \}) = \sigma (\{ X \}) = \{ \emptyset, X \} \)

(g) \( \sigma (\{ A \}) = \{ \emptyset, A, A^c, X \} \) for \( A \subset \Omega \).

(h) \( \sigma (\{ A, B \}) \) has at most 16 elements. They are \( \emptyset, A, B, A \cap B, A \cup B, A \setminus B, B \setminus A, A \Delta B \), and their corresponding complements. See also (3.11). Some of these 16 sets can be the same and hence the use of “at-most” in the statement.

(i) \( \sigma (A) = \sigma (A \cup \{ \emptyset \}) = \sigma (A \cup \{ X \}) = \sigma (A \cup \{ \emptyset, X \}) \)
(j) \( A \subset B \Rightarrow \sigma(A) \subset \sigma(B) \)

(k) \( \sigma(A), \sigma(B) \subset \sigma(A \cup B) \)

(l) \( \sigma(A) = \sigma(A \cup \{\emptyset\}) = \sigma(A \cup \{\Omega\}) = \sigma(A \cup \{\emptyset, \Omega\}) \)

3.9. For the decomposition described in (1.8), let the starting collection be \( \mathcal{C}_1 \) and the decomposed collection be \( \mathcal{C}_2 \).

- If \( \mathcal{C}_1 \) is finite, then \( \sigma(C_2) = \sigma(C_1) \).
- If \( \mathcal{C}_1 \) is countable, then \( \sigma(C_2) \subset \sigma(C_2) \).

3.10. **Construction of \( \sigma \)-algebra from countable partition**: An intermediate-sized \( \sigma \)-algebra (a \( \sigma \)-algebra smaller than \( 2^\Omega \)) can be constructed by first partitioning \( \Omega \) into subsets and then forming the power set of these subsets, with an individual subset now playing the role of an individual outcome \( \omega \).

Given a countable partition \( \Pi = \{A_i, i \in I\} \) of \( \Omega \), we can form a \( \sigma \)-algebra \( \mathcal{A} \) by including all unions of subcollections:

\[
\mathcal{A} = \{\bigcup_{\alpha \in S} A_\alpha : S \subset I\}
\]

[7] Ex 1.5 p. 39 where we define \( \bigcup_{i \in \emptyset} A_i = \emptyset \). It turns out that \( \mathcal{A} = \sigma(\Pi) \). Of course, a \( \sigma \)-algebra is also an algebra. Hence, (8) is also a way to construct an algebra. Note that, from (1.9), the necessary and sufficient condition for distinct \( S \) to produce distinct element in (8) is that none of the \( A_\alpha \)'s are empty.

In particular, for countable \( \Omega = \{x_i : i \in \mathbb{N}\} \), where \( x_i \)'s are distinct. If we want a \( \sigma \)-algebra which contains \( \{x_i\} \forall i \), then, the smallest one is \( 2^\Omega \) which happens to be the biggest one. So, \( 2^\Omega \) is the only \( \sigma \)-algebra which is “reasonable” to use.

3.11. **Generation of \( \sigma \)-algebra from finite partition**: If a finite collection \( \Pi = \{A_i, i \in I\} \) of non-empty sets forms a partition of \( \Omega \), then the algebra generated by \( \Pi \) is the same as the \( \sigma \)-algebra generated by \( \Pi \) and it is given by (8). Moreover, \( |\sigma(\Pi)| = 2^{|\Pi|} \), that is distinct sets \( S \) in (8) produce distinct member of the \( \sigma \)-algebra.

Therefore, given a finite collection of sets \( \mathcal{C} = \{C_1, C_2, \ldots, C_n\} \). To find an algebra or a \( \sigma \)-algebra generated by \( \mathcal{C} \), the first step is to use the \( C_i \)'s to create a partition of \( \Omega \). Using (1.8), the partition is given by

\[
\Pi = \{\cap_{i=1}^n B_i : B_i = C_i \text{ or } C_i^c\}.
\]

By (3.9), we know that \( \sigma(\pi) = \sigma(C) \). Note that there are seemingly \( 2^n \) sets in \( \Pi \) however, some of them can be \( \emptyset \). We can eliminate the empty set(s) from \( \Pi \) and it is still a partition. So the cardinality of \( \Pi \) in (9) (after empty-set elimination) is \( k \) where \( k \) is at most \( 2^n \). The partition \( \Pi \) is then a collection of \( k \) sets which can be renamed as \( A_1, \ldots, A_k \), all of which are non-empty. Applying the construction in (8) (with \( I = [n] \)), we then have \( \sigma(C) \) whose cardinality is \( 2^k \) which is \( \leq 2^{2^n} \). See also properties (3.8.7) and (3.8.8).

\[^2\text{In this case, } \Pi \text{ is countable if and only if } I \text{ is countable.}\]
Definition 3.12. In general, the \textbf{Borel $\sigma$-algebra} or Borel algebra $B$ is the $\sigma$-algebra generated by the open subsets of $\Omega$

- Call $B_\Omega$ the \textbf{\textit{$\sigma$-algebra of Borel subsets}} of $\Omega$ or $\sigma$-algebra of Borel sets on $\Omega$
- Call set $B \in B_\Omega$ \textbf{Borel set} of $\Omega$

(a) On $\Omega = \mathbb{R}$, the $\sigma$-algebra generated by any of the followings are \textbf{Borel $\sigma$-algebra}:

(i) Open sets
(ii) Closed sets
(iii) Intervals
(iv) Open intervals
(v) Closed intervals
(vi) Intervals of the form $(-\infty, a]$, where $a \in \mathbb{R}$
   - i. Can replace $\mathbb{R}$ by $\mathbb{Q}$
   - ii. Can replace $(-\infty, a]$ by $(-\infty, a)$, $[a, +\infty)$, or $(a, +\infty)$
   - iii. Can replace $(-\infty, a]$ by combination of those in (f)ii.

(b) For $\Omega \subset \mathbb{R}$, $B_\Omega = \{ A \in B_\mathbb{R} : A \subset \Omega \} = B_\mathbb{R} \cap \Omega$ where $B_\mathbb{R} \cap \Omega = \{ A \cap \Omega : A \in B_\mathbb{R} \}$

(c) Borel $\sigma$-algebra on the extended real line is the extended $\sigma$-algebra

$$B_\mathbb{R} = \{ A \cup B : A \in B_\mathbb{R}, B \in \{ \emptyset, \{-\infty\}, \{\infty\}, \{-\infty, \infty\} \} \}$$

It is generated by, for example

(i) $A \cup \{\{-\infty\}, \{\infty\}\}$ where $\sigma(A) = B_\mathbb{R}$
(ii) $\{[a, \hat{b}], \{\hat{a}, b\}\}, \{\hat{a}, b\}$
(iii) $\{[-\infty, b]\}, \{[-\infty, b]\}, \{\{a, +\infty\}\}$
(iv) $\{[-\infty, \hat{b}]\}, \{\{\hat{a}, +\infty\}\} \quad \{\{\hat{a}, +\infty\}\}$

Here, $a, b \in \mathbb{R}$ and $\hat{a}, \hat{b} \in \mathbb{R} \cup \{\pm\infty\}$

(d) Borel $\sigma$-algebra in $\mathbb{R}^k$ is generated by

(i) the class of open sets in $\mathbb{R}^k$
(ii) the class of closed sets in $\mathbb{R}^k$
(iii) the class of bounded semi-open rectangles (cells) $I$ of the form

$$I = \{ x \in \mathbb{R}^k : a_i < x_i \leq b_i, \; i = 1, \ldots, k \} .$$

Note that $I = \bigotimes_{i=1}^{k} (a_i, b_i)$ where $\otimes$ denotes the Cartesian product $\times$
(iv) the class of “southwest regions” $S_x$ of points “southwest” to $x \in \mathbb{R}^k$, i.e. $S_x = \{ y \in \mathbb{R}^k : y_i \leq x_i, \ i = 1, \ldots, k \}$

3.13. The Borel $\sigma$-algebra $\mathcal{B}$ of subsets of the reals is the usual algebra when we deal with real- or vector-valued quantities.

Our needs will not require us to delve into these issues beyond being assured that events we discuss are constructed out of intervals and repeated set operations on these intervals and these constructions will not lead us out of $\mathcal{B}$. In particular, countable unions of intervals, their complements, and much, much more are in $\mathcal{B}$.

3.2 Kolmogorov’s Axioms for Probability

Definition 3.14. Kolmogorov’s Axioms for Probability [13]: A set function satisfying K0–K4 is called a probability measure.

K0 Setup: The random experiment $\mathcal{E}$ is described by a probability space $(\Omega, \mathcal{A}, P)$ consisting of an event $\sigma$-algebra $\mathcal{A}$ and a real-valued function $P : \mathcal{A} \to \mathbb{R}$.

K1 Nonnegativity: $\forall A \in \mathcal{A}, P(A) \geq 0$.

K2 Unit normalization: $P(\Omega) = 1$.

K3 Finite additivity: If $A, B$ are disjoint, then $P(A \cup B) = P(A) + P(B)$.

K4 Monotone continuity: If $(\forall i > 1) A_{i+1} \subset A_i$ and $\cap_{i \in \mathbb{N}} A_i = \emptyset$ (a nested series of sets shrinking to the empty set), then

$$\lim_{i \to \infty} P(A_i) = 0.$$  

K4’ Countable or $\sigma$-additivity: If $(A_i)$ is a countable collection of pairwise disjoint (no overlap) events, then

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

- Note that there is never a problem with the convergence of the infinite sum; all partial sums of these non-negative summands are bounded above by 1.

- K4 is not a property of limits of sequences of relative frequencies nor meaningful in the finite sample space context of classical probability, is offered by Kolmogorov to ensure a degree of mathematical closure under limiting operations [7, p. 111].

- K4 is an idealization that is rejected by some accounts (usually subjectivist) of probability.

- If $P$ satisfies K0–K3, then it satisfies K4 if and only if it satisfies K4’ [7, Theorem 3.5.1 p. 111].
Proof. To show \(\Rightarrow\), consider disjoint \(A_1, A_2, \ldots\). Define \(B_n = \bigcup_{i>n} A_i\). Then, by (1.6), \(B_{n+1} \subset B_n\) and \(\bigcap_{n=1}^{\infty} B_n = \emptyset\). So, by K4, \(\lim_{n \to \infty} P(B_n) = 0\). Furthermore,

\[
\bigcup_{i=1}^{\infty} A_i = B_n \cup \left( \bigcup_{i=1}^{n} A_i \right)
\]

where all the sets on the RHS are disjoint. Hence, by finite additivity,

\[
P\left( \bigcup_{i=1}^{\infty} A_i \right) = P(B_n) + \sum_{i=1}^{n} P(A_i).
\]

Taking limiting as \(n \to \infty\) gives us K4'.

To show \(\Leftarrow\), see (3.17).

Equivalently, instead of K0-K4, we can define probability measure using P0-P2 below.

**Definition 3.15.** A **probability measure** defined on a \(\sigma\)-algebra \(\mathcal{A}\) of \(\Omega\) is a (set) function

(P0) \(P : \mathcal{A} \to [0, 1]\)

that satisfies:

(P1,K2) \(P(\Omega) = 1\)

(P2,K4') **Countable additivity:** For every countable sequence \((A_n)_{n=1}^{\infty}\) of disjoint elements of \(\mathcal{A}\), one has \(P\left( \bigcup_{n=1}^{\infty} A_n \right) = \sum_{n=1}^{\infty} P(A_n)\)

- \(P(\emptyset) = 0\)
- The number \(P(A)\) is called the **probability** of the event \(A\)
- The triple \((\Omega, \mathcal{A}, P)\) is called a **probability measure space**, or simply a **probability space**
- A **support** of \(P\) is any \(\mathcal{A}\)-set \(A\) for which \(P(A) = 1\)

### 3.3 Properties of Probability Measure

3.16. Properties of probability measures:

(a) \(P(\emptyset) = 0\)

(b) \(0 \leq P \leq 1\): For any \(A \in \mathcal{A}\), \(0 \leq P(A) \leq 1\)

(c) If \(P(A) = 1\), \(A\) is not necessary \(\Omega\).

(d) Additivity: \(A, B \in \mathcal{A}, A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)\)
(e) Monotonicity: \( A, B \in \mathcal{A}, A \subseteq B \Rightarrow P(A) \leq P(B) \) and \( P(B - A) = P(B) - P(A) \)

(f) \( P(A^c) = 1 - P(A) \)

(g) \( P(A) + P(B) = P(A \cup B) + P(A \cap B) = P(A - B) + 2P(A \cap B) + P(B - A) \).
\( P(A \cup B) = P(A) + P(B) - P(A \cap B) \).

(h) \( P(A \cup B) \geq \max(P(A), P(B)) \geq \min(P(A), P(B)) \geq P(A \cap B) \)

(i) **Inclusion-exclusion principle:**

\[
P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{i} P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1} P(A_1 \cap \ldots \cap A_n)
\]

In a more compact form,

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{\emptyset \neq I \subset \{1, \ldots, n\}} (-1)^{|I|+1} P\left(\bigcap_{i \in I} A_i\right).
\]

For example,

\[
P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).
\]

See also (8.2). Moreover, for any event \( B \), we have

\[
P\left(\bigcap_{k=1}^{n} A_k^c \cap B\right) = P(B) + \sum_{\emptyset \neq I \subset [n]} (-1)^{|I|} P\left(\bigcap_{i \in I} A_i \cap B\right).
\]

(j) **Finite additivity:** If \( A = \bigcup_{j=1}^{n} A_j \) with \( A_j \in \mathcal{A} \) disjoint, then \( P(A) = \sum_{j=1}^{n} P(A_j) \)

- If \( A \) and \( B \) are disjoint sets in \( \mathcal{A} \), then \( P(A \cup B) = P(A) + P(B) \)

(k) **Subadditivity or Boole’s Inequality:** If \( A_1, \ldots, A_n \) are events, not necessarily disjoint, then \( P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i) \)

(l) **\( \sigma \)-subadditivity:** If \( A_1, A_2, \ldots \) is a sequence of measurable sets, not necessarily disjoint, then \( P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i) \)

- This formula is known as the **union bound** in engineering.
- If \( A_1, A_2, \ldots \) is a sequence of events, not necessarily disjoint, then \( \forall \alpha \in (0, 1], \)
\[
P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \left(\sum_{i=1}^{\infty} P(A_i)\right)^{\alpha}.
\]

32
3.17. **Conditional continuity from above.** If \( B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots \) is a decreasing sequence of measurable sets, then \( P(\bigcap_{i=1}^{\infty} B_i) = \lim_{j \to \infty} P(B_j) \). In a more compact notation, if \( B_i \searrow B \), then \( P(B) = \lim_{j \to \infty} P(B_j) \).

*Proof.* Let \( B = \bigcap_{i=1}^{\infty} B_i \). Let \( A_k = B_k \setminus B_{k+1} \), i.e., the new part. We consider two partitions of \( B_1 \): (1) \( B_1 = B \cup \bigcup_{j=1}^{\infty} A_j \) and (2) \( B_1 = B_n \cup \bigcup_{j=1}^{n-1} A_j \). (1) implies \( P(B_1) - P(B) = \sum_{j=1}^{\infty} P(A_j) \).

(2) implies \( P(B_1) - P(B_n) = \sum_{j=1}^{n-1} P(A_j) \). We then have

\[
\lim_{n \to \infty} (P(B_1) - P(B_n)) = \sum_{j=1}^{\infty} P(A_j) = P(B_1) - P(B).
\]

\( \square \)

3.18. Let \( \mathcal{A} \) be a \( \sigma \)-algebra. Suppose that \( P : \mathcal{A} \to [0,1] \) satisfies (P1) and is finitely additive. Then, the following are equivalent:

- (P2): If \( A_n \in \mathcal{A} \), disjoint, then \( P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \)
- (K4) If \( A_n \in \mathcal{A} \), and \( A_n \searrow \emptyset \), then \( P(A_n) \searrow 0 \)
- (Continuity from above) If \( A_n \in \mathcal{A} \), and \( A_n \searrow A \), then \( P(A_n) \searrow P(A) \)
- If \( A_n \in \mathcal{A} \), and \( A_n \nearrow \Omega \), then \( P(A_n) \nearrow P(\Omega) = 1 \)
- (Continuity from below) If \( A_n \in \mathcal{A} \), and \( A_n \nearrow A \), then \( P(A_n) \nearrow P(A) \)

Hence, a probability measure satisfies all five properties above. Continuity from above and continuity from below are collectively called **sequential continuity properties**.

In fact, for probability measure \( P \), let \( A_n \) be a sequence of events in \( \mathcal{A} \) which converges to \( A \) (i.e. \( A_n \to A \)). Then \( A \in \mathcal{A} \) and \( \lim_{n \to \infty} P(A_n) = P(A) \). Of course, both \( A_n \searrow A \) and \( A_n \nearrow A \) imply \( A_n \to A \). Note also that

\[
P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{N \to \infty} P\left(\bigcup_{n=1}^{N} A_n\right),
\]

and

\[
P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{N \to \infty} P\left(\bigcap_{n=1}^{N} A_n\right).
\]

Alternative form of the sequential continuity properties are

\[
P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{N \to \infty} P(A_N), \text{ if } A_n \subset A_{n+1}
\]
and
\[
P \left( \bigcap_{n=1}^{\infty} A_n \right) = \lim_{N \to \infty} P(A_N), \quad \text{if } A_{n+1} \subset A_n.
\]

3.19. Given a common event algebra \( \mathcal{A} \), probability measures \( P_1, \ldots, P_m \), and the numbers \( \lambda_1, \ldots, \lambda_m, \lambda_i \geq 0, \sum_1^m \lambda_i = 1 \), a convex combination \( P = \sum \lambda_i P_i \) of probability measures is a probability measure.

3.20. \( \mathcal{A} \) can not contain an uncountable, disjoint collection of sets of positive probability.

Definition 3.21. Discrete probability measure \( P \) is a discrete probability measure if \( \exists \) finitely or countably many points \( \omega_k \) and nonnegative masses \( m_k \) such that \( \forall A \in \mathcal{A} \)
\[
P(A) = \sum_{k: \omega_k \in A} m_k = \sum_k m_k I_A(\omega_k)
\]
If there is just one of these points, say \( \omega_0 \), with mass \( m_0 = 1 \), then \( P \) is a unit mass at \( \omega_0 \). In this case, \( \forall A \in \mathcal{A}, P(A) = I_A(\omega_0) \).

Notation: \( P = \delta_{\omega_0} \)
- Here, \( \Omega \) can be uncountable.

3.4 Countable \( \Omega \)

A sample space \( \Omega \) is countable if it is either finite or countably infinite. It is countably infinite if it has as many elements as there are integers. In either case, the element of \( \Omega \) can be enumerated as, say, \( \omega_1, \omega_2, \ldots \). If the event algebra \( \mathcal{A} \) contains each singleton set \( \{\omega_k\} \) (from which it follows that \( \mathcal{A} \) is the power set of \( \Omega \)), then we specify probabilities satisfying the Kolmogorov axioms through a restriction to the set \( S = \{\{\omega_k\}\} \) of singleton events.

Definition 3.22. When \( \Omega \) is countable, a probability mass function (pmf) is any function \( p : \Omega \rightarrow [0, 1] \) such that
\[
\sum_{\omega \in \Omega} p(\omega) = 1.
\]
When the elements of \( \Omega \) are enumerated, then it is common to abbreviate \( p(\omega_i) = p_i \).

3.23. Every pmf \( p \) defines a probability measure \( P \) and conversely. Their relationship is given by
\[
p(\omega) = P(\{\omega\}), \quad (11)
P(A) = \sum_{\omega \in A} p(\omega). \quad (12)
\]
The convenience of a specification by pmf becomes clear when \( \Omega \) is a finite set of, say, \( n \) elements. Specifying \( P \) requires specifying \( 2^n \) values, one for each event in \( \mathcal{A} \), and doing so in a manner that is consistent with the Kolmogorov axioms. However, specifying \( p \) requires only providing \( n \) values, one for each element of \( \Omega \), satisfying the simple constraints of nonnegativity and addition to 1. The probability measure \( P \) satisfying (12) automatically satisfies the Kolmogorov axioms.

34
## 3.5 Independence

**Definition 3.24.** Independence between events and collections of events.

(a) Two events $A, B$ are called **independent** if $P(A \cap B) = P(A)P(B)$

(i) An event with probability 0 or 1 is independent of any event (including itself).
   In particular, $\emptyset$ and $\Omega$ are independent of any events.

(ii) Two events $A, B$ with positive probabilities are independent if and only if 
   $$P(B | A) = P(B),$$
   which is equivalent to $P(A | B) = P(A)$
   When $A$ and/or $B$ has zero probability, $A$ and $B$ are automatically independent.

(iii) An event $A$ is independent of itself if and only if $P(A)$ is 0 or 1.

(iv) If $A$ an $B$ are independent, then the two classes
   $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$ and
   $\sigma(\{B\}) = \{\emptyset, B, B^c, \Omega\}$ are independent.

(v) Suppose $A$ and $B$ are disjoint. $A$ and $B$ are independent if and only if $P(A) = 0$
   or $P(B) = 0$.

(vi) Suppose $A \subset B$. $A$ and $B$ are independent if and only if
   $$P(A) = \frac{P(B)}{P(B)+1}.$$

(b) Independence for finite collection $\{A_1, \ldots, A_n\}$ of sets:

\[
\therefore P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j) \quad \forall J \subset [n] \text{ and } |J| \geq 2
\]

\[\therefore\]

- Note that the case when $j = 1$ automatically holds. The case when $j = 0$
  can be regard as the $\emptyset$ event case, which is also trivially true.

- There are $\sum_{j=2}^{n} \binom{n}{j} = 2^n - 1 - n$ constraints.

- Example: $A_1, A_2, A_3$ are independent if and only if
  \[
  P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)
  \]
  \[
  P(A_1 \cap A_2) = P(A_1)P(A_2)
  \]
  \[
  P(A_1 \cap A_3) = P(A_1)P(A_3)
  \]
  \[
  P(A_2 \cap A_3) = P(A_2)P(A_3)
  \]

**Remark:** The first equality alone is not enough for independence. See a
counter example below. In fact, it is possible for the first equation to hold
while the last three fail as shown in (3.26b). It is also possible to construct
events such that the last three equations hold (pairwise independence), but
the first one does not as demonstrated in (3.26a).

\[
\equiv P(B_1 \cap B_2 \cap \cdots \cap B_n) = P(B_1)P(B_2)\cdots P(B_n) \text{ where } B_i = A_i \text{ or } B_i = \Omega
\]

(c) Independence for collection $\{A_\alpha : \alpha \in I\}$ of sets:

\[
\equiv \forall \text{ finite } J \subset I, \ P\left(\bigcap_{\alpha \in J} A_\alpha\right) = \prod_{\alpha \in J} P(A)
\]

35
Example 3.26.

(d) Independence for finite collection \( \{A_1, \ldots, A_n\} \) of classes:

\[\begin{align*}
\equiv & \quad \text{Each of the finite subcollection is independent.} \\
\equiv & \quad \text{the finite collection of sets } A_1, \ldots, A_n \text{ is independent where } A_i \in A_i. \\
\equiv & \quad P (B_1 \cap B_2 \cdots \cap B_n) = P (B_1) P (B_2) \cdots P (B_n) \text{ where } B_i \in A_i \text{ or } B_i = \Omega \\
\equiv & \quad P (B_1 \cap B_2 \cdots \cap B_n) = P (B_1) P (B_2) \cdots P (B_n) \text{ where } B_i \in A_i \cup \{\Omega\} \\
\equiv & \quad \forall i \forall B_i \subset A_i \quad B_1, \ldots, B_n \text{ are independent.} \\
\equiv & \quad A_i \cup \{\Omega\}, \ldots, A_n \cup \{\Omega\} \text{ are independent.} \\
\equiv & \quad \{\emptyset, \Omega\} \text{ are independent.}
\end{align*}\]

(e) Independence for collection \( \{A_\theta : \theta \in \Theta\} \) of classes:

\[\begin{align*}
\equiv & \quad \text{Any collection } \{A_\theta : \theta \in \Theta\} \text{ of sets is independent where } A_\theta \in A_\theta \\
\equiv & \quad \text{Any finite subcollection of classes is independent.} \\
\equiv & \quad \forall \text{ finite } \Lambda \subset \Theta, \quad P \left( \bigcap_{\theta \in \Lambda} A_\theta \right) = \prod_{\theta \in \Lambda} P (A_\theta) \\
\equiv & \quad \text{By definition, a subcollection of independent events is also independent.} \\
\equiv & \quad \text{The class } \{\emptyset, \Omega\} \text{ is independent from any class.}
\end{align*}\]

Definition 3.25. A collection of events \( \{A_\alpha\} \) is called pairwise independent if for every distinct events \( A_{\alpha_1}, A_{\alpha_2} \), we have \( P (A_{\alpha_1} \cap A_{\alpha_2}) = P (A_{\alpha_1}) P (A_{\alpha_2}) \)

- If a collection of events \( \{A_\alpha : \alpha \in I\} \) is independent, then it is pairwise independent. The converse is false. See (a) in example 3.26.

- For \( K \subset J \), \( P \left( \bigcap_{\alpha \in J} A_\alpha \right) = \prod_{\alpha \in J} P (A) \) does not imply \( P \left( \bigcap_{\alpha \in K} A_\alpha \right) = \prod_{\alpha \in K} P (A) \)

Example 3.26.

(a) Let \( \Omega = \{1, 2, 3, 4\}, A = 2^\Omega, P (i) = \frac{1}{3}, A_1 = \{1, 2\}, A_2 = \{1, 3\}, A_3 = \{2, 3\}. \) Then \( P (A_i \cap A_j) = P (A_i) P (A_j) \) for all \( i \neq j \) but \( P (A_1 \cap A_2 \cap A_3) \neq P (A_1) P (A_2) P (A_3) \)

(b) Let \( \Omega = \{1, 2, 3, 4, 5, 6\}, A = 2^\Omega, P (i) = \frac{1}{6}, A_1 = \{1, 2, 3\}, A_2 = A_3 = \{4, 5, 6\}. \) Then, \( P (A_1 \cap A_2 \cap A_3) = P (A_1) P (A_2) P (A_3) \) but \( P (A_i \cap A_j) \neq P (A_i) P (A_j) \) for all \( i \neq j \)

(c) The paradox of ”almost sure” events: Consider two random events with probabilities of 99% and 99.99%, respectively. One could say that the two probabilities are nearly the same, both events are almost sure to occur. Nevertheless the difference may become significant in certain cases. Consider, for instance, independent events which may occur on any day of the year with probability \( p = 99\% \); then the probability \( P \) that it will occur every day of the year is less than 3%, while if \( p = 99.99\% \) then \( P = 97\% \).
4 Random Element

4.1. A function \( X : \Omega \rightarrow E \) is said to be a \textit{random element} of \( E \) if and only if \( X \) is \textit{measurable} which is equivalent to each of the following statements.

\[ \equiv X^{-1}(B_E) \subset A \]
\[ \equiv \sigma(X) \subset A \]
\[ \equiv (\text{reduced form}) \ \exists C \subset B_E \text{ such that } \sigma(C) = B_E \text{ and } X^{-1}(C) \subset A \]

- When \( E \subset \mathbb{R} \), \( X \) is called a \textit{random variable}.
- When \( E \subset \mathbb{R}^d \), then \( X \) is called a \textit{random vector}.

Definition 4.2. \( X = Y \) \textit{almost surely} (a.s.) if \( P[X = Y] = 1 \).

- The a.s. equality is an equivalence relation.

4.3. \textbf{Law} of \( X \) or \textbf{Distribution} of \( X \): \( P^X = \mu_X = PX^{-1} = \mathcal{L}(X) : E \rightarrow [0,1] \)

\[ \mu_X(A) = P^X(A) = P(X^{-1}(A)) = P(\{\omega : X(\omega) \in A\}) = P(\{X \in A\}) \]

4.4. For \( X \in L^p \), \( \lim_{t \to \infty} t^p P[|X| \geq t] \to 0 \)

4.5. A Borel set \( S \) is called a \textbf{support} of \( X \) if \( P^X(S^c) = 0 \) (or equivalently \( P^X(S) = 1 \))

4.1 Random Variable

Definition 4.6. A real-valued function \( X(\omega) \) defined for points \( \omega \) in a sample space \( \Omega \) is called a \textit{random variable}.

- Random variables are important because they provide a compact way of referring to events via their numerical attributes.
- The abbreviation r.v. will be used for “real-valued random variables” [11] p. 1.
- Technically, a random variable must be \textit{measurable}.

4.7. At a certain point in most probability courses, the sample space is rarely mentioned anymore and we work directly with random variables. The sample space often “disappears” but it is really there in the background.

4.8. For \( B \in \mathbb{R} \), we use the shorthand

- \( [X \in B] = \{\omega \in \Omega : X(\omega) \in B\} \) and
- \( P[X \in B] = P([X \in B]) = P(\{\omega \in \Omega : X(\omega) \in B\}) \).
• In particular, \( P[X < x] \) is a shorthand for \( P(\{\omega \in \Omega : X(\omega) < x\}) \).

4.9. If \( X \) and \( Y \) are random variables, we use the shorthand

- \( [X \in B, Y \in C] = \{\omega \in \Omega : X(\omega) \in B \text{ and } Y(\omega) \in C\} = [X \in B] \cap [Y \in C] \).
- \( P[X \in B, Y \in C] = P([X \in B] \cap [Y \in C]) \).

4.10. Every random variable can be written as a sum of a discrete random variable and a continuous random variable.

4.11. A random variable can have at most countably many point \( x \) such that \( P[X = x] > 0 \).

4.12. Point masses probability measures / Dirac measures, usually written \( \delta_{\alpha} \), is used to denote point mass of size one at the point \( \alpha \). In this case,

- \( P^X \{\alpha\} = 1 \)
- \( P^X (\{\alpha\}^c) = 0 \)
- \( F_X (x) = 1_{[\alpha, \infty)} (x) \)

4.13. There exists distributions that are neither discrete nor continuous.
Let \( \mu (A) = \frac{1}{2} \mu_1 (A) + \frac{1}{2} \mu_2 (A) \) for \( \mu_1 \) discrete and \( \mu_2 \) coming from a density.

4.14. When \( X \) and \( Y \) take finitely many values, say \( x_1, \ldots, x_m \) and \( y_1, \ldots, y_n \), respectively, we can arrange the probabilities \( p_{X,Y}(x_i, y_j) \) in the \( m \times n \) matrix

\[
\begin{bmatrix}
p_{X,Y}(x_1, y_1) & p_{X,Y}(x_1, y_2) & \cdots & p_{X,Y}(x_1, y_n) \\
p_{X,Y}(x_2, y_1) & p_{X,Y}(x_2, y_2) & \cdots & p_{X,Y}(x_2, y_n) \\
\vdots & \vdots & \ddots & \vdots \\
p_{X,Y}(x_m, y_1) & p_{X,Y}(x_m, y_2) & \cdots & p_{X,Y}(x_m, y_n)
\end{bmatrix}.
\]

- The sum of the entries in the \( i \)th row is \( P_X (x_i) \), and the sum of the entries in the \( j \)th column is \( P_Y (y_j) \).
- The sum of all the entries in the matrix is one.

4.2 Distribution Function

4.15. The (cumulative) **distribution function** (cdf) induced by a probability \( P \) on \( (\mathbb{R}, \mathcal{B}) \) is the function \( F (x) = P (\{\omega \in \Omega : X(\omega) < x\}) \).

The (cumulative) **distribution function** (cdf) of the random variable \( X \) is the function \( F_X (x) = P^X (\{\omega \in \Omega : X(\omega) < x\}) = P[X \leq x] \).

- The distribution \( P^X \) can be obtained from the distribution function by setting \( P^X (\{\omega \in \Omega : X(\omega) \leq x\}) = F_X (x) \); that is \( F_X \) uniquely determines \( P_X \)
- \( 0 \leq F_X \leq 1 \)

38
C1 $F_X$ is non-decreasing

C2 $F_X$ is right continuous:

$$\forall x F_X (x^+) \equiv \lim_{y \to x^{-}} F_X (y) \equiv \lim_{y \to x^{-}} F_X (y) = F_X (x) = P [X \leq x]$$

- The function $g(x) = P[X < x]$ is left-continuous in $x$.

C3 \( \lim_{x \to -\infty} F_X (x) = 0 \) and \( \lim_{x \to \infty} F_X (x) = 1 \).

- \( \forall x F_X (x^-) \equiv \lim_{y \to x^{-}} F_X (y) \equiv \lim_{y \to x^{-}} F_X (y) = P_X (-\infty, x) = P [X < x] \)

- \( P [X = x] = P_X \{x\} = F (x) - F (x^-) \) is the jump or saltus in $F$ at $x$

- \( \forall x < y \)
  $$
  P ([x, y]) = F (y) - F (x) \\
  P ([x, y]) = F (y) - F (x^-) \\
  P ([x, y]) = F (y^-) - F (x^-) \\
  P ([x, y]) = F (y^-) - F (x) \\
  P \{x\} = F (x) - F (x^-)
  $$

- A function $F$ is the distribution function of some probability measure on $(\mathbb{R}, \mathcal{B})$ if and only if one has (C1), (C2), and (C3).

- If $F$ satisfies (C1), (C2), and (C3), then \( \exists \) a unique probability measure $P$ on $(\mathbb{R}, \mathcal{B})$ that has \( P (a, b] = F (b) - F (a) \) \( \forall a, b \in \mathbb{R} \)

- $F_X$ is continuous if and only if $P [X = x] = 0$

- $F_X$ is continuous if and only if $P_X$ is continuous.

- $F_X$ has at most countably many points of discontinuity.
Definition 4.16. It is traditional to write $X \sim F$ to indicate that “$X$ has distribution $F$” [23 p. 25].

Definition 4.17. $F_{X,A}(x) = P([X \leq x] \cap A)$

4.18. Left-continuous inverse: $g^{-1}(y) = \inf \{x \in \mathbb{R} : g(x) \geq y\}$, $y \in (0, 1)$

![Left-continuous inverse on (0,1)](image)

- **Trick**: Just flip the graph along the line $x = y$, then make the graph left-continuous.
- If $g$ is a cdf, then only consider $y \in (0, 1)$. It is called the inverse CDF [7 Def 8.4.1, p. 238] or quantile function.
  - In [23 Def 2.16, p. 25], the inverse CDF is defined using strict inequality “$>$” rather than “$\geq$”.
- See table 1 for examples.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$F$</th>
<th>$F^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$1 - e^{-\lambda x}$</td>
<td>$-\frac{1}{\lambda} \ln(u)$</td>
</tr>
<tr>
<td>Extreme value</td>
<td>$1 - e^{-\frac{x-a}{b}}$</td>
<td>$a + b \ln \ln u$</td>
</tr>
<tr>
<td>Geometric</td>
<td>$1 - (1 - p)^i$</td>
<td>$\frac{\ln u}{\ln(1-p)}$</td>
</tr>
<tr>
<td>Logistic</td>
<td>$1 - \frac{1}{1 + e^{-x/\mu}}$</td>
<td>$\mu - b \ln \left(\frac{1}{u} - 1\right)$</td>
</tr>
<tr>
<td>Pareto</td>
<td>$1 - u^{-\alpha}$</td>
<td>$u^{-\frac{1}{\alpha}}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$1 - e^{\left(-\frac{x}{\beta}\right)^{\alpha}}$</td>
<td>$a \left(\ln u\right)^{\frac{1}{\beta}}$</td>
</tr>
</tbody>
</table>

**Table 1**: Left-continuous inverse
Definition 4.19. Let $X$ be a random variable with distribution function $F$. Suppose that $p \in (0, 1)$. A value of $x$ such that $F(x^-) = P[X < x] \leq p$ and $F(x) = P[X \leq x] \geq p$ is called a quantile of order $p$ for the distribution. Roughly speaking, a quantile of order $p$ is a value where the cumulative distribution crosses $p$. Note that it is not unique. Suppose $F(x) = p$ on an interval $[a, b]$, then all $x \in [a, b]$ are quantile of order $p$.

A quantile of order $\frac{1}{2}$ is called a median of the distribution. When there is only one median, it is frequently used as a measure of the center of the distribution. A quantile of order $\frac{1}{4}$ is called a first quartile and the quantile of order $\frac{3}{4}$ is called a third quartile. A median is a second quartile.

Assuming uniqueness, let $q_1$, $q_2$, and $q_3$ denote the first, second, and third quartiles of $X$. The interquartile range is defined to be $q_3 - q_1$, and is sometimes used as a measure of the spread of the distribution with respect to the median. The five parameters $\max X, q_1, q_2, q_3, \min X$ are often referred to as the five-number summary. Graphically, the five numbers are often displayed as a boxplot.

4.20. If $F$ is non-decreasing, right continuous, with $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$, then $F$ is the CDF of some probability measure on $(\mathbb{R}, B)$.

In particular, let $U \sim U(0, 1)$ and $X = F^{-1}(U)$, then $F_X = F$. Here, $F^{-1}$ is the left-continuous inverse of $F$. Note that we just explicitly define a random variable $X(\omega)$ with distribution function $F$ on $\Omega = (0, 1)$.

- For example, to generate $X \sim \mathcal{E}(\lambda)$, set $X = -\frac{1}{\lambda} \ln(U)$

4.21. Random Variable Generation

(a) Inverse-Transform Method: To generate a random variable $X$ with CDF $F$, set $X = F^{-1}(U)$ where $U$ is uniform on $(0, 1)$. See also 4.20.

(b) Acceptance-Rejection Method: To generate a random variable $X$ with pdf $f$, first find an easy-to-generate pdf $g$ and a constant $C$ such that $Cg \geq f$.

1. Generate a random variable $Z$ from $g(z)$.
2. Generate a uniform random variable $U$ on $(0, 1)$ independently of $Z$.
3. If $U \leq \frac{f(Z)}{Cg(Z)}$, then return $X = Z$ (“accept”). Otherwise, go back to step (1) (“reject”).

4.3 Discrete random variable

Definition 4.22. A random variable $X$ is said to be a discrete random variable if there exists countable distinct real numbers $x_k$ such that

$$\sum_k P[X = x_k] = 1.$$
X is completely determined by the values \( \mu_X (\{x_1\}), \mu_X (\{x_2\}), \ldots \)

- \( p_i = p_X (x_i) = P [X = x_i] \)

**Definition 4.23.** When \( X \) is a discrete random variable taking distinct values \( x_k \), we define its **probability mass function** (pmf) by

\[
p_X(x_k) = P[X = x_k].
\]

- We can use stem plot to visualize \( p_X \).
- If \( \Omega \) is countable, then there can be only countably many value of \( X(\omega) \). So, any random variable defined on countable \( \Omega \) is discrete.
- Sometimes, we write \( p(x_k) \) or \( p_{x_k} \) in stead of \( p_X(x_k) \).
- \( P[X \in B] = \sum_{x_k \in B} P[X = x_k] \).
- \( F_X(x) = \sum_{x_k} p_X(x_k) U(x - x_k) \).

**Definition 4.24** (Discrete CDF). A cdf which can be written in the form \( F_d(x) = \sum_k p_k U(x - x_k) \) is called a discrete cdf [7, Def. 5.4.1, p. 163]. Here, \( U \) is the unit step function, \( \{x_k\} \) is an arbitrary countable set of real numbers, and \( \{p_k\} \) is a countable set of positive numbers that sum to 1.

**Definition 4.25.** An integer-valued random variable is a discrete random variable whose distinct values are \( x_k = k \).

For integer-valued random variables,

\[
P[X \in B] = \sum_{k \in B} P[X = k].
\]

**4.26.** Properties of pmf

- \( p : \Omega \to [0, 1] \).
- \( 0 \leq p_X \leq 1 \).
- \( \sum_k p_X(x_k) = 1 \).

**Definition 4.27.** Sometimes, it is convenient to work with the “pdf” of a discrete r.v. Given that \( X \) is a discrete random variable which is defined as in definition 4.23. Then, the “pdf” of \( X \) is

\[
f_X(x) = \sum_{x_k} p_X(x_k) \delta(x - x_k), \quad x \in \mathbb{R}.
\]

Although the delta function is not a well-defined function\(^3\), this technique does allow easy manipulation of mixed distribution. The definition of quantities involving discrete random variables and the corresponding properties can then be derived from the pdf and hence there is no need to talk about pmf at all.

\(^3\)Rigorously, it is a unit measure at 0.
4.4 Continuous random variable

Definition 4.28. A random variable $X$ is said to be a **continuous random variable** if and only if any one of the following equivalent conditions holds.

$\equiv \forall x, P[X = x] = 0$

$\equiv \forall$ countable set $C, P^X (C) = 0$

$\equiv F_X$ is continuous

4.29. $f$ is (**probability**) **density function** $f$ (with respect to Lebesgue measure) of a random variable $X$ (or the distribution $P^X$)

$\equiv P^X$ have density $f$ with respect to Lebesgue measure.

$\equiv P^X$ is **absolutely continuous** w.r.t. the Lebesgue measure ($P^X \ll \lambda$) with $f = \frac{dP^X}{dx}$, the Radon-Nikodym derivative.

$\equiv f$ is a nonnegative Borel function on $\mathbb{R}$ such that $\forall B \in \mathcal{B}_\mathbb{R} P^X (B) = \int_B f(x)dx = \int_B f d\lambda$ where $\lambda$ is the Lebesgue measure. (This extends nicely to the random vector case.)

$\equiv X$ is **absolutely continuous**

$\equiv X$ (or $F^X$) comes from the density $f$

$\equiv \forall x \in \mathbb{R} F^X (x) = \int_{-\infty}^{x} f(t)dt$

$\equiv \forall a, b F^X (b) - F^X (a) = \int_{a}^{b} f(x)dx$

4.30. If $F$ does differentiate to $f$ and $f$ is continuous, it follows by the fundamental theorem of calculus that $f$ is indeed a density for $F$. That is, if $F$ has a continuous derivative, this derivative can serve as the density $f$.

4.31. Suppose a random variable $X$ has a density $f$.

- $F$ need not differentiate to $f$ everywhere.
  - When $X \sim \mathcal{U}(a, b)$, $F_X$ is not differentiable at $a$ nor $b$.

- $\int f (x) dx = 1$

- $f$ is determined only Lebesgue-a.e. That is, If $g = f$ Lebesgue-a.e., then $g$ can also serve as a density for $X$ and $P^X$

- $f$ is nonnegative a.e. [9, stated on p. 138]
• $X$ is a continuous random variable

• $f$ at its continuity points must be the derivative of $F$

• $P[X \in [a, b]] = P[X \in [a, b)] = P[X \in (a, b)] = P[X \in (a, b)]$ because the corresponding integrals over an interval are not affected by whether or not the endpoints are included or excluded. In other words, $P[X = a] = P[X = b] = 0$.

• $P[f_X(X) = 0] = 0$

4.32. $f_X(x) = \mathbb{E}[\delta(X - x)]$

**Definition 4.33** (Absolutely Continuous CDF). An absolutely continuous cdf $F_{ac}$ can be written in the form

$$F_{ac}(x) = \int_{-\infty}^{x} f(z) dz,$$

where the integrand,

$$f(x) = \frac{d}{dx} F_{ac}(x),$$

is defined a.e., and is a nonnegative, integrable function (possibly having discontinuities) satisfying

$$\int f(x) dx = 1.$$

4.34. Any nonnegative function that integrates to one is a **probability density function** (pdf) [9, p. 139].

4.35. Remarks: Some useful intuitions

(a) Approximately, for a small $\Delta x$, $P[X \in [x, x + \Delta x]] = \int_{x}^{x+\Delta x} f_X(t) dt \approx f - X(x) \Delta x$.

This is why we call $f_X$ the density function.

(b) In fact, $f_X(x) = \lim_{\Delta x \to 0} \frac{P[x < X \leq x + \Delta x]}{\Delta x}$

4.36. Let $T$ be an absolutely continuous nonnegative random variable with cumulative distribution function $F$ and density $f$ on the interval $[0, \infty)$. The following terms are often used when $T$ denotes the lifetime of a device or system.

(a) Its survival-, survivor-, or reliability-function is:

$$R(t) = P[T > t] = \int_{t}^{\infty} f(x) dx = 1 - F(t).$$

• $R(0) = P[T > 0] = P[T \geq 0] = 1$.

(b) The mean time of failure (MTTF) = $\mathbb{E}[T] = \int_{0}^{\infty} R(t) dt$. 

44
(c) The (age-specific) failure rate or hazard function of a device or system with lifetime $T$ is

$$r(t) = \lim_{\delta \to 0} \frac{P[T \leq t + \delta | T > t]}{\delta} = -\frac{R'(t)}{R(t)} = \frac{f(t)}{R(t)} = \frac{d}{dt} \ln R(t).$$

(i) $r(t) \delta \approx P[T \in (t, t + \delta) | T > t]$

(ii) $R(t) = e^{-\int_0^t r(\tau)d\tau}$.

(iii) $f(t) = r(t)e^{-\int_0^t r(\tau)d\tau}$

- For $T \sim \mathcal{E}(\lambda), r(t) = \lambda$.

See also [9, section 5.7].

**Definition 4.37.** A random variable whose cdf is continuous but whose derivative is the zero function is said to be **singular**.

- See Cantor-type distribution in [5, p. 35–36].
- It has no density. (Otherwise, the cdf is the zero function.) So, $\exists$ continuous random variable $X$ with no density. Hence, $\exists$ random variable $X$ with no density.
- Even when we allow the use of delta function for the density as in the case of mixed r.v., it still has no density because there is no jump in the cdf.
- There exists singular r.v. whose cdf is strictly increasing.

**Definition 4.38.** $f_{X,A}(x) = \frac{d}{dx}F_{X,A}(x)$. See also definition 4.17.

### 4.5 Mixed/hybrid Distribution

There are many occasion where we have a random variable whose distribution is a combination of a normalized linear combination of discrete and absolutely continuous distributions. For convenience, we use the Dirac delta function to link the pmf to a pdf as in definition 4.27. Then, we only have to concern about the pdf of a mixed distribution/r.v.

**4.39.** By allowing density functions to contain impulses, the cdfs of mixed random variables can be expressed in the form $F(x) = \int_{(-\infty,x]} f(t)dt$.

**4.40.** Given a cdf $F_X$ of a mixed random variable $X$, the density $f_X$ is given by

$$f_X(x) = \tilde{f}_X(x) + \sum_k P[X = x_k] \delta(x - x_k),$$

where

- the $x_i$ are the distinct points at which $F_X$ has jump discontinuities, and
- $\tilde{f}_X(x) = \begin{cases} F'_X(x), & F_X \text{ is differentiable at } x \\ 0, & \text{otherwise.} \end{cases}$
In which case,
\[ \mathbb{E}[g(X)] = \int g(x)\,f_X(x)\,dx + \sum_k g(x_k)P[X = x_k]. \]

Note also that \( P[X = x_k] = F(x_k) - F(x^-_k) \)

**4.41.** Suppose the cdf \( F \) can be expressed in the form \( F(x) = G(x)U(x - x_0) \) for some function \( G \). Then, the density is \( f(x) = G'(x)U(x - x_0) + G(x)\delta(x - x_0) \). Note that \( G(x_0) = F(x_0) = P[X = x_0] \) is the jump of the cdf at \( x_0 \). When the random variable is continuous, \( G(x_0) = 0 \) and thus \( f(x) = G'(x)U(x - x_0) \).

### 4.6 Independence

**Definition 4.42.** A family of random variables \( \{X_i : i \in I\} \) is **independent** if \( \forall \) finite \( J \subset I \), the family of random variables \( \{X_i : i \in J\} \) is independent. In words, “an infinite collection of random elements is by definition independent if each finite subcollection is.” Hence, we only need to know how to test independence for finite collection.

(a) \((E_i, \mathcal{E}_i)\)'s are not required to be the same.

(b) The collection of random variables \( \{1_{A_i} : i \in I\} \) is independent iff the collection of events (sets) \( \{A_i : i \in I\} \) is independent.

**Definition 4.43.** Independence among finite collection of random variables: For finite \( I \), the following statements are equivalent

\[ \equiv (X_i)_{i \in I} \text{ are independent (or mutually independent [2 p 182])}. \]
\[ \equiv P[X_i \in H_i \forall i \in I] = \prod_{i \in I} P[X_i \in H_i] \text{where } H_i \in \mathcal{E}_i \]
\[ \equiv P\left[(X_i : i \in I) \in \times_{i \in I} H_i\right] = \prod_{i \in I} P[X_i \in H_i] \text{ where } H_i \in \mathcal{E}_i \]
\[ \equiv P^{(X_i: i \in I)}(\times_{i \in I} H_i) = \prod_{i \in I} P^{X_i}(H_i) \text{ where } H_i \in \mathcal{E}_i \]
\[ \equiv P[X_i \leq x_i \forall i \in I] = \prod_{i \in I} P[X_i \leq x_i] \]
\[ \equiv [\text{Factorization Criterion}] \quad F_{(X_i:i \in I)}((x_i : i \in I)) = \prod_{i \in I} F_{X_i}(x_i) \]
\[ \equiv X_i \text{ and } X_i^{i-1} \text{ are independent } \forall i \geq 2 \]
\[ \equiv \sigma(X_i) \text{ and } \sigma(X_i^{i-1}) \text{ are independent } \forall i \geq 2 \]
\[ \equiv \text{Discrete random variables } X_i \text{'s with countable range } E: \]
\[ P[X_i = x_i \forall i \in I] = \prod_{i \in I} P[X_i = x_i] \quad \forall x_i, E \forall i \in I \]
\[ \equiv \text{Absolutely continuous } X_i \text{ with density } f_{X_i} : f_{(X_i:i \in I)}((x_i : i \in I)) = \prod_{i \in I} f_{X_i}(x_i) \]
Definition 4.44. If the $X_\alpha$, $\alpha \in I$ are independent and each has the same marginal distribution with distribution $Q$, we say that the $X_\alpha$’s are iid (independent and identically distributed) and we write $X_\alpha \overset{iid}{\sim} Q$

- The abbreviation can be IID \[23, p 39\].

Definition 4.45. A pairwise independent collection of random variables is a set of random variables any two of which are independent.

(a) Any collection of (mutually) independent random variables is pairwise independent

(b) Some pairwise independent collections are not independent. See example (4.46).

Example 4.46. Let suppose $X$, $Y$, and $Z$ have the following joint probability distribution: $p_{X,Y,Z}(x,y,z) = \frac{1}{4}$ for $(x,y,z) \in \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$. This, for example, can be constructed by starting with independent $X$ and $Y$ that are Bernoulli-$\frac{1}{2}$. Then set $Z = X \oplus Y = X + Y \mod 2$.

(a) $X,Y,Z$ are pairwise independent.

(b) The combination of $X \parallel Z$ and $Y \parallel Z$ does not imply $(X,Y) \parallel Z$.

Definition 4.47. The convolution of probability measures $\mu_1$ and $\mu_2$ on $(\mathbb{R},\mathcal{B})$ is the measure $\mu_1 \ast \mu_2$ defined by

$$(\mu_1 \ast \mu_2)(H) = \int \mu_2(H - x)\mu_1(dx)$$ $H \in \mathcal{B}_\mathbb{R}$

(a) $\mu_X \ast \mu_Y = \mu_Y \ast \mu_X$ and $\mu_X \ast (\mu_Y \ast \mu_Z) = (\mu_X \ast \mu_Y) \ast \mu_Z$

(b) If $F_X$ and $G_X$ are distribution functions corresponding to $\mu_X$ and $\mu_Y$, the distribution function corresponding to $\mu_X \ast \mu_Y$ is

$$(\mu_X \ast \mu_Y)(-\infty, z] = \int F_Y(z - x)\mu_X(dx)$$

In this case, it is notationally convenient to replace $\mu_X(dx)$ by $dF_X(x)$ (Stieltjes Integral.) Then, $$(\mu_X \ast \mu_Y)(-\infty, z] = \int F_Y(z - x)dF_X(x).$$ This is denoted by $(F_X \ast F_Y)(z)$. That is

$$F_Z(z) = (F_X \ast F_Y)(z) = \int F_Y(z - x)dF_X(x)$$

(c) If density $f_Y$ exists, $F_X \ast F_Y$ has density $F_X \ast f_Y$, where

$$(F_X \ast f_Y)(z) = \int f_Y(z - x)dF_X(x)$$

(i) If $Y$ (or $F_Y$) is absolutely continuous with density $f_Y$, then for any $X$ (or $F_X$), $X + Y$ (or $F_X \ast F_Y$) is absolutely continuous with density $(F_X \ast f_Y)(z) = \int f_Y(z - x)dF_X(x)$

47
If, in addition, $F_X$ has density $f_X$. Then,

$$\int f_Y (z - x) dF_X (x) = \int f_Y (z - x) f_X (x) \, dx$$

This is denoted by $f_X \ast f_Y$.

In other words, if densities $f_X, f_Y$ exist, then $F_X \ast F_Y$ has density $f_X \ast f_Y$, where

$$(f_X \ast f_Y) (z) = \int f_Y (z - x) f_X (x) \, dx$$

4.48. If random variables $X$ and $Y$ are independent and have distribution $\mu_X$ and $\mu_Y$, then $X + Y$ has distribution $\mu_X \ast \mu_Y$.

4.49. Expectation and independence

(a) Let $X$ and $Y$ be nonnegative independent random variables on $(\Omega, \mathcal{A}, P)$, then $E [XY] = E X E Y$.

(b) If $X_1, X_2, \ldots, X_n$ are independent and $g_k$'s complex-valued measurable function. Then $g_k (X_k)$'s are independent. Moreover, if $g_k (X_k)$ is integrable, then $E \left[ \prod_{k=1}^{n} g_k (X_k) \right] = \prod_{k=1}^{n} E [g_k (X_k)]$.

(c) If $X_1, \ldots, X_n$ are independent and $\sum_{k=1}^{n} X_k$ has a finite second moment, then all $X_k$ have finite second moments as well. Moreover, $\text{Var} \left[ \sum_{k=1}^{n} X_k \right] = \sum_{k=1}^{n} \text{Var} X_k$.

(d) If pairwise independent $X_i \in L^2$, then $\text{Var} \left[ \sum_{i=1}^{k} X_i \right] = \sum_{i=1}^{k} \text{Var} [X_i]$.

4.7 Misc

4.50. The mode of a discrete probability distribution is the value at which its probability mass function takes its maximum value. The mode of a continuous probability distribution is the value at which its probability density function attains its maximum value.

- the mode is not necessarily unique, since the probability mass function or probability density function may achieve its maximum value at several points.

5 PMF Examples

The following pmf will be defined on its support $S$. For $\Omega$ larger than $S$, we will simply put the pmf to be 0.
Table 2: Examples of probability mass functions. Here, \( p, \beta \in (0, 1). \lambda > 0. \)

### 5.1 Random/Uniform

5.1. \( \mathcal{U}_n \)

When an experiment results in a finite number of “equally likely” or “totally random” outcomes, we model it with a uniform random variable. We say that \( X \) is uniformly distributed on \( [n] \) if

\[
P[X = k] = \frac{1}{n}, \quad k \in [n].
\]

We write \( X \sim \mathcal{U}_n \).

- \( p_i = \frac{1}{n} \) for \( i \in S = \{1, 2, \ldots, n\} \).
- Examples
  - classical game of chance / classical probability drawing at random
  - fair gaming devices (well-balanced coins and dice, well shuffled decks of cards)
  - experiment where
    * there are only \( n \) possible outcomes and they are all equally probable
    * there is a balance of information about outcomes

5.2. Uniform on a finite set: \( \mathcal{U}(S) \)

Suppose \( |S| = n \), then \( p(x) = \frac{1}{n} \) for all \( x \in S \).

Example 5.3. For \( X \) uniform on \([-M:1:M]\), we have \( \mathbb{E}X = 0 \) and \( \text{Var} X = \frac{M(M+1)}{3} \).

For \( X \) uniform on \([N:1:M]\), we have \( \mathbb{E}X = \frac{M-N}{2} \) and \( \text{Var} X = \frac{1}{12}(M-N)(M-N-2) \).

Example 5.4. Set \( S = 0, 1, 2, \ldots, M \), then the sum of two independent \( \mathcal{U}(S) \) has pmf

\[
p(k) = \frac{(M+1) - |k-M|}{(M+1)^2}
\]

for \( k = 0, 1, \ldots, 2M \). Note its triangular shape with maximum value at \( p(M) = \frac{1}{M+1} \). To visualize the pmf in MATLAB, try
k = 0:2*M;
P = (1/((M+1)^2))*ones(1,M+1);
P = conv(P,P); stem(k,P)

5.2 Bernoulli and Binary distributions

5.5. Bernoulli: \( B(1, p) \) or Bernoulli\((p)\)

- \( S = \{0, 1\} \)
- \( p_0 = q = 1 - p, \) \( p_1 = p \)
- \( \mathbb{E}X = \mathbb{E}[X^2] = p. \)
  \( \text{Var}[X] = p - p^2 = p(1-p). \) Note that the variance is maximized at \( p = 0.5. \)

5.6. Binary: Suppose \( X \) takes only two values \( a \) and \( b \) with \( b > a. \) \( P[X = b] = p. \)

(a) \( X \) can be expressed as \( X = (b - a)I + a, \) where \( I \) is a Bernoulli random variable with \( P[I = 1] = p. \)

(b) \( \text{Var} X = (b - a)^2 \text{Var} I = (b - a)^2 p (1-p). \) Note that it is still maximized at \( p = \frac{1}{2}. \)

(c) Suppose \( a = -b. \) Then, \( X = 2I + a = 2I - b. \) In which case, \( \text{Var} X = 2b^2p (1-p). \)

(d) Suppose \( X_k \) are independent random variables taking on two values \( b \) and \( a \) with \( P[X_k = p] = p = 1 - P[X_k = a] \) where \( b > a. \) Then, the sum \( S = \sum_{k=1}^{n} X_k \) is “binomial” on \( \{k(b - a) + an : k = 0, 1, \ldots, n\} \) where the probability at point \( \frac{k(b - a) + an}{n} \) is \( (\frac{n!}{k!(n-k)!})p^k(1-p)^{n-k}. \)
5.3 Binomial: $\mathcal{B}(n, p)$

5.7 Binomial distribution with size $n$ and parameter $p$, $p \in [0, 1]$.

(a) $p_i = \binom{n}{i} p^i (1-p)^{n-i}$ for $i \in S = \{0, 1, 2, \ldots, n\}$

- Use `binopdf(i,n,p)` in MATLAB.

(b) $X$ is the number of success in $n$ independent Bernoulli trials and hence the sum of $n$ independent, identically distributed Bernoulli r.v.

(c) $\varphi_X(u) = (1 - p + pe^{ju})^n$

(d) $\mathbb{E}X = np$

(e) $\mathbb{E}X^2 = (np)^2 + np(1-p)$

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

(g) Tail probability: $\sum_{r=k}^{n} \binom{n}{r} p^r (1-p)^{n-r} = I_p(k, n-k+1)$

(h) Maximum probability value happens at $k_{\text{max}} = \text{mode } X = \lfloor (n+1)p \rfloor \approx np$

- When $(n+1)p$ is an integer, then the maximum is achieved at $k_{\text{max}}$ and $k_{\text{max}} - 1$.

(i) By (2),

$$P[X \text{ is even}] = \frac{1}{2} (1 + (1 - 2p)^n), \quad \text{and} \quad P[X \text{ is odd}] = \frac{1}{2} (1 - (1 - 2p)^n).$$

(j) If we have $\mathcal{E}_1, \ldots, \mathcal{E}_n$, $n$ unlinked repetition of $\mathcal{E}$ and event $A$ for $\mathcal{E}$, the the distribution $\mathcal{B}(n,p)$ describe the probability that $A$ occurs $k$ times in $\mathcal{E}_1, \ldots, \mathcal{E}_n$.

(k) Gaussian Approximation for Binomial Probabilities: When $n$ is large, binomial distribution becomes difficult to compute directly because of the need to calculate factorial terms. We can use

$$P[X = k] \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}}, \quad (14)$$

which comes from approximating $X$ by Gaussian $Y$ with the same mean and variance and the relation

$$P[X = k] \approx P[X \leq k] - P[X \leq k-1] \approx P[Y \leq k] - P[Y \leq k-1] \approx f_Y(k).$$

See also (12.23).

(l) Approximation: $\binom{n}{k} p^k (1-p)^{n-k} = \frac{(np)^k}{k!} e^{-np} \left( 1 + O \left( np^2 \frac{k^2}{n} \right) \right)$.
5.4 Geometric: \(G(\beta)\)

A geometric distribution is defined by the fact that for some \(\beta \in [0, 1]\), \(p_{k+1} = \beta p_k\) for all \(k \in S\) where \(S\) can be either \(\mathbb{N}\) or \(\mathbb{N} \cup \{0\}\).

- When its support is \(\mathbb{N}\), \(p_k = (1 - \beta) \beta^{k-1}\). This is referred to as \(G_1(\beta)\) or \(\text{geometric}_1(\beta)\). In MATLAB, use \(\text{geopdf}(k-1, 1-\beta)\).
- When its support is \(\mathbb{N} \cup \{0\}\), \(p_k = (1 - \beta) \beta^k\). This is referred to as \(G_0(\beta)\) or \(\text{geometric}_0(\beta)\). In MATLAB, use \(\text{geopdf}(k, 1-\beta)\).

5.8. Consider \(X \sim G_0(\beta)\).

- \(p_i = (1 - \beta) \beta^i\), for \(S = \mathbb{N} \cup \{0\}\), \(0 \leq \beta < 1\)
- \(\beta = \frac{m}{m+1}\) where \(m\) = average waiting time/lifetime
- \(P[X = k] = P[k\ \text{failures followed by a success}] = (P[\text{failures}])^k P[\text{success}]\)
- \(P[X \geq k] = \beta^k\) = the probability of having at least \(k\) initial failure = the probability of having to perform at least \(k+1\) trials.
- \(P[X > k] = \beta^{k+1}\) = the probability of having at least \(k+1\) initial failure.

- Memoryless property:
  - \(P[X \geq k + c | X \geq k] = P[X \geq c], k, c > 0\).
  - \(P[X > k + c | X \geq k] = P[X > c], k, c > 0\).
  - If a success has not occurred in the first \(k\) trials (already fails for \(k\) times), then the probability of having to perform at least \(j\) more trials is the same the probability of initially having to perform at least \(j\) trials.
  - Each time a failure occurs, the system “forgets” and begins anew as if it were performing the first trial.
○ Geometric r.v. is the only discrete r.v. that satisfies the memoryless property.

• Ex.

○ lifetimes of components, measured in discrete time units, when the fail catastrophically (without degradation due to aging)

○ waiting times
  * for next customer in a queue
  * between radioactive disintegrations
  * between photon emission

○ number of repeated, unlinked random experiments that must be performed prior to the first occurrence of a given event $A$
  * number of coin tosses prior to the first appearance of a ‘head’
  * number of trials required to observe the first success

• The sum of independent $G_0(p)$ and $G_0(q)$ has pmf

\[
\begin{cases}
(1-p)(1-q) \frac{k+1-p^{k+1}}{q-p}, & p \neq q \\
(k+1)(1-p)^2 p^k, & p = q
\end{cases}
\]

for $k \in \mathbb{N} \cup \{0\}$.

5.9. Consider $X \sim G_1(\beta)$.

• $P[X > k] = \beta^k$

• Suppose independent $X_i \sim G_1(\beta_i)$. $\min(X_1, X_2, \ldots, X_n) \sim G_1(\prod_{i=1}^n \beta_i)$.

5.5 Poisson Distribution: $P(\lambda)$

5.10. Characterized by

• $p_X(k) = P[X = k] = e^{-\lambda} \frac{\lambda^k}{k!}$; or equivalently,

• $\varphi_X(u) = e^{\lambda(e^u - 1)}$, where $\lambda \in (0, \infty)$ is called the parameter or intensity parameter of the distribution. In MATLAB, use poisspdf(k,lambda).

5.11. Denoted by $P(\lambda)$.

5.12. In stead of $X$, Poisson random variable is usually denoted by $\Lambda$.

5.13. $EX = Var X = \lambda$.

5.14. Successive probabilities are connected via the relation $kp_X(k) = \lambda p_X(k - 1)$.

5.15. mode $X = \lfloor \lambda \rfloor$. 

53
| $0 < \lambda < 1$ | $0$ | $e^{-\lambda}$ |
| \(\lambda \in \mathbb{N}\) | $\lambda - 1, \lambda$ | $\frac{\lambda!}{\lambda} e^{-\lambda}$ |
| $\lambda \geq 1, \lambda \notin \mathbb{N}$ | $\lfloor \lambda \rfloor$ | $\frac{\lambda^{(\lambda)}}{\lambda!} e^{-\lambda}$ |

- Note that when $\lambda \in \mathbb{N}$, there are two maxima at $\lambda - 1$ and $\lambda$.
- When $\lambda \gg 1$, $p_x (\lfloor \lambda \rfloor) \approx \frac{1}{\sqrt{2\pi \lambda}}$ via the Stirling’s formula (1.13).

5.16. $P \{X \geq 2\} = 1 - e^{-\lambda} - \lambda e^{-\lambda} = O (\lambda^2)$.

The cumulative probabilities can be found by

$$P \{X \leq k\} \overset{(\ast)}{=} \mathbb{P} \left[ \sum_{i=1}^{k+1} X_i > 1 \right] = \frac{1}{\Gamma(k+1)} \int_0^\infty e^{-t^2} dt,$$

$$P \{X > k\} = P \{X \geq k + 1\} \overset{(\ast)}{=} P \left[ \sum_{i=1}^{k+1} X_i \leq 1 \right] = \frac{1}{\Gamma(k+1)} \int_0^\lambda e^{-t^2} dt,$$

where the $X_i$’s are i.i.d. $\mathcal{E} (\lambda)$. The equalities given by $(\ast)$ are easily obtained via counting the number of events from rate-$\lambda$ Poisson process on interval $[0, 1]$.

5.17. **Fano factor** (index of dispersion): $\frac{\text{Var} X}{\mathbb{E} X} = 1$

An important property of the Poisson and Compound Poisson laws is that their classes are close under convolution (independent summation). In particular, we have divisibility properties (5.22) and (5.32) which are straightforward to prove from their characteristic functions.

5.18 (Recursion equations). Suppose $X \sim \mathcal{P}(\lambda)$. Let $m_k (\lambda) = \mathbb{E} [X^k]$ and $\mu_k (\lambda) = \mathbb{E} [(X - \mathbb{E} X)^k]$.

$$m_{k+1} (\lambda) = \lambda (m_k (\lambda) + m'_k (\lambda)) \quad (15)$$

$$\mu_{k+1} (\lambda) = \lambda (k \mu_{k-1} (\lambda) + \mu'_k (\lambda)) \quad (16)$$

[15 p 112]. Starting with $m_1 = \lambda = \mu_2$ and $\mu_1 = 0$, the above equations lead to recursive determination of the moments $m_k$ and $\mu_k$.

5.19. $\mathbb{E} \left[ \frac{1}{X+1} \right] = \frac{1}{\lambda} (1 - e^{-\lambda})$. Because for $d \in \mathbb{N}$, $Y = \frac{1}{X+1} \left( \sum_{n=0}^{d} a_n X^n \right)$ can be expressed as $\left( \sum_{n=0}^{d-1} b_n X^n \right) + \frac{e}{X+1}$, the value of $\mathbb{E} Y$ is easy to find if we know $\mathbb{E} X^n$.

5.20. **Mixed Poisson distribution**: Let $X$ be Poisson with mean $\lambda$. Suppose, that the mean $\lambda$ is chosen in accord with a probability distribution whose characteristic function is $\varphi_{\lambda}$. Then,

$$\varphi_X (u) = \mathbb{E} \left[ \mathbb{E} [e^{iuX} | \Lambda] \right] = \mathbb{E} \left[ e^{\Lambda (e^{iu} - 1)} \right] = \mathbb{E} \left[ e^{i(-i(e^{iu} - 1))\Lambda} \right] = \varphi_{\Lambda} (-i (e^{iu} - 1)).$$
\[ \mathbb{E}X = \mathbb{E}\Lambda. \]
\[ \text{Var } X = \text{Var } \Lambda + \mathbb{E}\Lambda. \]
\[ \mathbb{E}[X^2] = \mathbb{E}[\Lambda^2] + \mathbb{E}\Lambda. \]
\[ \text{Var}[X|\Lambda] = \mathbb{E}[X|\Lambda] = \Lambda. \]
\[ \text{When } \Lambda \text{ is a nonnegative integer-valued random variable, we have } G_X(z) = G_\Lambda(e^{z-1}) \text{ and } P[X = 0] = G_\Lambda(1). \]
\[ \mathbb{E}[X\Lambda] = \mathbb{E}[\Lambda^2] \]
\[ \text{Cov } [X, \Lambda] = \text{Var } \Lambda. \]

5.21. **Thinned Poisson**: Suppose we have \( X \to Y \) where \( X \sim \mathcal{P}(\lambda) \). The box \( s \) is a binomial channel with success probability \( s \). (Each 1 in the \( X \) get through the channel with success probability \( s \).)

- Note that \( Y \) is in fact a random sum \( \sum_{i=1}^{X} I_i \) where i.i.d. \( I_i \) has Bernoulli distribution with parameter \( s \).
- \( Y \sim \mathcal{P}(s\lambda) \);
- \( p(x|y) = e^{-\lambda(1-s)}\frac{(\lambda(1-s))^{x-y}}{(x-y)!}; x \geq y \) (shifted Poisson);

[Levy and Baxter, 2002]

5.22. **Finite additivity**: Suppose we have independent \( \Lambda_i \sim \mathcal{P}(\lambda_i) \), then \( \sum_{i=1}^{n} \Lambda_i \sim \mathcal{P}(\sum_{i=1}^{n} \lambda_i) \).

5.23. **Raikov’s theorem**: independent random variables can have their sum Poisson-distributed only if every component of the sum is Poisson-distributed.

5.24. **Countable Additivity Theorem** [12, p 5]: Let \( (X_j : j \in \mathbb{N}) \) be independent random variables, and assume that \( X_j \) has the distribution \( \mathcal{P}(\mu_j) \) for each \( j \). If

\[
\sum_{j=1}^{\infty} \mu_j
\]

converges to \( \mu \), then \( S = \sum_{j=1}^{\infty} X_j \) converges with probability 1, and \( S \) has distribution \( \mathcal{P}(\mu) \).

If on the other hand [17] diverges, then \( S \) diverges with probability 1.

5.25. Let \( X_1, X_2, \ldots, X_n \) be independent, and let \( X_j \) have distribution \( \mathcal{P}(\mu_j) \) for all \( j \). Then \( S_n = \sum_{j=1}^{n} X_j \) has distribution \( \mathcal{P}(\mu) \), with \( \mu = \sum_{j=1}^{n} \mu_j \); and so, whenever \( \sum_{j=1}^{n} r_j = s \),

\[
P[X_j = r_j \forall j | S_n = s] = \frac{s!}{r_1!r_2!\cdots r_n!} \prod_{j=1}^{n} \left( \frac{\mu_j}{\mu} \right)^{r_j}
\]

which follows the multinomial distribution [12, p 6–7].
• If \( X \) and \( Y \) are independent Poisson random variables with respective parameters \( \lambda \) and \( \mu \), then (1) \( Z = X + Y \) is \( P(\lambda + \mu) \) and (2) conditioned on \( Z = z \), \( X \) is \( B(z, \frac{\lambda}{\lambda+\mu}) \).

So, \( \mathbb{E}[X|Z] = \frac{\lambda}{\lambda+\mu} Z \), \( \text{Var}[X|Z] = Z \frac{\lambda \mu}{(\lambda+\mu)^2} \), and \( \mathbb{E}[\text{Var}[X|Z]] = \frac{\lambda \mu}{\lambda+\mu} \).

5.26. One of the reasons why Poisson distribution is important is because many natural phenomena can be modeled by Poisson processes. For example, if we consider the number of occurrences \( \Lambda \) during a time interval of length \( \tau \) in a rate-\( \lambda \) homogeneous Poisson process, then \( \Lambda \sim P(\lambda \tau) \).

Example 5.27.

• The first use of the Poisson model is said to have been by a Prussian physician, von Bortkiewicz, who found that the annual number of late-19th-century Prussian soldiers kicked to death by horses followed a Poisson distribution [7, p 150].

• #photons emitted by a light source of intensity \( \lambda \) [photons/second] in time \( \tau \)

• #atoms of radioactive material undergoing decay in time \( \tau \)

• #clicks in a Geiger counter in \( \tau \) seconds when the average number of click in 1 second is \( \lambda \).

• #dopant atoms deposited to make a small device such as an FET

• #customers arriving in a queue or workstations requesting service from a file server in time \( \tau \)

• Counts of demands for telephone connections

• number of occurrences of rare events in time \( \tau \)

• #soldiers kicked to death by horses

• Counts of defects in a semiconductor chip.

5.28. Normal Approximation to Poisson Distribution with large \( \lambda \): Let \( X \sim P(\lambda) \). \( X \) can be thought of as a sum of i.i.d. \( X_i \sim P(\lambda_n) \), i.e., \( X = \sum_{i=1}^{n} X_i \), where \( n\lambda_n = \lambda \). Hence \( X \) is approximately normal \( N(\lambda, \lambda) \) for \( \lambda \) large.

Some says that the normal approximation is good when \( \lambda > 5 \).

5.29. Poisson distribution can be obtained as a limit from negative binomial distributions. Thus, the negative binomial distribution with parameters \( r \) and \( p \) can be approximated by the Poisson distribution with parameter \( \lambda = \frac{rp}{p} \) (mean-matching), provided that \( p \) is “sufficiently” close to 1 and \( r \) is “sufficiently” large.
5.30. Convergence of sum of Bernoulli random variables to the Poisson Law

Suppose that for each $n \in \mathbb{N}$

$$X_{n,1}, X_{n,2}, \ldots, X_{n,r_n}$$

are independent; the probability space for the sequence may change with $n$. Such a collection is called a triangular array [1] or double sequence [8] which captures the nature of the collection when it is arranged as

$$
\begin{array}{cccc}
X_{1,1}, & X_{1,2}, & \ldots, & X_{1,r_1}, \\
X_{2,1}, & X_{2,2}, & \ldots, & X_{2,r_2}, \\
\vdots & \vdots & \cdots & \vdots \\
X_{n,1}, & X_{n,2}, & \ldots, & X_{n,r_n}, \\
\vdots & \vdots & \cdots & \vdots \\
\end{array}
$$

where the random variables in each row are independent. Let $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,r_n}$ be the sum of the random variables in the $n^{th}$ row.

Consider a triangular array of Bernoulli random variables $X_{n,k}$ with $P[X_{n,k} = 1] = p_{n,k}$. If $\max_{1 \leq k \leq r_n} p_{n,k} \to 0$ and $\sum_{k=1}^{r_n} p_{n,k} \to \lambda$ as $n \to \infty$, then the sums $S_n$ converges in distribution to the Poisson law. In other words, Poisson distribution rare events limit of the binomial (large $n$, small $p$).

As a simple special case, consider a triangular array of Bernoulli random variables $X_{n,k}$ with $P[X_{n,k} = 1] = p_n$. If $np_n \to \lambda$ as $n \to \infty$, then the sums $S_n$ converges in distribution to the Poisson law.

To show this special case directly, we bound the first $i$ terms of $n!$ to get $\frac{(n-i)^i}{i!} \leq \binom{n}{i} \leq \frac{n^i}{i!}$. Using the upper bound,

$$
\binom{n}{i} p_n^i (1 - p_n)^{n-i} \leq \frac{1}{i!} (np_n)^i (1 - p_n)^{n-i} \left(1 - \frac{np_n}{n}\right)^n.
$$

The lower bound gives the same limit because $(n-i)^i = \left(\frac{n-i}{n}\right)^i n^i$ where the first term $\to 1$.

5.6 Compound Poisson

Given an arbitrary probability measure $\mu$ and a positive real number $\lambda$, the compound Poisson distribution $\mathcal{CP}(\lambda, \mu)$ is the distribution of the sum $\sum_{j=1}^{\Lambda} V_j$ where the $V_j$ are i.i.d. with distribution $\mu$ and $\Lambda$ is a $\mathcal{P}(\lambda)$ random variable, independent of the $V_j$.

Sometimes, it is written as $\text{POIS}(\lambda \mu)$. The parameter $\lambda$ is called the rate of $\mathcal{CP}(\lambda, \mu)$ and $\mu$ is called the base distribution.

5.31. The mean and variance of $\mathcal{CP}(\lambda, \mathcal{L}(V))$ are $\lambda \mathbb{E} V$ and $\lambda \mathbb{E} V^2$ respectively.

5.32. If $Z \sim \mathcal{CP}(\lambda, q)$, then $\varphi_Z(t) = e^{\lambda (\varphi_q(t) - 1)}$. 

57
An important property of the Poisson and Compound Poisson laws is that their classes are close under convolution (independent summation). In particular, we have divisibility properties (5.22) and (5.32) which are straightforward to prove from their characteristic functions.

5.33. Divisibility property of the compound Poisson law: Suppose we have independent \( \Lambda_i \sim \mathcal{CP} (\lambda_i, \mu^{(i)}) \), then \( \sum_{i=1}^{n} \Lambda_i \sim \mathcal{CP} \left( \lambda \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i \mu^{(i)} \right) \) where \( \lambda = \sum_{i=1}^{n} \lambda_i \).

Proof.

\[
\varphi_{\sum_{i=1}^{n} \Lambda_i} (t) = \prod_{i=1}^{n} e^{\lambda_i (\varphi_{\mu^{(i)}} (t)) - 1} = \exp \left( \sum_{i=1}^{n} \lambda_i (\varphi_{\mu^{(i)}} (t) - 1) \right) = \exp \left( \sum_{i=1}^{n} \lambda_i \varphi_{\mu^{(i)}} (t) - \lambda \right) = \exp \left( \frac{1}{\lambda} \sum_{i=1}^{n} \lambda_i \varphi_{\mu^{(i)}} (t) - 1 \right)
\]

\[\square\]

We usually focus on the case when \( \mu \) is a discrete probability measure on \( \mathbb{N} = \{1, 2, \ldots\} \). In which case, we usually refer to \( \mu \) by the pmf \( q \) on \( \mathbb{N} \); \( q \) is called the base pmf. Equivalently, \( \mathcal{CP} (\lambda, q) \) is also the distribution of the sum \( \sum_{i \in \mathbb{N}} \Lambda_i \) where \( \Lambda_i : i \in \mathbb{N} \) are independent with \( \Lambda_i \sim \mathcal{P} (\lambda q_i) \). Note that \( \sum_{i \in \mathbb{N}} \lambda q_i = \lambda \). The Poisson distribution is a special case of the compound Poisson distribution where we set \( q \) to be the point mass at 1.

5.34. The compound negative binomial [Bower, Gerber, Hickman, Jones, and Nesbitt, 1982, Ch 11] can be approximated by the compound Poisson distribution.

5.7 Hypergeometric

An urn contains \( N \) white balls and \( M \) black balls. One draws \( n \) balls without replacement, so \( n \leq N + M \). One gets \( X \) white balls and \( n - X \) black balls.

\[
P[X = x] = \begin{cases} \binom{n}{x} \binom{M}{n-x} / \binom{N+M}{n}, & 0 \leq x \leq N \text{ and } 0 \leq n - x \leq M \\ 0, & \text{otherwise} \end{cases}
\]

5.35. The hypergeometric distributions “converge” to the binomial distribution: Assume that \( n \) is fixed, while \( N \) and \( M \) increase to \(+\infty\) with \( \lim_{N,M \to \infty} \frac{N}{N+M} = p \), then

\[
p(x) \to \binom{n}{x} p^x (1-p)^{n-x} \text{ (binomial)}.
\]

Note that binomial is just drawing balls with replacement:

\[
p(x) = \binom{n}{x} N^x M^{n-x} (N+M)^n = \binom{n}{x} \left( \frac{N}{N+M} \right)^x \left( \frac{M}{N+M} \right)^{n-x}.
\]

Intuitively, when \( N \) and \( M \) large, there is not much difference in drawing \( n \) balls with or without replacement.
5.36. Extension: If we have \( m \) colors and \( N_i \) balls of color \( i \). The urn contains \( N = N_1 + \cdots + N_m \) balls. One draws \( n \) balls without replacement. Call \( X_i \) the number of balls of color \( i \) drawn among \( n \) balls. (Of course, \( X_1 + \cdots + X_m = n \).)

\[
P[X_1 = x_1, \ldots, X_m = x_m] = \begin{cases} 
\frac{(N_1)(N_2)\ldots(N_m)}{\binom{N}{x_1}} \cdot \frac{(N_2-x_1)\ldots(N_m-x_1)\ldots(N-n-x_1)}{\binom{n}{x_1}}, & x_1 + \cdots + x_m = n \text{ and } x_i \geq 0 \\
0, & \text{otherwise}
\end{cases}
\]

5.8 Negative Binomial Distribution (Pascal / Pólya distribution)

5.37. The probability that the \( r \)-th success occurs on the \((x + r)\)-th trial

\[
p\left(\begin{array}{c}
x + r - 1 \\
r - 1
\end{array}\right) p^{r-1} (1-p)^x = \left(\begin{array}{c}
x + r - 1 \\
x
\end{array}\right) p^r (1-p)^x
\]

i.e. among the first \((x + r - 1)\) trials, there are \( r - 1 \) successes and \( x \) failures.

- Fix \( r \).
- \( \phi_X(u) = p^r \frac{1}{1-(1-p)e^{iu}} \)
- \( \mathbb{E}X = \frac{rq}{p} \) and \( \text{Var} [X] = \frac{rq}{p^2} \), where \( q = 1 - p \).
- Note that if we define \( \binom{n}{x} \equiv \frac{n(n-1)\ldots(n-(x-1))}{x(x-1)\ldots1} \). Then,

\[
\binom{-r}{x} \equiv (-1)^x \frac{r(r+1)\cdots(r+(x-1))}{x(x-1)\cdots1} = (-1)^x \left(\begin{array}{c}
x + r - 1 \\
x
\end{array}\right).
\]

- If independent \( X_i \sim \text{NegBin}(r_i, p) \), then \( \sum_i X_i \sim \text{NegBin} \left(\sum_i r_i, p\right) \). This is easy to see from the characteristic function.
- When \( r = 1 \), we have geometric distribution. Hence, when \( r \) is a positive integer, negative binomial is an independent sum of i.i.d. geometric.
- \( p(x) = \frac{\Gamma(r+x)}{\Gamma(r)x!} p^r (1-p)^x \)

5.38. A negative binomial distribution can arise as a mixture of Poisson distributions with mean distributed as a gamma distribution \( \Gamma \left( q = r, \lambda = \frac{p}{1-p} \right) \).

Let \( X \) be Poisson with mean \( \lambda \). Suppose, that the mean \( \lambda \) is chosen in accord with a probability distribution \( F_\lambda(\lambda) \). Then, \( \phi_X(u) = \phi_\lambda (-i(e^{iu} - 1)) \) [see compound Poisson distribution]. Here, \( \lambda \sim \Gamma(q, \lambda_0) \); hence, \( \phi_\lambda(u) = \frac{1}{(1-i \lambda_0 u)^q} \).

So, \( \phi_X(u) = \left(1 - e^{iu} - 1\right)^{-q} = \left(\frac{\lambda_0}{\lambda_0 + 1}e^{iu}\right)^q \), which is negative binomial with \( p = \frac{\lambda_0}{\lambda_0 + 1} \).

So, a.k.a. Poisson-gamma distribution, or simply compound Poisson distribution.
5.9 Beta-binomial distribution

A variable with a beta binomial distribution is distributed as a binomial distribution with parameter $p$, where $p$ is distribution with a beta distribution with parameters $\alpha$ and $\beta$.

- $P(k|p) = \binom{n}{k} p^k (1-p)^{n-k}$.
- $f(p) = f_{\beta_{q_1,q_2}}(p) = \frac{\Gamma(q_1+q_2)}{\Gamma(q_1)\Gamma(q_2)} p^{q_1-1} (1-p)^{q_2-1} I_{(0,1)}(p)$
- pmf: $P(k) = \binom{n}{k} \frac{\Gamma(q_1+q_2)}{\Gamma(q_1)\Gamma(q_2)} \frac{\Gamma(k+q_1)\Gamma(n-k+q_2)}{\Gamma(q_1+q_2+n)} = \binom{n}{k} \frac{\beta(k+q_1,n-k+q_2)}{\beta(q_1,q_2)}$
- $\mathbb{E}X = \frac{nq_1}{q_1+q_2}$
- $\text{Var} X = \frac{nq_2}{(q_1+q_2)^2(1+q_1+q_2)}$

5.10 Zipf or zeta random variable

5.39. $P[X = k] = \frac{1}{\xi(p)^p} \frac{1}{k^p}$ where $k \in \mathbb{N}$, $p > 1$, and $\xi$ is the zeta function defined in (A.13).

5.40. $\mathbb{E}[X^n] = \frac{\xi(p-n)}{\xi(p)}$ is finite for $n < p - 1$, and $\mathbb{E}[X^n] = \infty$ for $n \geq p - 1$.

6 PDF Examples

6.1 Uniform Distribution

6.1. Characterization for uniform $[a,b]$:

- $f(x) = \frac{1}{b-a} U(x-a) U(b-x) = \begin{cases} 0 & x < a, x > b \\ \frac{1}{b-a} & a \leq x \leq b \end{cases}$

- $F(x) = \begin{cases} 0 & x < a, x > b \\ \frac{x-a}{b-a} & a \leq x \leq b \end{cases}$

- $\varphi_X(u) = e^{iu(b-a)} \frac{\sin(u(b-a))}{u(b-a)}$

- $M_X(s) = \frac{e^{sb} - e^{sa}}{s(b-a)}$.

6.2. For most purpose, it does not matter whether the value of the density $f$ at the endpoints are 0 or $\frac{1}{b-a}$.

Example 6.3.

- Phase of oscillators $\Rightarrow \left[-\pi, \pi\right]$ or $[0,2\pi]$
- Phase of received signals in incoherent communications $\rightarrow$ usual broadcast carrier phase $\phi \sim U(-\pi,\pi)$
- Mobile cellular communication: multipath $\rightarrow$ path phases $\phi_c \sim U(-\pi,\pi)$
Table 3: Examples of probability density functions. Here, $c, \alpha, q, q_1, q_2, \sigma, \lambda$ are all strictly positive and $d > \frac{1}{2}$. $\gamma = -\psi(1) \approx .5772$ is the Euler-constant. $\psi(z) = \frac{d}{dz} \log \Gamma(z) = (\log e) \frac{\Gamma''(z)}{\Gamma(z)}$ is the digamma function. $B(q_1, q_2) = \frac{\Gamma(q_1)\Gamma(q_2)}{\Gamma(q_1+q_2)}$ is the beta function.
• Use with caution to represent ignorance about a parameter taking value in \([a,b]\).

**6.4.** \(\mathbb{E}X = \frac{a+b}{2} , \text{Var} \, X = \frac{(b-a)^2}{12} , \mathbb{E}[X^2] = \frac{1}{3} (b^2 + ab + a^2)\).

**6.5.** The product \(X\) of two independent \(\mathcal{U}[0, 1]\) has

\[ f_X (x) = - (\ln (x)) 1_{[0,1]} (x) \]

and

\[ F_X (x) = x - x \ln x \]

on \([0,1]\). This comes from \(P [X > x] = \int _0^1 1 - F_U (\frac{z}{t}) \, dt = \int _x^1 \frac{1}{t} \, dt\).

**6.2 Gaussian Distribution**

**6.6.** Gaussian distribution:

(a) Denoted by \(\mathcal{N}(m, \sigma^2)\). \(\mathcal{N}(0,1)\) is the standard Gaussian (normal) distribution.

(b) \(f_X (x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{x-m}{\sigma} \right)^2} \).

(c) \(F_X(x) = \text{normcdf}(x,m,\text{sigma})\).

• The standard normal cdf is sometimes denoted by \(\Phi(x)\). It inherits all properties of cdf. Moreover, note that \(\Phi(-x) = 1 - \Phi(x)\).

(d) \(\varphi_X (v) = \mathbb{E}[e^{jvX}] = e^{jmvm-\frac{1}{2}v^2\sigma^2} \).

(e) \(M_X(s) = e^{sm+\frac{1}{2}s^2\sigma^2} \)

(f) Fourier transform: \(\mathcal{F}\{f_X\} = \int _{-\infty}^{\infty} f_X (x) e^{-j\omega x} \, dx = e^{-j\omega m - \frac{1}{2}\omega^2\sigma^2} \).

(g) \(P [X > x] = P [X \geq x] = Q \left( \frac{x-m}{\sigma} \right) = 1 - \Phi \left( \frac{x-m}{\sigma} \right) = \Phi \left( -\frac{x-m}{\sigma} \right) = P [X < x] = P [X \leq x] = 1 - Q \left( \frac{x-m}{\sigma} \right) = Q \left( -\frac{x-m}{\sigma} \right) = \Phi \left( \frac{x-m}{\sigma} \right) \).

![Figure 14: Probability density function of \(X \sim \mathcal{N}(m, \sigma^2)\).](image)

**6.7.** Properties

62
(a) \[ P[|X - \mu| < \sigma] = 0.6827; \quad P[|X - \mu| > \sigma] = 0.3173 \]
\[ P[|X - \mu| > 2\sigma] = 0.0455; \quad P[|X - \mu| < 2\sigma] = 0.9545 \]

(b) Moments and central moments:

(i) \( \mathbb{E}[(X - \mu)^k] = (k - 1) \mathbb{E}[(X - \mu)^{k-2}] = \begin{cases} 0, & k \text{ odd} \\ 1 \cdot 3 \cdot 5 \cdots (k - 1) \sigma^k, & k \text{ even} \end{cases} \)

(ii) \( \mathbb{E}[|X - \mu|^k] = \begin{cases} 2 \cdot 4 \cdot 6 \cdots (k - 1) \sigma^k \sqrt{\frac{2}{\pi}}, & k \text{ odd} \\ 1 \cdot 3 \cdot 5 \cdots (k - 1) \sigma^k, & k \text{ even} \end{cases} \)

(iii) \( \text{Var}[X^2] = 4\mu^2\sigma^2 + 2\sigma^4 \)

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \mathbb{E}X^n )</th>
<th>( \mu )</th>
<th>( \mu^2 + \sigma^2 )</th>
<th>( \mu(\mu^2 + 3\sigma^2) )</th>
<th>( \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>( \mu )</td>
<td>( \sigma^2 )</td>
<td>0</td>
<td>( 3\sigma^4 )</td>
</tr>
</tbody>
</table>

(c) For \( \mathcal{N}(0, 1) \) and \( k \geq 1 \),

\[ \mathbb{E}[X^k] = (k - 1) \mathbb{E}[X^{k-2}] = \begin{cases} 0, & k \text{ odd} \\ 1 \cdot 3 \cdot 5 \cdots (k - 1), & k \text{ even} \end{cases} \]

The first equality comes from integration by parts. Observe also that

\[ \mathbb{E}[X^{2m}] = \frac{(2m)!}{2^m m!}. \]

(d) Lévy–Cramér theorem: If the sum of two independent non-constant random variables is normally distributed, then each of the summands is normally distributed.

- Note that \( \int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}}. \)

6.8 (Length bound). For \( X \sim \mathcal{N}(0, 1) \) and any (Borel) set \( B \),

\[ P[X \in B] \leq \int_{|B|/2}^{\infty} f_X(x) = 1 - 2Q\left(\frac{|B|}{2}\right), \]

where \( |B| \) is the length (Lebesgue measure) of the set \( B \). This is because the probability is concentrated around 0. More generally, for \( X \sim \mathcal{N}(m, \sigma^2) \)

\[ P[X \in B] \leq 1 - 2Q\left(\frac{|B|}{2\sigma}\right). \]

6.9 (Stein’s Lemma). Let \( X \sim \mathcal{N}(\mu, \sigma^2) \), and let \( g \) be a differentiable function satisfying \( \mathbb{E}|g'(X)| < \infty \). Then

\[ \mathbb{E}[g(X)(X - \mu)] = \sigma^2 \mathbb{E}[g'(X)]. \]

[2] Lemma 3.6.5 p 124. Note that this is simply integration by parts with \( u = g(x) \) and \( dv = (x - \mu)f_X(x)dx \).
• $E[(X - \mu)^k] = E[(X - \mu)^{k-1}(X - \mu)] = \sigma^2(k - 1)E[(X - \mu)^{k-2}]$.

6.10. **Q-function**: $Q(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ corresponds to $P[X > z]$ where $X \sim \mathcal{N}(0, 1)$; that is $Q(z)$ is the probability of the “tail” of $\mathcal{N}(0, 1)$. The $Q$ function is then a complementary cdf (ccdf).

![Figure 15: Q-function](image)

(a) $Q$ is a decreasing function with $Q(0) = \frac{1}{2}$.

(b) $Q(-z) = 1 - Q(z) = \Phi(z)$

(c) $Q^{-1}(1 - Q(z)) = -z$

(d) Craig’s formula: $Q(x) = \frac{1}{\pi} \int_{0}^{\pi} e^{-\frac{x^2}{2 \sin^2 \theta}} d\theta = \frac{1}{\pi} \int_{0}^{\pi} e^{-\frac{x^2}{2 \cos^2 \theta}} d\theta$, $x \geq 0$.

To see this, consider $X, Y \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. Then,

$$Q(z) = \int \int_{(x,y) \in (z, \infty) \times \mathbb{R}} f_{X,Y}(x,y) dx dy = 2 \int \int_{0}^{\pi} f_{X,Y}(r \cos \theta, r \sin \theta) dr d\theta.$$ 

where we evaluate the double integral using polar coordinates [9, Q7.22 p 322].

(e) $Q^2(x) = \frac{1}{\pi} \int_{0}^{\pi} e^{-\frac{x^2}{2 \sin^2 \theta}} d\theta$

(f) $\frac{d}{dx} Q(x) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

(g) $\frac{d}{dx} Q(f(x)) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{(f(x))^2}{2}} \frac{df}{dx} f(x)$

(h) $\int Q(f(x)) g(x) dx = Q(f(x)) \int g(x) dx + \int \frac{1}{\sqrt{2\pi}} e^{-\frac{(f(x))^2}{2}} \left( \frac{df}{dx} f(x) \right) \left( \int_{a}^{x} g(t) dt \right) dx$
(i) \( P[X > x] = Q \left( \frac{x-m}{\sigma} \right) \)
\( P[X < x] = 1 - Q \left( \frac{x-m}{\sigma} \right) = Q \left( -\frac{x-m}{\sigma} \right) \).

(j) Approximation:

1. \( Q(z) \approx \left[ \frac{1}{1-\frac{2}{\pi z + a + \sqrt{b+z^2}}} \right] \frac{1}{\sqrt{2\pi}} e^{-z^2/2}; \quad a = \frac{1}{\pi}, \quad b = 2\pi 
2. \( (1 - \frac{1}{x^2}) \frac{e^{-x^2}}{x\sqrt{2\pi}} \leq Q(x) \leq \frac{1}{2} e^{-\frac{x^2}{2}} \)
3. \( Q(z) \approx \frac{1}{\sqrt{2\pi}} \left( 1 - \frac{0.7}{z^2} \right) e^{-\frac{z^2}{2}}; \quad z > 2 \)

6.11. Error function \( \text{(MATLAB)}: \text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx = 1 - 2Q(\sqrt{2}z) \)

(a) It is an odd function of \( z \).

(b) For \( z \geq 0 \), it corresponds to \( P[|X| < z] \) where \( X \sim \mathcal{N}(0, \frac{1}{2}) \).

(c) \( \lim_{z \to \infty} \text{erf}(z) = 1 \)

(d) \( \text{erf}(-z) = -\text{erf}(z) \)

(e) \( Q(z) = \frac{1}{2} \text{erfc} \left( \frac{z}{\sqrt{2}} \right) = \frac{1}{2} \left( 1 - \text{erf} \left( \frac{z}{\sqrt{2}} \right) \right) \)

(f) \( \Phi(x) = \frac{1}{2} \left( 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right) = \frac{1}{2} \text{erfc} \left( -\frac{x}{\sqrt{2}} \right) \)

(g) \( Q^{-1}(q) = \sqrt{2} \text{erfc}^{-1}(2q) \)

(h) The complementary error function: \( \text{erfc}(z) = 1 - \text{erf}(z) = 2Q(\sqrt{2}z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-x^2} dx \)

\[ \mathcal{N}(0, \frac{1}{2}) \]

Figure 16: erf-function and Q-function
6.3 Exponential Distribution

6.12. Denoted by $E(\lambda)$. It is in fact $\Gamma(1, \lambda)$. $\lambda > 0$ is a parameter of the distribution, often called the rate parameter.

6.13. Characterized by

- $f_X(x) = \lambda e^{-\lambda x} U(x)$;
- $F_X(x) = (1 - e^{-\lambda x}) U(x)$;
- Survival-, survivor-, or reliability-function: $P[X > x] = e^{-\lambda x} 1_{[0, \infty)}(x) + 1_{(-\infty,0)}(x)$;
- $\varphi_X(u) = \frac{\lambda}{\lambda - iu}$.
- $M_X(s) = \frac{\lambda}{\lambda - s}$ for $\text{Re}\{s\} < \lambda$.

6.14. $\mathbb{E}X = \sigma_X = \frac{1}{\lambda}$, $\text{Var}\{X\} = \frac{1}{\lambda^2}$.

6.15. $\text{median}(X) = \frac{1}{\lambda} \ln 2$, $\text{mode}(X) = 0$, $\mathbb{E}[X^n] = \frac{n!}{\lambda^n}$.

6.16. Coefficient of variation: $\text{CV} = \frac{\sigma_X}{\mathbb{E}X} = 1$

6.17. It is a continuous version of geometric distribution. In fact, $[X] \sim G_0(e^{-\lambda})$ and $[X] \sim G_1(e^{-\lambda})$

6.18. $X \sim E(\lambda)$ is simply $\frac{1}{\lambda} X_1$ where $X_1 \sim E(1)$.

6.19. Suppose $X_1 \sim E(1)$. Then $\mathbb{E}[X^n] = n!$ for $n \in \mathbb{N} \cup \{0\}$. In general, for $X \sim E(\lambda)$, we have $\mathbb{E}[X^n] = \frac{1}{\lambda^n} \Gamma(\alpha + 1)$ for any $\alpha > -1$. In particular, for $n \in \mathbb{N} \cup \{0\}$, the moment $\mathbb{E}[X^n] = \frac{n!}{\lambda^n}$.

6.20. $\mu_3 = \mathbb{E}[(X - \mathbb{E}X)^3] = \frac{3}{\lambda^3}$ and $\mu_4 = \frac{9}{\lambda^4}$.

6.21. Hazard function: $\frac{f_X(x)}{P[X > x]} = \lambda$.

6.22. $h(X) = \log \frac{\lambda}{X}$.

6.23. Can be generated by $X = -\frac{1}{\lambda} \ln U$ where $U \sim U(0,1)$.

6.24. MATLAB:

- $X = \text{exprnd}(1/\text{lambda})$
- $f_X(x) = \text{exppdf}(x,1/\text{lambda})$
- $F_X(x) = \text{expcdf}(x,1/\text{lambda})$

6.25. Memoryless property: The exponential r.v. is the only continuous r.v. on $[0, \infty)$ that satisfies the memoryless property:

$$\mathbb{P}[X > s + x \mid X > s] = \mathbb{P}[X > x]$$

for all $x > 0$ and all $s > 0$ [13 p. 157–159]. In words, the future is independent of the past. The fact that it hasn’t happened yet, tells us nothing about how much longer it will take before it does happen.
• In particular, suppose we define the set \( B + x \) to be \( \{ x + b : b \in B \} \). For any \( x > 0 \) and set \( B \subset [0, \infty) \), we have

\[
P[X \in B + x | X > x] = P[X \in B]
\]

because

\[
\frac{P[X \in B + x]}{P[X > x]} = \frac{\int_{B+x} \lambda e^{-\lambda t} dt}{e^{-\lambda x}} = \frac{\int_B \lambda e^{-\lambda(t+x)} dt}{e^{-\lambda x}}.
\]

6.26. The difference of two independent \( E(\lambda) \) is \( L(\lambda) \). In particular, suppose \( X \) and \( Y \) are i.i.d. \( E(\lambda) \), then \( \mathbb{E}[X - Y] = \sigma_X = \frac{1}{\lambda} \) and \( \sqrt{\mathbb{E}[(X - Y)^2]} = \frac{\sqrt{2}}{\lambda} \).

6.27. Consider independent \( X_i \sim E(\lambda) \). Let \( S_n = \sum_{i=1}^{n} X_i \).

(a) \( S_n \sim \Gamma(n, \lambda) \), i.e. its has \( n \)-Erlang distribution.

(b) Let \( N = \inf \{ n : S_n \geq s \} \). Then, \( N \sim \Lambda + 1 \) where \( \Lambda \sim \mathcal{P}(\lambda s) \).

6.28. If independent \( X_i \sim E(\lambda_i) \), then

(a) \( \min_i X_i \sim E\left(\sum_i \lambda_i\right) \).

Recall order statistics. Let \( Y_1 = \min_i X_i \)

(b) \( P\left[ \min_i X_i = j \right] = \frac{\lambda_j}{\sum_i \lambda_i} \). Note that this is \( \int_0^\infty f_{X_j}(t) \prod_{i \neq j} P[X_i > t] dt \).

6.29. If \( S_i, T_i \sim \text{i.i.d.} E(\alpha) \), then

\[
P\left( \sum_{i=1}^{m} S_i > \sum_{j=1}^{n} T_j \right) = \sum_{i=0}^{m-1} \binom{n + m - 1}{i} \left( \frac{1}{2} \right)^{n+m-1} \]

\[
= \sum_{i=0}^{m-1} \binom{n + i - 1}{i} \left( \frac{1}{2} \right)^{n+i}.
\]

Note that we can set up two Poisson processes. Consider the superposed process. We want the \( n \)-th arrival from the \( T \) processes to come before the \( m \)-th one of the \( S \) process.

6.4 Pareto: \( \text{Par}(\alpha) \)-heavy-tailed model/density

6.30. Characterizations: Fix \( \alpha > 0 \).

(a) \( f(x) = \alpha x^{-\alpha-1} U(x - 1) \)

(b) \( F(x) = \left( 1 - \frac{1}{x^\alpha} \right) U(x - 1) = \begin{cases} 
0 & x < 1 \\
1 - \frac{1}{x^\alpha} & x \geq 1 
\end{cases} \)

Example 6.31.
• distribution of wealth
• flood heights of the Nile river
• designing dam height
• (discrete) sizes of files requested by web users
• waiting times between successive keystrokes at computer terminals
• (discrete) sizes of files stored on Unix system file servers
• running times for NP-hard problems as a function of certain parameters

6.5 Laplacian: \( L(\alpha) \)

6.32. Characterization: \( \alpha > 0 \)
(a) Also known as Laplace or double exponential.
(b) \( f(x) = \frac{\alpha}{2}e^{-\alpha|x|} \)
(c) \( F(x) = \begin{cases} \frac{1}{2}e^{\alpha x} & x < 0 \\ 1 - \frac{1}{2}e^{-\alpha x} & x \geq 0 \end{cases} \)
(d) \( \varphi_X(u) = \frac{\alpha^2}{\alpha^2 + u^2} \)
(e) \( M_X(s) = \frac{\lambda^2}{\lambda^2 - s^2}, -\lambda < \text{Re}\{s\} < \lambda. \)

6.33. \( \mathbb{E}X = 0, \text{Var } X = \frac{2}{\alpha^2}, \mathbb{E}|X| = \frac{1}{\alpha}. \)

Example 6.34.
• amplitudes of speech signals
• amplitudes of differences of intensities between adjacent pixels in an image
• If \( X \) and \( Y \) are independent \( \mathcal{E}(\lambda) \), then \( X - Y \) is \( L(\lambda) \). (Easy proof via ch.f.)

6.6 Rayleigh

6.35. Characterizations:
(a) \( F(x) = \left(1 - e^{-\alpha x^2}\right)U(x) \)
(b) \( f(x) = 2\alpha xe^{-\alpha x^2}u(x) \)
(c) \( P[X > t] = 1 - F(t) = \begin{cases} e^{-\alpha t^2}, & t \geq 0 \\ 1, & t < 0 \end{cases} \)
(d) Use \( \sqrt{-2\sigma^2 \ln U} \) to generate Rayleigh \( \left(\frac{1}{\sigma^2}\right) \) from \( U \sim \mathcal{U}(0,1). \)
Example 6.37.

- Noise $X$ at the output of AM envelope detector when no signal is present

6.38. Relationship with other distributions

(a) Let $X$ be a Rayleigh$(\alpha)$ r.v., then $Y = X^2$ is $\mathcal{E}(\alpha)$. Hence,

\[
\mathcal{E}(\alpha) \sim \frac{\sqrt{\alpha^2}}{\alpha^2} \text{ Rayleigh}(\alpha).
\] (18)

(b) Suppose $X, Y \overset{i.i.d.}{\sim} N(0, \sigma^2)$. $R = \sqrt{X^2 + Y^2}$ has a Rayleigh distribution with density

\[
f_R(r) = 2r \frac{1}{2\sigma^2} e^{-\frac{1}{2\sigma^2} r^2}.
\] (19)

- Note that $X^2, Y^2 \overset{i.i.d.}{\sim} \Gamma\left(\frac{1}{2}, \frac{1}{2\sigma^2}\right)$. Hence, $X^2 + Y^2 \sim \Gamma\left(1, \alpha = \frac{1}{2\sigma^2}\right)$, exponential. By (18), $\sqrt{X^2 + Y^2}$ is a Rayleigh r.v. $\alpha = \frac{1}{2\sigma^2}$.

- Alternatively, the transformation from Cartesian coordinates $(x, y)$ to polar coordinates $(r, \theta)$ gives

\[
f_{R,\Theta}(r, \theta) = r f_{X,Y}(r \cos \theta, r \sin \theta) = r \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} \left(\frac{r \cos \theta}{\sigma}\right)^2} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} \left(\frac{r \sin \theta}{\sigma}\right)^2}
\]

\[
= \left(\frac{1}{2\pi}\right) \left(2r \frac{1}{2\sigma^2} e^{-\frac{1}{2\sigma^2} r^2}\right).
\]

Hence, the radius $R$ and the angle $\Theta$ are independent, with the radius $R$ having a Rayleigh distribution while the angle $\Theta$ is uniformly distributed in the interval $(0, 2\pi)$.

6.7 Cauchy

6.39. Characterizations: Fix $\alpha > 0$.

(a) $f_X(x) = \frac{\alpha}{\pi} \frac{1}{\alpha^2 + x^2}, \quad \alpha > 0$

(b) $F_X(x) = \frac{1}{\pi} \tan^{-1} \left(\frac{x}{\alpha}\right) + \frac{1}{2}$

(c) $\varphi_X(u) = e^{-\alpha |u|}$

Note that $\text{Cau}(\alpha)$ is simply $\alpha X_1$ where $X_1 \sim \text{Cau}(1)$. Also, because $f_X$ is even, $f_{-X} = f_X$ and thus $-X$ is still $\sim \text{Cau}(\alpha)$.

6.40. Odd moments are not defined; even moments are infinite.

- Because the first moment is not defined, central moments, including the variance, are not defined.
• Mean and variance do not exist.
• Note that even though the pdf of Cauchy distribution is an even function, the mean is not 0.

6.41. Suppose $X_i$ are independent Cau$(\alpha_i)$. Then, $\sum a_i X_i \sim \text{Cau}(|a_i| \alpha_i)$.

6.42. Suppose $X \sim \text{Cau}(\alpha)$. Then, $X \sim \text{Cau}(1, \alpha)$.

6.8 More PDFs

6.43. Beta
(a) $f_{\beta_{q_1,q_2}}(x) = \frac{\Gamma(q_1+q_2)}{\Gamma(q_1)\Gamma(q_2)} x^{q_1-1} (1-x)^{q_2-1} 1_{(0,1)}(x), \ x \in (0,1)$
(b) MATLAB: $X = \text{betarnd}(q_1,q_2), f_X(x) = \text{betapdf}(x,q_1,q_2)$, and $F_X(x) = \text{betacdf}(x,q_1,q_2)$
(c) For symmetric distributions $q_1$ and $q_2$ are the same. As their values increase the distributions become more peaked.
(d) The uniform distribution has $q_1 = q_2 = 1$.
(e) $E[X^i (1-X)^j] = \frac{\Gamma(q_1+q_2)}{\Gamma(q_1+q_2+i+j)} \frac{\Gamma(q_1+i)}{\Gamma(q_1)} \frac{\Gamma(q_2+j)}{\Gamma(q_2)}$
(f) $E[X] = \frac{q_1}{q_1+q_2}$ and $\text{Var} X = \frac{q_1 q_2}{(q_1+q_2)^2(q_1+q_2+1)}$.
(g) Parameters for random variable with mean $m$ and variance $\sigma^2$:

(i) Suppose we want $E[X] = m$ and $\text{Var} X = \sigma^2$, then we need $q_1 = m \left(\frac{(1-m)m}{\sigma^2} - 1\right)$ and $q_2 = (1-m) \left(\frac{(1-m)m}{\sigma^2} - 1\right)$
(ii) The support of $Y = aX$ is $[0,a]$. Suppose we want $E[Y] = m$ and $\text{Var} Y = \sigma^2$, then we set $q_1 = m \left(\frac{(a-m)m}{a^2} - 1\right)$ and $q_2 = (1 - m) \left(\frac{(a-m)m}{a^2} - 1\right)$

6.44. Rice/Rician
(a) Characterizations: fix $v \geq 0$ and $\sigma > 0$,

$$f_R(r) = \frac{r}{\sigma^2} \exp \left(-\frac{r^2 + v^2}{2\sigma^2}\right) I_0 \left(\frac{rv}{\sigma^2}\right) 1[r > 0]$$

where $I_0$ is the modified Bessel function\textsuperscript{4} of the first kind with order zero.
(b) When $v = 0$, the distribution reduces to a Rayleigh distribution given in (19).

\textsuperscript{4} $I_0(z) = \frac{1}{\pi} \int_0^\pi e^{z \cos \theta} d\theta$. In MATLAB, use $\text{besseli}(0,z)$ for $I_0(z)$. 

70
(c) Suppose we have independent \( N_i \sim \mathcal{N}(m_i, \sigma^2) \), then \( R = \sqrt{N_1^2 + N_2^2} \) is Rician with \( v = \sqrt{m_1^2 + m_2^2} \) and the same \( \sigma \). To see this, use (30) and the fact that we can express \( m_1 = v \cos \phi \) and \( m_2 = v \sin \phi \) for some \( \phi \).

(d) \( \mathbb{E}[R^2] = 2\sigma^2 + v^2, \mathbb{E}[R^4] = 8\sigma^4 + 8\sigma^2 v^2 + v^4 \).

### 6.45 Weibull:
For \( \lambda > 0 \) and \( p > 0 \), the Weibull(\( p, \lambda \)) distribution \(^9\) is characterized by

(a) \( X = (\frac{Y}{\lambda})^\frac{1}{p} \) where \( Y \sim \mathcal{E}(1) \).

(b) \( f_X(x) = \lambda px^{p-1} e^{-\lambda x^p}, x > 0 \)

(c) \( F_X(x) = 1 - e^{-\lambda x^p}, x > 0 \)

(d) \( \mathbb{E}[X^n] = \frac{\Gamma(1 + \frac{n}{p})}{\lambda^p} \).

### 7 Expectation

Consider probability space \((\Omega, \mathcal{A}, P)\)

7.1. Let \( X^+ = \max(X, 0) \), and \( X^- = -\min(X, 0) = \max(-X, 0) \). Then, \( X = X^+ - X^- \), and \( X^+, X^- \) are nonnegative r.v.’s. Also, \( |X| = X^+ + X^- \)

7.2. A random variable \( X \) is integrable if and only if

\( \equiv X \) has a finite expectation
\( \equiv \) both \( \mathbb{E}[X^+] \) and \( \mathbb{E}[X^-] \) are finite
\( \equiv \mathbb{E}|X| \) is finite.
\( \equiv \mathbb{E}X \) is finite \( \equiv \mathbb{E}X \) is defined.
\( \equiv X \in \mathcal{L} \)
\( \equiv |X| \in \mathcal{L} \)

In which case,

\[
\mathbb{E}X = \mathbb{E}[X^+] - \mathbb{E}[X^-] = \int X(\omega)P(d\omega) = \int XdP = \int xdP^X(x) = \int xP^X(dx)
\]

and

\[
\int_A XdP = \mathbb{E}[1_A X].
\]
Definition 7.3. A r.v. $X$ admits (has) an expectation if $\mathbb{E}[X^+]$ and $\mathbb{E}[X^-]$ are not both equal to $+\infty$. Then, the expectation of $X$ is still given by $\mathbb{E}X = \mathbb{E}[X^+] - \mathbb{E}[X^-]$ with the conventions $+\infty + a = +\infty$ and $-\infty + a = -\infty$ when $a \in \mathbb{R}$.

7.4. $L^1 = L^1(\Omega, \mathcal{A}, P)$ = the set of all integrable random variables.

7.5. For $1 \leq p < \infty$, the following are equivalent:
   
   (a) $(\mathbb{E}[X^p])^{\frac{1}{p}} = 0$
   
   (b) $\mathbb{E}[X^p] = 0$
   
   (c) $X = 0$ a.s.

7.6. $X \overset{a.s.}{=} Y \Rightarrow \mathbb{E}X = \mathbb{E}Y$

7.7. $\mathbb{E}[1_{B}(X)] = P(X^{-1}(B)) = P^X(B) = P[X \in B]$

   * $F_X(x) = \mathbb{E}[1_{(-\infty,x)}(X)]$

7.8. Expectation rule: Let $X$ be a r.v. on $(\Omega, \mathcal{A}, P)$, with values in $(E, \mathcal{E})$, and distribution $P^X$. Let $h : (E, \mathcal{E}) \rightarrow (\mathbb{R}, \mathcal{B})$ be measurable. If

   * $X \geq 0$ or
   * $h(X) \in L^1(\Omega, \mathcal{A}, P)$ which is equivalent to $h \in L^1(E, \mathcal{E}, P^X)$

then

   * $\mathbb{E}[h(X)] = \int h(X(\omega)) \, P(d\omega) = \int h(x) \, P^X(dx)$
   
   * $\int_{[X \in G]} h(X(\omega)) \, P(d\omega) = \int_{G} h(x) \, P^X(dx)$

7.9. Expectation of an absolutely continuous random variable: Suppose $X$ has density $f_X$, then $h$ is $P^X$-integrable if and only if $h \cdot f_X$ is integrable w.r.t. Lebesgue measure. In which case,

   $\mathbb{E}[h(X)] = \int h(x) \, P^X(dx) = \int h(x) \, f_X(x) \, dx$

   and

   $\int_{G} h(x) \, P^X(dx) = \int_{G} h(x) \, f_X(x) \, dx$

   * Caution: Suppose $h$ is an odd function and $f_X$ is an even function, we can not conclude that $\mathbb{E}[h(X)] = 0$. One obvious odd-function $h$ is $h(x) = x$. For example, in (6.40), when $X$ is Cauchy, the expectation does not exist even though the pdf is an even function. Of course, in general, if we also know that $h(X)$ is integrable, then $\mathbb{E}[h(X)]$ is 0.
Expectation of a discrete random variable: Suppose $x$ is a discrete random variable.

$$\mathbb{E}X = \sum xP[X = x]$$

and

$$\mathbb{E}[g(X)] = \sum g(x)P[X = x].$$

Similarly,

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

These are called the law/rule of the lazy statistician (LOTUS) [23, Thm 3.6 p 48],[9, p. 149].

7.10. $\int E P[X \geq t] dt = \int E P[X > t] dt$ and $\int E P[X \leq t] dt = \int E P[X < t] dt$

7.11. **Expectation and cdf**:

(a) For nonnegative $X$,

$$\mathbb{E}X = \int_{0}^{\infty} P[X > y] dy = \int_{0}^{\infty} (1 - F_X(y)) dy = \int_{0}^{\infty} P[X \geq y] dy$$

For $p > 0$,

$$\mathbb{E}[X^p] = \int_{0}^{\infty} px^{p-1} P[X > x]dx.$$

(b) For integrable $X$,

$$\mathbb{E}X = \int_{0}^{\infty} (1 - F_X(x)) dx - \int_{-\infty}^{0} F_X(x) dx$$

(c) For nonnegative integer-valued $X$, $\mathbb{E}X = \sum_{k=0}^{\infty} P[X > k] = \sum_{k=1}^{\infty} P[X \geq k]$.

**Definition 7.12.**

(a) **Absolute moment**: $\mathbb{E}\left[|X|^k\right] = \int |x|^k P^X(dx)$, where we define $\mathbb{E}\left[|X|^0\right] = 1$

(b) **Moment**: If $\mathbb{E}\left[|X|^k\right] < \infty$, then $m_k = \mathbb{E}[X^k] = \int x^k P^X(dx)$ is the $k^{th}$ moment of $X$

(c) **Variance**: If $\mathbb{E}[X^2] < \infty$, then we define

$$\text{Var}X = \mathbb{E}\left[(X - \mathbb{E}X)^2\right] = \int (x - \mathbb{E}X)^2 P^X(dx) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

$$= \mathbb{E}[X(X - \mathbb{E}X)]$$
• Notation: $D_X$, or $\sigma^2(X)$, or $\sigma^2_X$, or $\nabla X \ [23\ p.\ 51]$

• If $X_i \in L^2$, then $\sum_{i=1}^k X_i \in L^2$, $\text{Var} \left[ \sum_{i=1}^k X_i \right]$ exists, and $\mathbb{E} \left[ \sum_{i=1}^k X_i \right] = \sum_{i=1}^k \mathbb{E}X_i$

• Suppose $\mathbb{E}[X^2] < \infty$. If $\text{Var} X = 0$, then $X = \mathbb{E}X$ a.s.

(d) **Standard Deviation**: $\sigma_X = \sqrt{\text{Var}[X]}$.

(e) **Coefficient of Variation**: $CV_X = \frac{\sigma_X}{\mathbb{E}X}$.

• It is the standard deviation of the “normalized” random variable $\frac{X}{\mathbb{E}X}$.

• 1 for exponential.

(f) **Fano Factor** (index of dispersion): $\frac{\text{Var}X}{\mathbb{E}X}$.

• 1 for Poisson.

(g) **Central Moments**: the $n$th central moment is $\mu_n = \mathbb{E}[(X - \mathbb{E}X)^n]$.

(i) $\mu_1 = \mathbb{E}[X - \mathbb{E}X] = 0$.

(ii) $\mu_2 = \sigma^2_X = \text{Var} X$

(iii) $\mu_n = \sum_{k=1}^n \left( \begin{array}{c} n \\ k \end{array} \right) m_{n-k} (-m_1)^k$

(iv) $m_n = \sum_{k=1}^n \left( \begin{array}{c} n \\ k \end{array} \right) \mu_{n-k} m_1^k$

(h) Skewness coefficient: $\gamma_X = \frac{\mu_3}{\sigma_X}$

(i) Describe the deviation of the distribution from a symmetric shape (around the mean)

(ii) 0 for any symmetric distribution

(i) Kurtosis: $\kappa_X = \frac{\mu_4}{\sigma_X^4}$.

• $\kappa_X = 3$ for Gaussian distribution

(j) **Excess coefficient**: $\varepsilon_X = \kappa_X - 3 = \frac{\mu_4}{\sigma_X^4} - 3$

(k) **Cumulants** or semivariants: For one variable: $\gamma_k = \frac{1}{j^k} \frac{\partial^k}{\partial v^k} \ln (\varphi_X(v)) \bigg|_{v=0}$.

(i) $\gamma_1 = \mathbb{E}X = m_1$, $\gamma_2 = \mathbb{E}[X - \mathbb{E}X]^2 = m_2 - m_1^2 = \mu_2$

(ii) $m_1 = \gamma_1$

$\gamma_2 = m_2 + \gamma_1^2$

$\gamma_3 = m_3 + 3\gamma_1 \gamma_2 + \gamma_1^3$

$\gamma_4 = m_4 + 3m_2^2 - 4m_1 m_3 + 12m_1^2 m_2 - 6m_1^4 = \mu_4 - 3\mu_2^2$
TABLE 2.1 Product-Moments of Random Variables

<table>
<thead>
<tr>
<th>Moment Measure of De</th>
<th>Continuous variable</th>
<th>Discrete variable</th>
<th>Sample estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Location</td>
<td>$E(X) = \mu_x$</td>
<td>$\mu_x = \int_{-\infty}^{\infty} xf_X(x) , dx$</td>
<td>$\bar{x} = \sum x_i/n$</td>
</tr>
<tr>
<td>Second Dispersion</td>
<td>$\text{Var}(X) = \mu_2 = \sigma^2$</td>
<td>$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x) , dx$</td>
<td>$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$</td>
</tr>
<tr>
<td>Coefficient of variation, $\Omega_2$</td>
<td>$\Omega_2 = \sigma_2/\mu_2$</td>
<td>$\Omega_2 = \sigma_2/\mu_2$</td>
<td>$C_v = s/\bar{x}$</td>
</tr>
<tr>
<td>Third Asymmetry</td>
<td>Skewness</td>
<td>$\gamma_x = \mu_3/\sigma_3^3$</td>
<td>$g = m_3/\sigma^3$</td>
</tr>
<tr>
<td>Fourth Peakedness</td>
<td>Kurtosis, $\kappa_x$</td>
<td>$\kappa_x = \mu_4/\sigma_4^4$</td>
<td>$k = m_4/\sigma^4$</td>
</tr>
<tr>
<td></td>
<td>Excess coefficient, $\epsilon_x$</td>
<td>$\epsilon_x = \kappa_x - 3$</td>
<td>$\epsilon_x = \kappa_x - 3$</td>
</tr>
</tbody>
</table>

Figure 17: Product-Moments of Random Variables [22]

Table 4: Expectations and Variances

<table>
<thead>
<tr>
<th>Model</th>
<th>$E[X]$</th>
<th>$\text{Var}[X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{U}(0,\ldots,n-1)$</td>
<td>$\frac{n-1}{2}$</td>
<td>$\frac{n^2-1}{12}$</td>
</tr>
<tr>
<td>$\mathcal{B}(n, p)$</td>
<td>$np$</td>
<td>$np(1-p)$</td>
</tr>
<tr>
<td>$\mathcal{G}(\beta)$</td>
<td>$1-\beta$</td>
<td>$\beta (1-\beta)^2$</td>
</tr>
<tr>
<td>$\mathcal{P}(\lambda)$</td>
<td>$\lambda$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>$\mathcal{U}(a, b)$</td>
<td>$\frac{a+b}{2}$</td>
<td>$\frac{(b-a)^2}{12}$</td>
</tr>
<tr>
<td>$\mathcal{E}(\lambda)$</td>
<td>$\frac{1}{\lambda}$</td>
<td>$\frac{1}{\lambda^2}$</td>
</tr>
<tr>
<td>$\text{Par}(\alpha)$</td>
<td>$\left{ \begin{array}{ll} \frac{\alpha}{\alpha-1}, &amp; \alpha &gt; 1 \ \infty, &amp; 0 &lt; \alpha \leq 1 \end{array} \right.$</td>
<td>$\left{ \begin{array}{ll} \text{undefined,} &amp; 0 &lt; \alpha &lt; 1 \ \infty, &amp; 1 &lt; \alpha &lt; 2 \end{array} \right.$</td>
</tr>
<tr>
<td>$\mathcal{L}(\alpha)$</td>
<td>$0$</td>
<td>$\frac{1}{\alpha^2}$</td>
</tr>
<tr>
<td>$\mathcal{N}(m, \sigma^2)$</td>
<td>$m$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>$\mathcal{N}(m, \Lambda)$</td>
<td>$m$</td>
<td>$\Lambda = \text{Cov}[X_1, X_2]$</td>
</tr>
<tr>
<td>$\Gamma(q, \lambda)$</td>
<td>$\frac{q}{\lambda}$</td>
<td>$\frac{q}{\lambda^2}$</td>
</tr>
</tbody>
</table>

Table 4: Expectations and Variances

7.13.

- For $c \in \mathbb{R}$, $E[c] = c$
- $E[\cdot]$ is a linear operator: $E[aX + bY] = aE[X] + bE[Y]$
- In general, $\text{Var}[\cdot]$ is not a linear operator.

7.14. All pairs of mean and variance are possible. A random variable $X$ with $E[X] = m$ and $\text{Var}[X] = \sigma^2$ can be constructed by setting $P[X = m - a] = P[X = m + a] = \frac{1}{2}$.

Definition 7.15.

- Correlation between $X$ and $Y$: $E[XY]$. 
• Covariance between $X$ and $Y$:
\[
\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y
\]
\[
= \mathbb{E}[X(Y - \mathbb{E}Y)] = \mathbb{E}[Y(X - \mathbb{E}X)].
\]

• $X$ and $Y$ are said to be **uncorrelated** if and only if $\text{Cov}[X, Y] = 0$.
\[
\equiv \mathbb{E}[XY] = \mathbb{E}X\mathbb{E}Y
\]

• $X$ and $Y$ are said to be **orthogonal** if $\mathbb{E}[XY] = 0$.

• **Correlation coefficient, autocorrelation, normalized covariance**:
\[
\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y} = \mathbb{E} \left[ \left( \frac{X - \mathbb{E}X}{\sigma_X} \right) \left( \frac{Y - \mathbb{E}Y}{\sigma_Y} \right) \right] = \frac{\mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y}{\sigma_X \sigma_Y}.
\]

7.16. Properties

(a) $\text{Var} X = \text{Cov}[X, X]$, $\rho_{X,X} = 1$

(b) $\text{Var}[aX] = a^2 \text{Var} X$, $\sigma_{aX} = |a| \sigma_X$.

(c) If $X \perp Y$, then $\text{Cov}[X, Y] = 0$. The converse is not true.

(d) $\rho_{XY} \in [-1, 1]$.

(e) By Caychy-Schwartz inequality, $(\text{Cov}[X, Y])^2 \leq \sigma_X^2 \sigma_Y^2$ with equality if and only if $\sigma_Y^2 (X - \mathbb{E}X)^2 = \sigma_X^2 (Y - \mathbb{E}Y)^2$ a.s.

• This implies $|\rho_{X,Y}| \leq 1$.

When $\sigma_Y, \sigma_X > 0$, equality occurs if and only if the following conditions holds
\[
\equiv \exists a \neq 0 \text{ such that } (X - \mathbb{E}X) = a(Y - \mathbb{E}Y)
\]
\[
\equiv \exists c \neq 0 \text{ and } b \in \mathbb{R} \text{ such that } Y = cX + b
\]
\[
\equiv \exists a \neq 0 \text{ and } b \in \mathbb{R} \text{ such that } X = aY + b
\]
\[
\equiv |\rho_{XY}| = 1
\]

In which case, $|a| = \frac{\sigma_X}{\sigma_Y}$ and $\rho_{XY} = \frac{a}{|a|} = \text{sgn } a$. Hence, $\rho_{XY}$ is used to quantify linear dependence between $X$ and $Y$. The closer $|\rho_{XY}|$ to 1, the higher degree of linear dependence between $X$ and $Y$.

(f) Linearity:

(i) Let $Y_i = a_i X_i + b_i$.

i. $\text{Cov}[Y_1, Y_2] = \text{Cov}[a_1 X_1 + b_1, a_2 X_2 + b_2] = a_1 a_2 \text{Cov}[X_1, X_2]$.

ii. The $\rho$ is preserved under linear transformation:
\[
\rho_{Y_1,Y_2} = \rho_{X_1,X_2}.
\]
(ii) \( \text{Cov} [a_1X + b_1, a_2X + b_2] = a_1a_2 \text{Var} X. \)

(iii) \( \rho_{a_1X+b_1,a_2X+b_2} = 1. \) In particular, if \( Y = aX + b, \) then \( \rho_{X,Y} = 1. \)

(g) \( \rho_{X,Y} = 0 \) if and only if \( X \) and \( Y \) are uncorrelated.

(h) When \( \mathbb{E}X = 0 \) or \( \mathbb{E}Y = 0, \) orthogonality is equivalent to uncorrelatedness.

(i) For finite index set \( I, \)

\[
\text{Var} \left[ \sum_{i \in I} a_iX_i \right] = \sum_{i \in I} a_i^2 \text{Var} X_i + 2 \sum_{(i,j) \in I \times I, i \neq j} a_i a_j \text{Cov} [X_i, X_j].
\]

In particular

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

and

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

(j) For finite index set \( I \) and \( J, \)

\[
\text{Cov} \left[ \sum_{i \in I} a_iX_i, \sum_{j \in J} b_jY_j \right] = \sum_{i \in I} \sum_{j \in J} a_i b_j \text{Cov} [X_i, Y_j].
\]

(k) Covariance Inequality: Let \( X \) be any random variable and \( g \) and \( h \) any function such that \( \mathbb{E} [g(X)], \mathbb{E} [h(X)], \) and \( \mathbb{E} [g(X)h(X)] \) exist.

- If \( g \) and \( h \) are either both non-decreasing or non-increasing, then
  \[
  \text{Cov} [g(X), h(X)] \geq 0. \tag{21}
  \]

- If \( g \) is non-decreasing and \( h \) is non-increasing, then
  \[
  \text{Cov} [g(X), h(X)] \leq 0. \tag{22}
  \]

See also \[2\] p. 191–192 and \[8.13\].

(l) Being uncorrelated does not imply independence

- Discrete: Suppose \( p_X \) is an even function with \( p_X(0) = 0. \) Let \( Y = g(X) \) where \( g \) is also an even function. Then, \( \mathbb{E} [XY] = \mathbb{E} [X] = \mathbb{E} [X] \mathbb{E} [Y] = \text{Cov} [X, Y] = 0. \) Consider a point \( x_0 \) such that \( p_X(x_0) > 0. \) Then, \( p_{XY}(x_0, g(x_0)) = p_X(x_0). \) We only need to show that \( p_Y(g(x_0)) \neq 1 \) to show that \( X \) and \( Y \) are not independent. For example, let \( X \) be uniform on \( \{ \pm 1, \pm 2 \} \) and \( Y = |X| \). Consider the point \( x_0 = 1. \)

- Continuous: Let \( \Theta \) be uniform on an interval of length \( 2\pi. \) Set \( X = \cos \Theta \) and \( Y = \sin \Theta. \) See \[11.6\].
7.17. Suppose $X$ and $Y$ are i.i.d. random variables. Then,

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

7.18. See (4.49) for relationships between expectation and independence.

Example 7.19 (Martingale betting strategy). Fix $a > 0$. Suppose $X_0, X_1, X_2, \ldots$ are independent random variables with $P[X_i = 1] = p$ and $P[X_i = 0] = 1 - p$. Let $N = \inf \{ i : X_i = 1 \}$. Also define

$$L(N) = \begin{cases} 
0, & N = 0 \\
 a \left( \sum_{i=0}^{N-1} r^i \right), & N \in \mathbb{N} 
\end{cases}$$

and $G(N) = ar^N - L(N)$. To have $G(N) > 0$, need $\sum_{i=0}^{k-1} r^i < r^k \forall k \in \mathbb{N}$ which turns out to require $r \geq 2$. In fact, for $r \geq 2$, we have $G(N) \geq a \forall N \in \mathbb{N} \cup \{0\}$. Hence, $\mathbb{E}[G(N)] \geq a$. It is exactly $a$ when $r = 2$.

Now, $\mathbb{E}[L(N)] = a \sum_{n=1}^{\infty} \frac{r^n - 1}{r-1} p (1 - p)^n = \infty$ if and only if $r(1 - p) \geq 1$. When $1 - p \leq \frac{1}{2}$, because we already have $r \geq 2$, it is true that $r(1 - p) \geq 1$.

8 Inequalities

8.1. Let $(A_i : i \in I)$ be a finite family of events. Then

$$\frac{\left( \sum_i P(A_i) \right)^2}{\sum_i \sum_j P(A_i \cap A_j)} \leq P\left( \bigcup_i A_i \right) \leq \sum_i P(A_i)$$

8.2. [20, p. 14]

$$- \left( 1 - \frac{1}{n} \right)^n \leq P\left( \bigcap_{i=1}^{n} A_i \right) - \prod_{i=1}^{n} P(A_i) \leq (n - 1) n^{-\frac{n}{n-1}}.$$ 

\[ \Downarrow \]

$$-\frac{1}{e} \approx -0.37$$

See figure 18

- $|P(A_1 \cap A_2) - P(A_1)P(A_2)| \leq \frac{1}{4}$.

8.3. Markov’s Inequality: $P[|X| \geq a] \leq \frac{1}{a} \mathbb{E}|X|, a > 0$.

(a) Useless when $a \leq \mathbb{E}|X|$. Hence, good for bounding the “tails” of a distribution.

(b) Remark: $P[|X| > a] \leq P[|X| \geq a]$

(c) $P[|X| \geq a\mathbb{E}|X|] \leq \frac{1}{a}, a > 0$. 

78
In a group of 23 randomly selected people, the probability that at least two will share a birthday (assuming birthdays are equally likely to occur on any given day of the year) is about 0.5. See also (3).

The approximation comes from $1 - e^{-\frac{365 \times 23}{365}} \approx 0.5$. From the approximation, to have $pu(n,r) = 0$, we need $n \geq 365$. Hence, good for bounding the "tails" of a distribution.

Figure 18: Bound for $P \left( \bigcap_{i=1}^{n} A_i \right) - \prod_{i=1}^{n} P(A_i)$.

Figure 19: Proof of Markov's Inequality
(d) Suppose \( g \) is a nonnegative function. Then, \( \forall \alpha > 0 \) and \( p > 0 \), we have

(i) \( P [g(X) \geq \alpha] \leq \frac{1}{\alpha^p} (\mathbb{E} [(g(X))^p]) \)

(ii) \( P [g(X - \mathbb{E}X) \geq \alpha] \leq \frac{1}{\alpha^p} (\mathbb{E} [(g(X - \mathbb{E}X))^p]) \)

(e) **Chebyshev’s Inequality:** \( P [|X| > a] \leq P [|X| \geq a] \leq \frac{1}{a^2} \mathbb{E}X^2, a > 0. \)

(i) \( P [|X| \geq a] \leq \frac{1}{a^p} (\mathbb{E} [|X|^p]) \)

(ii) \( P [|X - \mathbb{E}X| \geq \alpha] \leq \frac{\sigma_X^2}{\alpha^2}; \) that is \( P [|X - \mathbb{E}X| \geq n\sigma_X] \leq \frac{1}{n^2} \)

- Useful only when \( \alpha > \sigma_X \)

(iii) For \( a < b \), \( P [a \leq X \leq b] \geq 1 - \frac{4}{(b-a)^2} \left( \frac{\sigma_X^2}{\mathbb{E}X} \right) \)

(f) One-sided Chebyshev inequalities: If \( X \in L^2 \), for \( a > 0 \),

(i) If \( \mathbb{E}X = 0 \), \( P [X \geq a] \leq \frac{\mathbb{E}X^2}{\mathbb{E}X^2 + a^2} \)

(ii) For general \( X \),

i. \( P [X \geq \mathbb{E}X + a] \leq \frac{\sigma_X^2}{\mathbb{E}X^2 + a^2}; \) that is \( P [X \geq \mathbb{E}X + n\sigma_X] \leq \frac{1}{1+n^2} \)

ii. \( P [X \leq \mathbb{E}X - a] \leq \frac{\sigma_X^2}{\mathbb{E}X^2 + a^2}; \) that is \( P [X \leq \mathbb{E}X - n\sigma_X] \leq \frac{1}{1+n^2} \)

iii. \( P [|X - \mathbb{E}X| \geq a] \leq \frac{2\sigma_X^2}{\mathbb{E}X^2 + a^2}; \) that is \( P [|X - \mathbb{E}X| \geq n\sigma_X] \leq \frac{2}{1+n^2} \) This is a better bound than \( \frac{\sigma_X^2}{a^2} \) iff \( \sigma_X > a \)

(g) Chernoff bounds:

(i) \( P [X \leq b] \leq \frac{\mathbb{E}[e^{-\theta X}]}{e^{-\theta b}} \quad \forall \theta > 0 \)

(ii) \( P [X \geq b] \leq \frac{\mathbb{E}[e^{\theta X}]}{e^{\theta b}} \quad \forall \theta > 0 \)

This can be optimized over \( \theta \)

8.4. Suppose \( |X| \leq M \) a.s., then \( P [|X| \geq a] \geq \frac{\mathbb{E}[X|a]}{M-a} \forall a \in [0, M) \)

8.5. \( X \geq 0 \) and \( \mathbb{E}[X^2] < \infty \Rightarrow P [X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \)

**Definition 8.6.** If \( p \) and \( q \) are positive real numbers such that \( p + q = pq \), or equivalently, \( \frac{1}{p} + \frac{1}{q} = 1 \), then we call \( p \) and \( qa \) pair of conjugate exponents.

- \( 1 < p, q < \infty \)

- As \( p \to 1, q \to \infty \). Consequently, \( 1 \) and \( \infty \) are also regarded as a pair of conjugate exponents.

8.7. **Hölder’s Inequality:** \( X \in L^p, Y \in L^q, p > 1, \frac{1}{p} + \frac{1}{q} = 1 \). Then,

(a) \( XY \in L^1 \)
(b) \( \mathbb{E} ||XY|| \leq (\mathbb{E} ||X||^p)^{\frac{1}{p}} (\mathbb{E} ||Y||^q)^{\frac{1}{q}} \) with equality if and only if
\[
\mathbb{E} ||Y||^q |X(\omega)|^p = \mathbb{E} ||X||^p |Y(\omega)|^q \text{ a.s.}
\]

8.8. Cauchy-Bunyakovskii-Schwartz Inequality: If \( X, Y \in L^2 \), then \( XY \in L^1 \) and
\[
|\mathbb{E}[XY]| \leq |\mathbb{E}[XY]| \leq (\mathbb{E} ||X||^2)^{\frac{1}{2}} (\mathbb{E} ||Y||^2)^{\frac{1}{2}} \text{ or equivalently,}
\]
\[
(\mathbb{E}[XY])^2 \leq (\mathbb{E}[X^2]) (\mathbb{E}[Y^2])
\]
with equality if and only if \( \mathbb{E}[Y^2] X^2 = \mathbb{E}[X^2] Y^2 \) a.s.

(a) \( |\text{Cov}(X,Y)| \leq \sigma_X \sigma_Y \).
(b) \( (EX)^2 \leq EX \)
(c) \( (P(A \cap B))^2 \leq P(A)P(B) \)

8.9. Minkowski’s Inequality: \( p \geq 1, X, Y \in L^p \Rightarrow X + Y \in L^p \) and
\[
(\mathbb{E} ||X+Y||^p)^{\frac{1}{p}} \leq (\mathbb{E} ||X||^p)^{\frac{1}{p}} + (\mathbb{E} ||Y||^p)^{\frac{1}{p}}
\]

8.10. \( p > q > 0 \)
(a) \( \mathbb{E} ||X||^q \leq 1 + \mathbb{E} ||X||^p \)
(b) Lyapounov’s inequality: \( (\mathbb{E} ||X||^q)^{\frac{1}{q}} \leq (\mathbb{E} ||X||^p)^{\frac{1}{p}} \)

\[\begin{align*}
\mathbb{E} ||X|| &\leq \sqrt{\mathbb{E} ||X||^2} \leq (\mathbb{E} ||X||^3)^{\frac{1}{3}} \leq \cdots \\
\mathbb{E}|X-Y| &\leq \sqrt{\mathbb{E} [(X-Y)^2]} \\
&\quad \circ \text{If } X \text{ and } Y \text{ are independent, then the RHS is } \sqrt{\sigma_X^2 + \sigma_Y^2 + (EX-\mathbb{E}Y)^2}.
&\quad \circ \text{If } X \text{ and } Y \text{ are i.i.d., then the RHS is } \sqrt{2}\sigma_X.
\end{align*}\]

8.11. Jensen’s Inequality: For a random variable \( X \), if 1) \( X \in L^1 \) (and \( \varphi(X) \in L^1 \));
2) \( X \in (a,b) \) a.s.; and 3) \( \varphi \) is convex on \( (a,b) \), then \( \varphi(\mathbb{E}X) \leq \mathbb{E}[\varphi(X)] \)

\[\begin{align*}
&\quad \circ \text{For } X > 0 \text{ (a.s.), } \mathbb{E}\left(\frac{1}{X}\right) \geq \frac{1}{\mathbb{E}X}.
\end{align*}\]

8.12.
\[\begin{align*}
&\quad \circ \text{For } p \in (0,1], \mathbb{E} ||X+Y||^p \leq \mathbb{E} ||X||^p + \mathbb{E} ||Y||^p.
&\quad \circ \text{For } p \geq 1, \mathbb{E} ||X+Y||^p \leq 2^{p-1} (\mathbb{E} ||X||^p + \mathbb{E} ||Y||^p).
\end{align*}\]

8.13 (Covariance Inequality). Let \( X \) be any random variable and \( g \) and \( h \) any function such that \( \mathbb{E}[g(X)], \mathbb{E}[h(X)], \text{ and } \mathbb{E}[g(X)h(X)] \) exist.

\[\begin{align*}
&\quad \circ \text{If } g \text{ and } h \text{ are either both non-decreasing or non-increasing, then}
&\quad \quad \mathbb{E}[g(X)h(X)] \geq \mathbb{E}[g(X)] \mathbb{E}[h(X)].
&\quad \circ \text{In particular, for nondecreasing } g, \mathbb{E}[g(X)(X-\mathbb{E}X)] \geq 0.
&\quad \circ \text{If } g \text{ is non-decreasing and } h \text{ is non-increasing, then}
&\quad \quad \mathbb{E}[g(X)h(X)] \leq \mathbb{E}[g(X)] \mathbb{E}[h(X)].
\end{align*}\]

See also \([21]\), \([22]\), and \([2]\) p. 191–192.

81
9 Random Vectors

In this article, a vector is a column matrix with dimension \( n \times 1 \) for some \( n \in \mathbb{N} \). We use \( 1 \) to denote a vector with all element being 1. Note that \( 1(1^T) \) is a square matrix with all element being 1. Finally, for any matrix \( A \) and constant \( a \), we define the matrix \( A + a \) to be the matrix \( A \) with each of the components are added by \( a \). If \( A \) is a square matrix, then \( A + a = A + a1(1^T) \).

**Definition 9.1.** Suppose \( I \) is an index set. When \( X_i \)'s are random variables, we define a random vector \( X_I \) by \( X_I = (X_i : i \in I) \). For example, if \( I = [n] \), we have \( X_I = (X_1, X_2, \ldots, X_n) \). Note also that \( X_{[n]} \) is usually denoted by \( X^n \). Sometimes, we simply write \( X \) to denote \( X^n \).

- For disjoint \( A, B \), \( X_{A \cup B} = (X_A, X_B) \).
- For vector \( x_1, y_1 \), we say \( x \leq y \) if \( \forall i \in I, x_i \leq y_i \) [7, p 206].
- When the dimension of \( X \) is implicit, we simply write \( X \) and \( x \) to represent \( X^n \) and \( x_n \), respectively.
- For random vector \( X, Y \), we use \((X, Y)\) to represent the random vector \((X^T) \) or equivalently \([XT YT]^T \).

**Definition 9.2.** *Half-open cell* or *bounded rectangle* in \( \mathbb{R}^k \) is set of the form \( I_{a,b} = \{ x : a_i < x_i \leq b_i, \forall i \in [k] \} = \bigwedge_{i=1}^{k} (a_i, b_i) \). For a real function \( F \) on \( \mathbb{R}^k \), the difference of \( F \) around the vertices of \( I_{a,b} \) is

\[
\Delta_{I_{a,b}} F = \sum_{v} \left( \text{sgn}_{I_{a,b}} (v) \right) F (v) = \sum_{v} (-1)^{|\{ i : v_i = a_i \}|} F (v) \tag{23}
\]

where the sum extending over the \( 2^k \) vertices \( v \) of \( I_{a,b} \). (The \( i \)th coordinate of the vertex \( v \) could be either \( a_i \) or \( b_i \).) In particular, for \( k = 2 \), we have

\[
F (b_1, b_2) - F (a_1, b_2) - F (b_1, a_2) + F (a_1, a_2) .
\]

9.3 (Joint cdf).

\[
F_X (x) = F_{X^k} (x^k) = P [X_1 \leq x_1, \ldots, X_k \leq x_k] = P \left[ X \in S_x^k \right] = P^X (S_x)
\]

where \( S_x = \{ y : y_i \leq x_i, \ i = 1, \ldots, k \} \) consists of the points “southwest” of \( x \).

- \( \Delta_{I_{a,b}} F_X \geq 0 \)
- The set \( S_x \) is an orthanlike or semi-infinite corner with “northeast” vertex (vertex in the direction of the first orthant) specified by the point \( x \) [7, p 206].

C1 \( F_X \) is nondecreasing in each variable. Suppose \( \forall i y_i \geq x_i \), then \( F_X (y) \geq F_X (x) \)

C2 \( F_X \) is continuous from above: \( \lim_{h \searrow 0} F_X (x_1 + h, \ldots, x_k + h) = F_X (x) \)
C3 $x_i \to -\infty$ for some $i$ (the other coordinates held fixed), then $F_X(x) \to 0$
If $\forall i \ x_i \to \infty$, then $F_X(x) \to 1$.

- $\lim_{h_i \to 0} F_X(x_1 - h, \ldots, x_k - h) = P^X(S_x^0)$ where $S_x^0 = \{y : y_i < x_i, \ i = 1, \ldots, k\}$ is the interior of $S_x$.

- Given $a \leq b$, $P [X \in I_{a,b}] = \Delta_{I_{a,b}} F_X$.
  This comes from (10) with $A_i = [X_i \leq a_i]$ and $B = [\forall X_i \leq b_i]$. Note that
  $P [\bigcap_i \in I A_i \cap B] = F(v)$ where $v_i = \begin{cases} a_i, & i \in I, \\ b_i, & \text{otherwise}. \end{cases}$

- For any function $F$ on $\mathbb{R}^k$ with satisfies (C1), (C2), and (C3), there is a unique probability measure $\mu$ on $B_{\mathbb{R}^k}$ such that $\forall a, \forall b \in \mathbb{R}^k$ with $a \leq b$, we have $\mu(I_{a,b}) = \Delta_{I_{a,b}} F$ (and $\forall x \in \mathbb{R}^k \mu(S_x) = F(x)$).

- TFAE:
  (a) $F_X$ is continuous at $x$
  (b) $F_X$ is continuous from below
  (c) $F_X(x) = P^X(S_x^0)$
  (d) $P^X(S_x) = P^X(S_x^0)$
  (e) $P^X(\partial S_x) = 0$ where $\partial S_x = S_x - S_x^0 = \{y : y_i \leq x_i \forall i, \exists j y_j = x_j\}$

- If $k > 1$, $F_X$ can have discontinuity points even if $P^X$ has no point masses.
- $F_X$ can be discontinuous at uncountably many points.
- The continuity points of $F_X$ are dense.
- For any $j$, we have $\lim_{x_j \to \infty} F_X(x) = F_X|_{\{j\}}(x|_{\{j\}})$

9.4 (Joint pdf). A function $f$ is a multivariate or joint pdf (commonly called a density) if and only if it satisfies the following two conditions:

(a) $f \geq 0$;
(b) $\int f(x)d(x) = 1$.

- The integrability of the pdf $f$ implies that for all $i \in I$ 
  $\lim_{x_i \to \pm \infty} f(x_I) = 0$.

- $P[X \in A] = \int_A f_X(x)dx$.
- Remarks: Roughly, we may say the following:

  (a) $f_X(x) = \lim_{\forall i, \Delta x_i \to 0} \frac{P[\forall i, x_i < X_i \leq x_i + \Delta x_i]}{\prod_i \Delta x_i} = \lim_{\Delta x_i \to 0} \frac{P[x < X_i \leq x + \delta x_i]}{\prod_i \Delta x_i}$

83
For $I = [n]$, 
\[ F_X(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(x_1, \ldots, x_n) \, dx_1 \cdots dx_n \]
\[ f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \cdots \partial x_n} \]
\[ \frac{\partial}{\partial u} F_X(u, \ldots, u) = \sum_{k \in \mathbb{N}} \int_{-\infty}^{u} f_X(v_k, y_{k+1}, \ldots, y_{n}) \, dy \]
\[ f_X^n(x^n) = \mathbb{E} \left[ \prod_{i=1}^{n} \delta(X_i - x_i) \right]. \]

- The **level sets** of a density are sets where density is constant.

9.5. Consider two random vectors $X : \Omega \to \mathbb{R}^d_1$ and $Y : \Omega \to \mathbb{R}^{d_2}$. Define $Z = (X, Y) : \Omega \to \mathbb{R}^{d_1+d_2}$. Suppose that $Z$ has density $f_{X,Y}(x, y)$.

(a) **Marginal Density**: $f_Y(y) = \int_{\mathbb{R}^{d_1}} f_{X,Y}(x, y) \, dx$ and $f_X(x) = \int_{\mathbb{R}^{d_2}} f_{X,Y}(x, y) \, dy$.

- In other words, to obtain the marginal densities, integrate out the unwanted variables.

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

(c) $F_{Y|X}(y|x) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_{d_2}} f_{Y|X}(t|x) \, dt_{d_2} \cdots dt_{d_1}$.

9.6. $P[(X + a_1, X + b_1) \cap (Y + a_2, Y + b_2) \neq \emptyset] = \int_A f_{X,Y}(x, y) \, dxdy$ where $A$ is defined in (1.10).

9.7. Expectation and covariance:

(a) The expectation of a random vector $X$ is defined to be the vector of expectations of its entries. $\mathbb{E}X$ is usually denoted by $\mu_X$ or $m_X$.

(b) For non-random matrix $A$, $B$, $C$ and a random vector $X$, $\mathbb{E}[AXB + C] = A \mathbb{E}XB + C$.

(c) The correlation matrix $R_X$ of a random vector $X$ is defined by

\[ R_X = \mathbb{E}[XX^T]. \]

Note that it is symmetric.

(d) The covariance matrix $C_X$ of a random vector $X$ is defined as

\[ C_X = \Lambda_X = \text{Cov}[X] = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] = \mathbb{E}[XX^T] - (\mathbb{E}X)(\mathbb{E}X)^T = R_X - (\mathbb{E}X)(\mathbb{E}X)^T. \]

(i) The $ij$-entry of the Cov $[X]$ is simply Cov $[X_i, X_j]$. 

84
(ii) $A_X$ is symmetric.

i. Properties of symmetric matrix
   A. All eigenvalues are real.
   B. Eigenvectors corresponding to different eigenvalues are not just linearly independent, but mutually orthogonal.
   C. Diagonalizable.

ii. **Spectral theorem**: The following equivalent statements hold for symmetric matrix.
   A. There exists a complete set of eigenvectors; that is there exists an orthonormal basis $u^{(1)}, \ldots, u^{(n)}$ of $\mathbb{R}$ with $C_X u^{(k)} = \lambda_k u^{(k)}$.
   B. $C_X$ is diagonalizable by an orthogonal matrix $U$ ($U U^T = U^T U = I$).
   C. $C_X$ can be represented as $C_X = U \Lambda U^T$ where $U$ is an orthogonal matrix whose columns are eigenvectors of $C_X$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with the eigenvalues of $C_X$.

(iii) Always nonnegative definite (positive semidefinite). That is $\forall a \in \mathbb{R}^n$ where $n$ is the dimension of $X$, $a^T C_X a = \mathbb{E} \left[ (a^T (X - \mu_X))^2 \right] \geq 0$.
   - $\det (C_X) \geq 0$.

(iv) We can define $C_X^{\frac{1}{2}} = \sqrt{C_X}$ to be $\sqrt{C_X} = U \sqrt{\Lambda} U^T$ where $\sqrt{\Lambda} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$.
   i. $\det \sqrt{C_X} = \sqrt{\det C_X}$.
   ii. $\sqrt{C_X}$ is nonnegative definite.
   iii. $\sqrt{C_X}^2 = \sqrt{C_X} \sqrt{C_X} = C_X$.

(v) Suppose, furthermore, that $C_X$ is positive definite.
   i. $C_X^{-1} = U \Lambda^{-1} U^T$ where $\Lambda^{-1} = \text{diag}(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n})$.
   ii. $C_X^{-\frac{1}{2}} = \sqrt{C_X^{-1}} = (\sqrt{C_X})^{-1} = U D U^T$ where $D = \left( \frac{1}{\sqrt{\lambda_1}}, \ldots, \frac{1}{\sqrt{\lambda_n}} \right)$
   iii. $\sqrt{C_X} C_X^{-1} \sqrt{C_X} = I$.
   iv. $C_X^{-1}, C_X^{\frac{1}{2}}, C_X^{-\frac{1}{2}}$ are all positive definite (and hence are all symmetric).
   v. $\left( C_X^{-\frac{1}{2}} \right)^2 = C_X^{-1}$.
   vi. Let $Y = C_X^{-\frac{1}{2}} (X - \mathbb{E} X)$. Then, $\mathbb{E} Y = 0$ and $C_Y = I$.

(vi) For i.i.d. $X_i$ with each with variance $\sigma^2$, $\Lambda_X = \sigma^2 I$.

(e) $\text{Cov} [A X + b] = A \text{Cov} [X] A^T$
   - $\text{Cov} [X^T h] = \text{Cov} [h^T X] = h^T \text{Cov} [X] h$ where $h$ is a vector with the same dimension as $X$.

(f) For $Y = X + Z$, $\Lambda_Y = \Lambda_X + 2 \Lambda_{XZ} + \Lambda_Z$.
   - When $X$ and $Z$ are independent, $\Lambda_Y = \Lambda_X + \Lambda_Z$.
• For $Y_i = X + Z_i$ where $X$ and $Z$ are independent, $\Lambda_Y = \sigma_X^2 + \Lambda_Z$.

(g) $\Lambda_{X+Y} + \Lambda_{X-Y} = 2\Lambda_X + 2\Lambda_Y$

(h) $\det (\Lambda_{X+Y}) \leq 2^n \det (\Lambda_X + \Lambda_Y)$ where $n$ is the dimension of $X$ and $Y$.

(i) $Y = (X, X, \ldots, X)$ where $X$ is a random variable with variance $\sigma_X^2$, then $\Lambda_Y = \sigma_X^2 \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$. Note that $Y = 1X$ where $1$ has the same dimension as $Y$.

(j) Let $X$ be a zero-mean random vector whose covariance matrix is singular. Then, one of the $X_i$ is a deterministic linear combination of the remaining components. In other words, there is a nonzero vector $a$ such that $a^T X = 0$. In general, if $\Lambda_X$ is singular, then there is a nonzero vector $a$ such that $a^T X = a^T E X$.

(k) If $X$ and $Y$ are both random vectors (not necessarily of the same dimension), then their cross-covariance matrix is

$\Lambda_{XY} = C_{XY} = \text{Cov} [X, Y] = E [(X - E X)(Y - E Y)^T]$.

Note that the $ij$-entry of $C_{XY}$ is $\text{Cov} [X_i, Y_j]$.

• $C_{YX} = (C_{XY})^T$.

(l) $R_{XY} = \mathbb{E} [XY^T]$.

(m) If we stack $X$ and $Y$ in to a composite vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$, then

$C_Z = \begin{pmatrix} C_X & C_{XY} \\ C_{YX} & C_Y \end{pmatrix}$.

(n) $X$ and $Y$ are said to be uncorrelated if $C_{XY} = 0$, the zero matrix. In which case,$

C_{\begin{bmatrix} X \\ Y \end{bmatrix}} = \begin{pmatrix} C_X & 0 \\ 0 & C_Y \end{pmatrix},$

a block diagonal matrix.

9.8. The **joint characteristic function** of an $n$-dimensional random vector $X$ is defined by

$\varphi_X(v) = \mathbb{E} [e^{ju^T X}] = \mathbb{E} [e^{j \sum_i v_i X_i}]$.

When $X$ has a joint density $f_X$, $\varphi_X$ is just the $n$-dimensional Fourier transform:

$\varphi_X(v) = \int e^{ju^T x} f_X(x) dx,$

and the joint density can be recovered using the multivariate inverse Fourier transform:

$f_X(x) = \frac{1}{(2\pi)^n} \int e^{-ju^T x} \varphi_X(v) dv.$
(a) \( \varphi_X (u) = \mathbb{E} e^{iu^T X} \).

(b) \( f_X (x) = \frac{1}{(2\pi)^{n/2}} \int e^{-jv^T x} \varphi_X (v) dv. \)

(c) For \( Y = AX + b, \varphi_Y (u) = e^{ib^T u} \varphi_X (A^T u) \).

(d) \( \varphi_X (-u) = \overline{\varphi_X (u)} \).

(e) \( \varphi_X (u) = \varphi_{X,Y} (u, 0) \).

(f) Moment:

\[
\frac{\partial}{\partial v_i} \varphi_X (0) = j^n \mathbb{E} \left( \prod_{i=1}^{n} X_i^{v_i} \right)
\]

(i) \( \frac{\partial}{\partial v_i} \varphi_X (0) = j\mathbb{E} X_i. \)

(ii) \( \frac{\partial^2}{\partial v_i \partial v_j} \varphi_X (0) = j^2 \mathbb{E} [X_i X_j]. \)

(g) Central Moment:

(i) \( \frac{\partial}{\partial v_i} \ln (\varphi_X (v)) \bigg|_{v=0} = j\mathbb{E} X_i. \)

(ii) \( \frac{\partial^2}{\partial v_i \partial v_j} \ln (\varphi_X (v)) \bigg|_{v=0} = -\text{Cov} \left[X_i, X_j \right]. \)

(iii) \( \frac{\partial^3}{\partial v_i \partial v_j \partial v_k} \ln (\varphi_X (v)) \bigg|_{v=0} = j^3 \mathbb{E} \left[ (X_i - \mathbb{E} X_i) (X_j - \mathbb{E} X_j) (X_k - \mathbb{E} X_k) \right]. \)

(iv) \( \mathbb{E} \left[ (X_i - \mathbb{E} X_i) (X_j - \mathbb{E} X_j) (X_k - \mathbb{E} X_k) (X_\ell - \mathbb{E} X_\ell) \right] = \Psi_{ijk\ell} + \Psi_{ij} \Psi_{k\ell} + \Psi_{ik} \Psi_{j\ell} + \Psi_{i\ell} \Psi_{jk} \text{ where } \Psi_{ijk\ell} = \frac{\partial^4}{\partial v_i \partial v_j \partial v_k \partial v_\ell} \ln (\varphi_X (v)) \bigg|_{v=0}. \)

Remark: we do not require that any or all of \( i, j, k, \) and \( \lambda \) be distinct.

9.9 (Decorrelation and the Karhunen-Loève expansion). Let \( X \) be an \( n \)-dimensional random vector with zero mean and covariance matrix \( C \). \( X \) has the representation \( X = PY \), where the components of \( Y \) are uncorrelated and \( P \) is an \( n \times n \) orthonormal matrix. This representation is called the Karhunen-Loève expansion.

- \( Y = P^T X \)
- \( P^T = P^{-1} \) is called a decorrelating transformation.
- Diagonalize \( C = PD P^T \) where \( D = \text{diag}(\lambda_1) \). Then, \( \text{Cov} \left[Y\right] = D \).
- In MATLAB, use \([P, D] = \text{eig}(C)\). To extract the diagonal elements of \( D \) as a vector, use the command \( d = \text{diag}(D) \).
- If \( C \) is singular (equivalently, if some of the \( \lambda_i \) are zero), we only need to keep around the \( Y_i \) for which \( \lambda_i > 0 \) and can throw away the other components of \( Y \) without any loss of information. This is because \( \lambda_i = \mathbb{E} Y_i^2 \) and \( \mathbb{E} Y_i^2 = 0 \) if and only if \( Y_i \equiv 0 \) a.s.

9.1 Random Sequence

9.10. Given a countable family \( X_1, X_2, \ldots \) of random r.v.’s, their statistical properties are regarded as defined by prescribing, for each integer \( n \geq 1 \) and every finite set \( I \subset \mathbb{N} \), the joint distribution function \( F_{X_i} \) of the random vector \( X_I = (X_i : i \in I) \). Of course, some consistency requirements must be imposed upon the infinite family \( F_{X_I} \), namely, that for \( j \in I \)

(a) \( F_{X_{I\setminus\{j\}}} (x_{I\setminus\{j\}}) = \lim_{x_j \to \infty} F_{X_I} (x_I) \) and that

(b) the distribution function obtained from \( F_{X_I} (x_I) \) by interchanging two of the indices \( i_1, i_2 \in I \) and the corresponding variable \( x_{i_1} \) and \( x_{i_2} \) should be invariant. This simply means that the manner of labeling the random variables \( X_1, X_2, \ldots \) is not relevant.

The joint distributions \( \{F_{X_I}\} \) are called the finite-dimensional distributions associated with \( X_N = (X_n)_{n=1}^\infty \).

10 Transform Methods

10.1 Probability Generating Function

Definition 10.1. Let \( X \) be a discrete random variable taking only nonnegative integer values. The probability generating function (pgf) of \( X \) is

\[
G_X(z) = E [z^X] = \sum_{k=0}^{\infty} z^k P[X = k].
\]

- In the summation, the first term (the \( k = 0 \) term) is \( P[X = 0] \) even when \( z = 0 \).
- \( G_X(0) = P[X = 0] \)
- \( G(z^{-1}) \) is the \( z \) transform of the pmf.
- \( G_X(1) = 1. \)
- The names derives from the fact that it can be used to compute the pmf.
- It is finite at least for any complex \( z \) with \( |z| \leq 1 \). Hence pgf is well defined for \( |z| \leq 1 \).

Definition 10.2. \( G_X^{(k)}(1) = \lim_{z \to 1} G_X^{(k)}(z) \).

10.3. Properties

(a) \( G_X \) is infinitely differentiable at least for \( |z| < 1 \).

(b) Probability generating property:

\[
\frac{1}{k!} \frac{d^k}{dz^k} G_X(z) \bigg|_{z=0} = P[X = k].
\]
(c) Moment generating property:

\[ \frac{d^{(k)}}{dz^{(k)}} G_X(z) \bigg|_{z=1} = \mathbb{E} \left[ \prod_{i=0}^{k-1} (X - i) \right]. \]

The RHS is called the \( k \)th factorial moment of \( X \).

(d) In particular,

\[ \mathbb{E} X = G'_X(1) \]
\[ \mathbb{E} X^2 = G''_X(1) + G'_X(1) \]
\[ \text{Var} X = G''_X(1) + G'_X(1) - (G'_X(1))^2 \]

(e) pgf of a sum of independent random variables is the product of the individual pgfs. Let \( S = \sum_{i=1}^{n} X_i \) where the \( X_i \)'s are independent.

\[ G_S(z) = \prod_{i=1}^{n} G_{X_i}(z). \]

### 10.2 Moment Generating Function

**Definition 10.4.** The **moment generating function** of a random variable \( X \) is defined as \( M_X(s) = \mathbb{E} [e^{sX}] = \int e^{sx} P^X(dx) \) for all \( s \) for which this is finite.

**10.5. Properties of moment generating function**

(a) \( M_X(s) \) is defined on some interval containing 0. It is possible that the interval consists of 0 alone.

(i) If \( X \geq 0 \), this interval contains \((-\infty, 0]\).

(ii) If \( X \leq 0 \), this interval contains \([0, \infty)\).

(b) Suppose that \( M(s) \) is defined throughout an interval \((-s_0, s_0)\) where \( s_0 > 0 \), i.e. it exists (is finite) in some neighborhood of 0. Then,

(i) \( X \) has finite moments of all order: \( \mathbb{E} \left[ |X|^k \right] < \infty \forall k \geq 0 \)

(ii) \( M(s) = \sum_{k=0}^{\infty} \frac{s^k}{k!} \mathbb{E} \left[ X^k \right] \), for complex-valued \( s \) with \(|s| < s_0\) [eqn (21.22) p 278].

Thus \( M(s) \) has a Taylor expansion about 0 with positive radius of convergence.

i. If \( M(s) \) can somehow be calculated and expanded in a series \( \sum_{k=0}^{\infty} a_k s^k \), and if the coefficients \( a_k \) can be identified, then \( a_k = \frac{1}{k!} \mathbb{E} \left[ X^k \right] \). That is \( \mathbb{E} \left[ X^k \right] = k! a_k \)

(iii) \( M^{(k)}(0) = \mathbb{E} \left[ X^k \right] = \int x^k P^X(dx) \)

(c) If \( M \) is defined in some neighborhood of \( s \), then \( M^{(k)}(s) = \int x^k e^{sx} P^X(dx) \)

(d) See also Chernoff bound.
10.3 One-Sided Laplace Transform

10.6. The **one-sided Laplace transform** of a nonnegative random variable $X$ is defined for $s \geq 0$ by $L(s) = M(s) = \mathbb{E}[e^{-sx}] = \int_{0,\infty} e^{-sx}P^X(dx)$

- Note that 0 is included in the range of integration.
- Always finite because $e^{-sx} \leq 1$. In fact, it is a decreasing function of $s$
- $L(0) = 1$
- $L(s) \in [0, 1]$ 

(a) **Derivative**: For $s > 0$, $L^{(k)}(s) = (-1)^k \int x^k e^{-sx}P^X(dx) = (-1)^k \mathbb{E}[X^k e^{-sx}]$

(i) $\lim_{s \to 0} \frac{d^n}{ds^n} L(s) = (-1)^n \mathbb{E}[X^n]$
- Because the value at 0 can be $\infty$, it does not make sense to talk about $\frac{d^n}{ds^n} L(0)$ for $n > 1$

(ii) $\forall s \geq 0$ $L(s)$ is differentiable and $\frac{d}{ds} L(s) = -\mathbb{E}[Xe^{-sx}]$, where at 0, this is the right-derivative.
- $\frac{d}{ds} L(0) = -\mathbb{E}[X] \in [0, \infty]$

(b) **Inversion formula**: If $F_X$ is continuous at $t > 0$, then $F_X(t) = \lim_{s \to \infty} \sum_{k=0}^{\lfloor st \rfloor} (-1)^k \frac{1}{k!} s^k L^{(k)}(s)$

(c) $F_X$ and $P^X$ are determined by $L_X(s)$

(i) In fact, they are determined by the values of $L_X(s)$ for $s$ beyond any arbitrary $s_0$. (That is we don’t need to know $L_X(s)$ for small $s$.) Also, knowing $L_X(s)$ on $\mathbb{N}$ is also sufficient.

(ii) Let $\mu$ and $\nu$ be probability measures on $[0, \infty)$. If $\exists s_0 \geq 0$ such that $\int e^{-sx}\mu(dx) = \int e^{-sx}\nu(dx) \forall s \geq s_0$, then $\mu = \nu$

(iii) Let $f_1, f_2$ be real functions on $[0, \infty)$. If $\exists s_0 \geq 0$ such that $\forall s \geq s_0 \int e^{-sx}f_1(x)dx = \int_{[0,\infty)} e^{-sx}f_2(x)dx$, then $f_1 = f_2$ Lebesgue-a.e.

(d) Let $X_1, \ldots, X_n$ be independent nonnegative random variables, then $L_{\sum_{i=1}^{n} X_i}(s) = \prod_{i=1}^{n} L_X(s)$

(e) Suppose $F$ is a distribution function with corresponding Laplace transform $L$. Then

(i) $\int_{[0,\infty)} e^{-sx}F(x)dx = \frac{1}{s}L(s)$

(ii) $\int_{[0,\infty)} e^{-sx}(1 - F(x))dx = \frac{1}{s}(1 - L(s))$

10.4 Characteristic Function

10.7. The **characteristic function** (abbreviated c.f. or ch.f.) of a probability measure \( \mu \) on the line is defined for real \( t \) by \( \varphi(t) = \mathbb{E}[e^{itX}] = \int e^{itx} \mu(dx) \)

A random variable \( X \) has characteristic function \( \varphi_X(t) = \mathbb{E}[e^{itX}] = \int e^{itx} f_X(x) dx \)

(a) Always exists because \( |\varphi(t)| \leq \int |e^{itx}| \mu(dx) = \int 1 \mu(dx) = 1 < \infty \)

(b) If \( X \) has a density, then \( \varphi_X(t) = \int e^{itx} f_X(x) dx \)

(c) \( \varphi(0) = 1 \)

(d) \( \forall t \in \mathbb{R} \) \( |\varphi(t)| \leq 1 \)

(e) \( \varphi \) is uniformly continuous.

(f) Suppose that all moments of \( X \) exists and \( \forall t \in \mathbb{R} \), \( \mathbb{E}[e^{itX}] < \infty \), then \( \varphi_X(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \mu(X^k) \)

(g) If \( \mathbb{E}[|X|^k] < \infty \), then \( \varphi^{(k)}(t) = i^k \mathbb{E}[X^k e^{itX}] \) and \( \varphi^{(k)}(0) = i^k \mathbb{E}[X^k] \).

(h) **Riemann-Lebesgue theorem:** If \( X \) has a density, then \( \varphi_X(t) \to 0 \) as \( |t| \to \infty \)

(i) \( \varphi_{aX+b}(t) = e^{itb} \varphi(at) \)

(j) **Conjugate Symmetry Property:** \( \varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)} \)

\( \bullet \) \( X \overset{D}{=} -X \) iff \( \varphi_X \) is real-valued.

\( \bullet \) \( |\varphi_X| \) is even.

(k) \( X \) is a.s. integer-valued if and only if \( \varphi_X(2\pi) = 1 \)

(l) If \( X_1, X_2, \ldots, X_n \) are independent, then \( \varphi \sum_{j=1}^{n} X_j(t) = \prod_{j=1}^{n} \varphi_X(t) \)

(m) **Inversion**

(i) **The inversion formula:** If the probability measure \( \mu \) has characteristic function \( \varphi \) and if \( \mu \{a\} = \mu \{b\} = 0 \), then \( \mu(a,b) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \)

i. In fact, if \( a < b \), then \( \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt = \mu(a,b) + \frac{1}{2} \mu \{a,b\} \)

ii. Equivalently, if \( F \) is the distribution function, and \( a, b \) are continuity points of \( F \), then \( F(b) - F(a) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \)

(ii) **Fourier inversion:** Suppose that \( \int \varphi_X(t) |dt| < \infty \), then \( X \) is absolutely continuous with
i. bounded continuous density \( f(x) = \frac{1}{2\pi} \int e^{-itx} \varphi(t) dt \)

ii. \( \mu(a,b) = \frac{1}{2\pi} \int \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt \)

(n) **Continuity Theorem**: \( X_n \overset{D}{\to} X \) if and only if \( \forall t \ varphi_X(t) \to \varphi_X(t) \) (pointwise).

(o) \( \varphi_X \) on complex plane for \( X \geq 0 \)

(i) \( \varphi_X(z) \) is defined in the complex plane for \( \text{Im} z \geq 0 \)

• \( |\varphi_X(z)| \leq 1 \) for such \( z \)

(ii) In the domain \( \text{Im} \{z\} > 0 \), \( \varphi_X \) is analytic and continuous including the boundary \( \text{Im} \{z\} = 0 \)

(iii) \( \varphi_X \) determines uniquely a function \( L_X(s) \) of real argument \( s \geq 0 \) which is equal to \( L_X(s) = \varphi_X(is) = \mathbb{E}e^{-sx} \). Conversely, \( L_X(s) \) on the half-line \( s \geq 0 \) determines uniquely \( \varphi_X \)

(p) If \( \mathbb{E}|X|^n < \infty \), then

\[
\left| \varphi_X(t) - \sum_{k=0}^{n} \frac{(it)^k}{k!} \mathbb{E}X^k \right| \leq \mathbb{E} \left[ \min \left\{ \frac{|t|^{n+1}}{(n+1)!} |X|^{n+1}, \frac{2|t|^n}{n!} |X|^n \right\} \right] \quad (24)
\]

and \( \varphi_X(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} \mathbb{E}X^k + t^n \beta(t) \) where \( \lim_{|t| \to 0} \beta(t) = 0 \) or equivalently,

\[
\varphi_X(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} \mathbb{E}X^k + o(t^n) \quad (25)
\]

(i) \( |\varphi_X(t) - 1| \leq \mathbb{E} \left[ \min \{ |tX|, 2 \} \right] \)

• If \( \mathbb{E}X = 0 \), then \( |\varphi_X(t) - 1| \leq \mathbb{E} \left[ \min \left\{ \frac{t^2}{2} X^2, 2 |t| |X| \right\} \right] \leq \frac{t^2}{2} \mathbb{E}X^2 \)

(ii) For integrable \( X \), \( |\varphi_X(t) - 1 - it\mathbb{E}X| \leq \mathbb{E} \left[ \min \left\{ \frac{t^2}{2} X^2, 2 |t| |X| \right\} \right] \leq \frac{t^2}{2} \mathbb{E}X^2 \)

(iii) For \( X \) with finite \( \mathbb{E}X^2 \),

• \( |\varphi_X(t) - 1 - it\mathbb{E}X + \frac{1}{2} t^2 \mathbb{E}X^2| \leq \mathbb{E} \min \left\{ \frac{|t|^3}{6} |X|^3, t^2 |X|^2 \right\} \)

10.8. \( \phi_X(u) = M_X(\text{ju}) \).

10.9. If \( X \) is a continuous r.v. with density \( f_X \), then \( \varphi_X(t) = \int e^{itx} f_X(x) dx \) (Fourier transform) and \( f_X(x) = \frac{1}{2\pi} \int e^{-itx} \varphi_X(t) dt \) (Fourier inversion formula).

• \( \varphi_X(t) \) is the Fourier transform of \( f_X \) evaluated at \( -t \).

• \( \varphi_X \) inherit the properties of a Fourier transform.

(a) For nonnegative \( a_i \) such that \( \sum_i a_i = 1 \), if \( f_Y = \sum_i a_i f_{X_i} \), then \( \varphi_X = \sum_i a_i \varphi_{X_i} \).
(b) If \( f_X \) is even, then \( \varphi_X \) is also even.

- If \( f_X \) is even, \( \varphi_X = \varphi_{-X} \).

10.10. Linear combination of independent random variables: Suppose \( X_1, \ldots, X_n \) are independent. Let \( Y = \sum_{i=1}^n a_i X_i \). Then, \( \varphi_Y(t) = \prod_{i=1}^n \varphi_X(a_i t) \). Furthermore, if \( |a_i| = 1 \) and all \( f_{X_i} \) are even, then \( \varphi_Y(t) = \prod_{i=1}^n \varphi_X(t) \).

(a) \( \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \).

(b) \( \varphi_{X-Y}(t) = \varphi_X(t)\varphi_Y(-t) \).

(c) If \( f_Y \) is even, \( \varphi_{X-Y} = \varphi_{X+Y} = \varphi_X \varphi_Y \).

10.11. Characteristic function for sum of distribution: Consider nonnegative \( a_i \) such that \( \sum_i a_i = 1 \). Let \( P_i \) be probability measure with corresponding ch.f. \( \varphi_i \). Then, the ch.f. of \( \sum_i a_i P_i \) is \( \sum_i a_i \varphi_i \).

(a) Discrete r.v.: Suppose \( p_i \) is pmf with corresponding ch.f. \( \varphi_i \). Then, the ch.f. of \( \sum_i a_i p_i \) is \( \sum_i a_i \varphi_i \).

(b) Absolutely continuous r.v.: Suppose \( f_i \) is pdf with corresponding ch.f. \( \varphi_i \). Then, the ch.f. of \( \sum_i a_i f_i \) is \( \sum_i a_i \varphi_i \).

11 Functions of random variables

Definition 11.1. The preimage or inverse image of a set \( B \) is defined by \( g^{-1}(B) = \{ x : g(x) = B \} \).

11.2. For discrete \( X \), suppose \( Y = g(X) \). Then, \( p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x) \). The joint pmf of \( Y \) and \( X \) is given by \( p_{X,Y}(x,y) = p_X(x)1[y = g(x)] \).

- In most cases, we can show that \( X \) and \( Y \) are not independent, pick a point \( x^{(0)} \) such that \( p_X(x^{(0)}) > 0 \). Pick a point \( y^{(1)} \) such that \( y^{(1)} \neq g(x^{(0)}) \) and \( p_Y(y^{(1)}) > 0 \). Then, \( p_{X,Y}(x^{(0)},y^{(1)}) = 0 \) but \( p_X(x^{(0)}) p_Y(y^{(1)}) > 0 \). Note that this technique does not always work. For example, if \( g \) is a constant function which maps all values of \( x \) to a constant \( c \). Then, we will not be able to find \( y^{(1)} \). Of course, this is to be expected because we know that a constant is always independent of other random variables.

11.1 SISO case

There are many techniques for finding the cdf and pdf of \( Y = g(X) \).

(a) One may first find \( F_Y(y) = P[g(X) \leq y] \) first and then find \( f_Y \) from \( F_Y(y) \). In which case, the Leibniz’ rule in \( [52] \) will be useful.

(b) Formula \( [27] \) below provides a convenient way of arriving at \( f_Y \) from \( f_X \) without going through \( F_Y \).
11.3 (Linear transformation). \( Y = aX + b \) where \( a \neq 0 \).

\[
F_Y(y) = \begin{cases} 
F_X \left( \frac{y-b}{a} \right), & a > 0 \\
1 - F_X \left( \left( \frac{y-b}{a} \right)^- \right), & a < 0 
\end{cases}
\]

(a) Suppose \( X \) is absolutely continuous,

\[
f_Y(y) = \frac{1}{|a|} f_X \left( \frac{y-b}{a} \right)
\]  \hfill (26)

In fact (26) holds even for mixed r.v. if we allow delta function because \( \frac{1}{|a|} \delta \left( \frac{y-b}{a} - x_k \right) = \delta (y - (ax_k + b)) \).

(b) Suppose \( X \) is discrete,

\[
p_Y(y) = p_X \left( \frac{y-b}{a} \right)
\]

If we write \( f_X(x) = \sum_k p_X(x_k) \delta(x - x_k) \), we have

\[
f_Y(y) = \sum_k p(x_k) \delta(y - (ax_k + b)).
\]

11.4 (Power Law Function). \( Y = X^n \), \( n \in \mathbb{N} \) or \( n \in (0, \infty) \).

(a) \( n \) odd: \( F_Y(y) = F_X \left( y^{\frac{1}{n}} \right) \) and \( f_Y(y) = \frac{1}{n} y^{\frac{1}{n}-1} f_X \left( y^{\frac{1}{n}} \right) \).

(b) \( n \) even:

\[
F_Y(y) = \begin{cases} 
F_X \left( y^{\frac{1}{n}} \right) - F_X \left( \left( -y^{\frac{1}{n}} \right)^- \right), & y \geq 0 \\
0, & y < 0 
\end{cases}
\]

and

\[
f_Y(y) = \begin{cases} 
\frac{1}{n} y^{\frac{1}{n}-1} \left( f_X \left( y^{\frac{1}{n}} \right) + f_X \left( -y^{\frac{1}{n}} \right) \right), & y \geq 0 \\
0, & y < 0 
\end{cases}
\]

Again, the density \( f_Y \) in the above formula holds when \( X \) is absolutely continuous. Note that when \( n < 1 \), \( f_Y \) is not defined at 0. If we allow delta functions, then the density formula above are also valid for mixed r.v. because \( \frac{1}{n} y^{\frac{1}{n}-1} \delta \left( \pm y^{\frac{1}{n}} - x_k \right) = \delta (y - (\pm x_k)^n) \).

- Let \( X \) be an absolutely continuous random variable. The density of \( Y = X^2 \) is

\[
f_Y(y) = \begin{cases} 
0, & y < 0 \\
\frac{1}{2 \sqrt{y}} f_X \left( \sqrt{y} \right) + \frac{1}{2 \sqrt{y}} f_X \left( -\sqrt{y} \right), & y \geq 0 
\end{cases}
\]

11.5. In general, for \( Y = g(X) \), we solve the equation \( y = g(x) \). Denoting its real roots by \( x_k \). Then,

\[
f_Y(y) = \sum_k \frac{f_X(x_k)}{|g'(x_k)|} \]  \hfill (27)

If \( g(x) = c = \text{constant for every } x \) in the interval \((a, b)\), then \( F_Y(y) \) is discontinuous for \( y = c \). Hence, \( f_Y(y) \) contains an impulse \( (F_X(b) - F_X(a)) \delta(y - c) \) \[15\] p. 93–94.
To see this, consider when there is unique $x$ such that $g(x) = y$. Then, for small $\Delta x$ and $\Delta y$, $P[y, y < Y \leq y + \Delta y] = P[x < X \leq x + \Delta x]$ where $(y + \Delta y) = g((x, x + \Delta x))$ is the image of the interval $(x, x + \Delta x)$. (Equivalently, $(x, x + \Delta x]$ is the inverse image of $y + \Delta y$..) This gives $f_Y(y)\Delta y = f_X(x)\Delta x$.

The joint density $f_{X,Y}$ is

$$f_{X,Y}(x, y) = f_X(x) \delta(y - g(x)).$$

Let the $x_k$ be the solutions for $x$ of $g(x) = y$. Then, by integrating (28) w.r.t. $x$, we have (27) via the use of (4).

When $g$ bijective,

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y)).$$

For $Y = \frac{X}{a}$, $f_Y(y) = \left| \frac{a}{y} \right| f_X \left( \frac{a}{y} \right)$.

Suppose $X$ is nonnegative. For $Y = \sqrt{X}$,

$$f_Y(y) = 2y f_X(y^2).$$

### 11.6.

Given $Y = g(X)$ where $X \sim U(a,b)$. Then, to get $f_Y(y_0)$, plot $g$ on $(a,b)$. Let $A = g^{-1}(\{y_0\})$ be the set of all points $x$ such that $g(x) = y_0$. Suppose $A$ can be written as a countable disjoint union $A = B \cup \bigcup_i I_i$ where $B$ is countable and the $I_i$'s are intervals. We have

$$f_Y(y) = \frac{1}{b-a} \frac{1}{|g'(x)|} + \left( \sum_i \ell(I_i) \right) \delta(y - y_0)$$

at $y = y_0$ where $\ell(I)$ is the length of the interval $I$.

Suppose $\Theta$ is uniform on an interval of length $2\pi$. $Y_1 = \cos \Theta$ and $Y_2 = \sin \Theta$ are both arcsine random variables with $F_{Y_i}(y) = 1 - \frac{1}{\pi} \cos^{-1} y = \frac{1}{2} + \frac{1}{\pi} \sin^{-1}(y)$ and $f_{Y_i}(y) = \frac{1}{\sqrt{1-y^2}}$ for $y \in [-1, 1]$. Note also that $\mathbb{E}[Y_1 Y_2] = \mathbb{E}Y_i = \text{Cov}[Y_1, Y_2] = 0$.

Hence, $Y_1$ and $Y_2$ are uncorrelated. However, it is easy to see that $Y_1$ and $Y_2$ are not independent by considering the joint and marginal densities at $y_1 = y_2 = 0$.

### Example 11.7.

$Y = X^2 1_{\{X \geq 0\}}$

(a) $F_Y(y) = \begin{cases} 0, & y < 0 \\ F_X(0), & y = 0 \\ F_X(\sqrt{y}), & y > 0 \end{cases}$

(b) $f_Y(y) = \begin{cases} 0, & y < 0 \\ \frac{1}{2\sqrt{y}} f_X(\sqrt{y}), & y > 0 \end{cases} + F_X(0) \delta(y)$
Figure 2.15: Relationships among univariate distributions. (After Leemis, 1986.)

Figure 20: Relationships among univariate distributions [22]
\[ X_i - \Gamma(q_i, \lambda_i) \Rightarrow \sum X_i - \Gamma \left( \sum q_i, \lambda \right). \]
\[ X - \Gamma(q, \lambda) \Rightarrow Y = aX - \Gamma \left( \frac{q}{a}, \frac{\lambda}{a} \right). \]
\[ X - \mathcal{U}(0,1) \Rightarrow -\frac{1}{\lambda} \ln(X) - \mathcal{E}(\lambda). \]
\[ X_i - \mathcal{E}(\lambda) \Rightarrow \min(X_i) - \mathcal{E} \left( \sum \lambda_i \right). \]

**Uniform:** \( \mathcal{U}(a, b) \)
\[ f_x(x) = \frac{1}{b - a} I_{[a,b]}(x) \]
\[ X - \mathcal{U}(0,1) \Rightarrow -\frac{1}{\lambda} \ln(X) - \mathcal{E}(\lambda). \]

**Exponential:** \( \mathcal{E}(\lambda) \)
\[ f_x(x) = \lambda e^{-\lambda x} I_{[0,\infty)}(x) \]

**Normal/Gaussian:** \( \mathcal{N}(m, \sigma^2) \)
\[ f_x(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \]
\[ X^2 + Y^2 \text{ with } X, Y \sim \mathcal{N}(0, 1) \]

**Chi-squared:** \( \chi^2 \)
\[ f_x(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)} I_{[0,\infty)}(x) \]
\[ X - \mathcal{N}(0, \sigma^2) \Rightarrow X^2 - \Gamma \left( \frac{1}{2}, \frac{1}{2\sigma^2} \right) \]

**Gamma:** \( \Gamma(q, \lambda) \)
\[ f_x(x) = \frac{x^{q-1}e^{-\lambda x}}{\Gamma(q)} I_{[0,\infty)}(x) \]
\[ q = \frac{n \lambda}{2} = \frac{1}{2\sigma^2} \]
\[ X_i - \Gamma(q_i, \lambda_i) \]

**Beta:**
\[ f_{\alpha, \beta}(x) = \frac{\Gamma(q_i + q_j) x^{q_i-1} (1-x)^{q_j-1}}{\Gamma(q_i) \Gamma(q_j)} I_{[0,\infty)}(x) \]
\[ \frac{X_1}{X_1 + X_2} \text{ with } X_i - \Gamma(q_i, \lambda_i) \]

**Rayleigh:**
\[ f(x) = 2ax e^{-ax^2} I_{[0,\infty)}(x) \]

\[ X^2 + Y^2 \text{ with } X, Y \sim \mathcal{N}(0, 1) \]

**Figure 21:** Another diagram demonstrating relationship among univariate distributions
11.2 MISO case

11.8. If \( X \) and \( Y \) are jointly continuous random variables with joint density \( f_{X,Y} \). The following two methods give the density of \( Z = g(X,Y) \).

- Condition on one of the variable, say \( Y = y \). Then, begin conditioned, \( Z \) is simply a function of one variable \( g(X,y) \); hence, we can use the one-variable technique to find \( f_{Z|Y}(z|y) \). Finally, \( f_Z(z) = \int f_{Z|Y}(z|y)f_Y(y)dy \).

- Directly find the joint density of the random vector \( (Z,Y) = (g(X,Y),Y) \). Observe that the Jacobian is \( \begin{pmatrix} \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \\ 0 & 1 \end{pmatrix} \). Hence, the magnitude of the determinant is \( \left| \frac{\partial g}{\partial x} \right| \).

Of course, the standard way of finding the pdf of \( Z \) is by finding the derivative of the cdf \( F_Z(z) = \int_{(x,y):x^2+y^2\leq z} f_{X,Y}(x,y)d(x,y) \). This is still good for solving specific examples. It is also a good starting point for those who haven’t learned conditional probability nor Jacobian.

Let the \( x^{(k)} \) be the solutions of \( g(x, y) = z \) for fixed \( z \) and \( y \). The first method gives

\[
f_{Z|Y}(z|y) = \sum_k f_{X|Y}(x|y) \left| \frac{\partial g}{\partial x}(x, y) \right|_{x=x^{(k)}} .
\]

Hence,

\[
f_{Z,Y}(z, y) = \sum_k f_{X,Y}(x, y) \left| \frac{\partial g}{\partial x}(x, y) \right|_{x=x^{(k)}} ,
\]

which comes out of the second method directly. Both methods then gives

\[
f_Z(z) = \int \sum_k f_{X,Y}(x, y) \left| \frac{\partial g}{\partial x}(x, y) \right|_{x=x^{(k)}} dy.
\]

The integration for a given \( z \) is only on the value of \( y \) such that there is at least a solution for \( x \) in \( z = g(x, y) \). If there is no such solution, \( f_Z(z) = 0 \). The same technique works for a function of more than one random variables \( Z = g(X_1, \ldots, X_n) \). For any \( j \in [n] \), let the \( x_j^{(k)} \) be the solutions for \( x_j \) in \( z = g(x_1, \ldots, x_n) \). Then,

\[
f_Z(z) = \int \sum_k f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \left| \frac{\partial g}{\partial x_j}(x_1, \ldots, x_n) \right|_{x_j=x_j^{(k)}} dx_{[n]\{j\}}.
\]

For the second method, we consider the random vector \( (h_r(X_1, \ldots, X_n), r \in [n]) \) where \( h_r(X_1, \ldots, X_n) = X_r \) for \( r \neq j \) and \( h_j = g \). The Jacobian is of the form

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
\frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} & \frac{\partial g}{\partial x_j} & \frac{\partial g}{\partial x_n} \\
0 & 0 & 0 & 1
\end{pmatrix}.
\]
By swapping the row with all the partial derivatives to the first row, the magnitude of the determinant is unchanged and we also end up with upper triangular matrix whose determinant is simply the product of the diagonal elements.

(a) For $Z = aX + bY$,

$$f_Z(z) = \int \frac{1}{|a|} f_{X,Y} \left( \frac{z - by}{a}, y \right) dy = \int \frac{1}{|b|} f_{X,Y} \left( x, \frac{z - ax}{b} \right) dx.$$  

- Note that Jacobian $\left( \begin{array}{c} \frac{ax + by}{y} \\ y \end{array}, \begin{array}{c} x \\ y \end{array} \right) = \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix}$.

(i) When $a = 1, b = -1$,

$$f_{X-Y}(z) = \int f_{X,Y}(z + y, y) dy = \int f_{X,Y}(x, x - z) dx.$$  

(ii) Note that when $X$ and $Y$ are independent and $a = b = 1$, we have the convolution formula

$$f_Z(z) = \int f_X(z - y) f_Y(y) dy = \int f_X(x) f_Y(z - x) dx.$$  

(b) For $Z = XY$,

$$f_Z(z) = \int f_{X,Y} \left( x, \frac{z}{x} \right) \frac{1}{|x|} dx = \int f_{X,Y} \left( \frac{z}{y}, y \right) \frac{1}{|y|} dy.$$  

[9] Ex 7.2, 7.11, 7.15]. Note that Jacobian $\left( \begin{array}{c} \frac{xy}{y} \\ \frac{x}{y} \end{array}, \begin{array}{c} y \\ x \end{array} \right) = \begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix}$.

(c) For $Z = X^2 + Y^2$,

$$f_Z(z) = \sqrt{z} \int_{-\sqrt{z}}^{\sqrt{z}} f_{X|Y} \left( \sqrt{z - y^2} \right) + f_{X|Y} \left( -\sqrt{z - y^2} \right) \frac{y}{2\sqrt{z - y^2}} dy f_Y(y) dy.$$  

[9] Ex 7.16]. Alternatively, applying (33), we have

$$f_Z(z) = \frac{1}{2} \int_0^{2\pi} f_{X,Y}(\sqrt{z} \cos \theta, \sqrt{z} \sin \theta) d\theta, \quad z > 0 \quad (29)$$  


- This can be used to show that when $X,Y \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $Z = X^2 + Y^2 \sim \mathcal{E} \left( \frac{1}{2} \right)$.

(d) For $R = \sqrt{X^2 + Y^2}$, applying (32), we have

$$f_R(r) = r \int_0^{2\pi} f_{X,Y}(r \cos \theta, r \sin \theta) d\theta, \quad r > 0 \quad (30)$$  

(e) For $Z = \frac{Y}{X}$,
$$f_Z(z) = \int \frac{|y|}{z} f_{X,Y} \left(\frac{y}{z}, y\right) dy = \int |x| f_{X,Y}(x, xz) dx.$$

Similarly, when $Z = \frac{X}{Y}$,
$$f_Z(z) = \int |y| f_{X,Y}(yz, y) dy.$$

(f) For $Z = \min(X, Y) \max(X, Y)$ where $X$ and $Y$ are strictly positive,
$$F_Z(z) = \int_0^\infty F_{Y|X}(zx|x) f_X(x) dx + \int_0^\infty F_{X|Y}(zy|y) f_Y(y) dy,$$
$$f_Z(z) = \int_0^z x f_{Y|X}(zx|x) f_X(x) dx + \int_0^z y f_{X|Y}(zy|y) f_Y(y) dy, \quad 0 < z < 1.$$

[9, Ex 7.17].

11.9 (Random sum). Let $S = \sum_{i=1}^N V_i$ where $V_i$’s are i.i.d. $\sim V$ independent of $N$.

(a) $\varphi_S(u) = \varphi_N(-i \ln(\varphi_V(u)))$.
- $\varphi_S(u) = G_N(\varphi_V(u))$
- For non-negative integer-valued summands, we have $G_S(z) = G_N(G_V(z))$

(b) $\mathbb{E}S = \mathbb{E}N \mathbb{E}V$.

(c) $\text{Var}[S] = \mathbb{E}N (\text{Var} V) + (\mathbb{E}V)^2 (\text{Var} N)$.

Remark: If $N \sim \mathcal{P}(\lambda)$, then $\varphi_S(u) = \exp(\lambda(\varphi_V(u) - 1))$, the compound Poisson distribution $\mathcal{C}P(\lambda, \mathcal{L}(V))$. Hence, the mean and variance of $\mathcal{C}P(\lambda, \mathcal{L}(V))$ are $\lambda \mathbb{E}V$ and $\lambda \mathbb{E}V^2$ respectively.

11.3 MIMO case

Definition 11.10 (Jacobian). In vector calculus, the Jacobian is shorthand for either the Jacobian matrix or its determinant, the Jacobian determinant. Let $g$ be a function from a subset $D$ of $\mathbb{R}^n$ to $\mathbb{R}^m$. If $g$ is differentiable at $z \in D$, then all partial derivatives exist at $z$ and the Jacobian matrix of $g$ at a point $z \in D$ is

$$dg(z) = \begin{pmatrix}
\frac{\partial g_1}{\partial x_1}(z) & \cdots & \frac{\partial g_1}{\partial x_n}(z) \\
\vdots & \ddots & \vdots \\
\frac{\partial g_m}{\partial x_1}(z) & \cdots & \frac{\partial g_m}{\partial x_n}(z)
\end{pmatrix} = \left(\frac{\partial g}{\partial x_1}(z), \ldots, \frac{\partial g}{\partial x_n}(z)\right).$$

Alternative notations for the Jacobian matrix are $J_g$, $\frac{\partial (g_1, \ldots, g_m)}{\partial (x_1, \ldots, x_n)}$ [7, p 242], $J_g(x)$ where the it is assumed that the Jacobian matrix is evaluated at $z = x = (x_1, \ldots, x_n)$. 

100
• Let $A$ be an $n$-dimensional "box" defined by the corners $x$ and $x + \Delta x$. The "volume" of the image $g(A)$ is $|\prod_i \Delta x_i| |\det dg(x)|$. Hence, the magnitude of the Jacobian determinant gives the ratios (scaling factor) of $n$-dimensional volumes (contents). In other words,

$$dy_1 \cdots dy_n = \left| \frac{\partial(y_1, \ldots, y_n)}{\partial(x_1, \ldots, x_n)} \right| dx_1 \cdots dx_n.$$ 

• $d(g^{-1}(y))$ is the Jacobian of the inverse transformation.

• In MATLAB, use `jacobian`.

See also (A.16).

11.11 (Jacobian formulas). Suppose $g$ is a vector-valued function of $x \in \mathbb{R}^n$, and $X$ is an $\mathbb{R}^n$-valued random vector. Define $Y = g(X)$. (Then, $Y$ is also an $\mathbb{R}^n$-valued random vector.) If $X$ has joint density $f_X$, and $g$ is a suitable invertible mapping (such that the inversion mapping theorem is applicable), then

$$f_Y(y) = \frac{1}{|\det (dg(g^{-1}(y)))|} f_X(g^{-1}(y)) = |\det (d(g^{-1})(y))| f_X(g^{-1}(y)).$$

• Note that for any matrix $A$, $\det(A) = \det(A^T)$. Hence, the formula above could tolerate the incorrect "Jacobian".

In general, let $X = (X_1, X_2, \ldots, X_n)$ be a random vector with pdf $f_X(x)$. Let $S = \{x : f_X(x) > 0\}$. Consider a new random vector $Y = (Y_1, Y_2, \ldots, Y_n)$, defined by $Y_i = g_i(X)$. Suppose that $A_0, A_1, \ldots, A_r$ form a partition of $S$ with these properties. The set $A_0$, which may be empty, satisfies $P[X \in A_0] = 0$. The transformation $Y = g(X) = (g_1(X), \ldots, g_n(X))$ is a one-to-one transformation from $A_i$ onto some common set $B$ for each $i \in [k]$. Then, for each $i$, the inverse functions from $B$ to $A_i$ can be found. Denote the $k$th inverse $x = h^{(k)}(u)$ by $x_j = h^{(k)}_{ij}(y)$. This $k$th inverse gives, for $y \in B$, the unique $x \in A_k$ such that $y = g(x)$. Assuming that the Jacobians $\det(dh^{(k)}(y))$ do not vanish identically on $B$, we have

$$f_Y(y) = \sum_{k=1}^r f_X(h^{(k)}(y)) |\det(dh^{(k)}(y))|, \quad y \in B$$


• Suppose for some $k$, $Y_k$ is some functions of other $Y_i$. In particular, suppose $Y_k = h(y_I)$ for some index set $I$ and some deterministic function $h$. Then, the $k$th row of the Jacobian matrix is a linear combination of other rows. In particular,

$$\frac{\partial y_k}{\partial x_j} = \sum_{i \in I} \left( \frac{\partial}{\partial y_i} h(y_I) \right) \frac{\partial y_i}{\partial x_j}.$$

Hence, the Jacobian determinant is 0.
11.12. Suppose $Y = g(X)$ where both $X$ and $Y$ have the same dimension, then the joint density of $X$ and $Y$ is 

$$f_{X,Y}(x, y) = f_X(x)\delta(y - g(x)).$$

- In most cases, we can show that $X$ and $Y$ are not independent, pick a point $x^{(0)}$ such that $f_X(x^{(0)}) > 0$. Pick a point $y^{(1)}$ such that $y^{(1)} \neq g(x^{(0)})$ and $f_Y(y^{(1)}) > 0$. Then, $f_{X,Y}(x^{(0)}, y^{(1)}) = 0$ but $f_X(x^{(0)}) f_Y(y^{(1)}) > 0$. Note that this technique does not always work. For example, if $g$ is a constant function which maps all values of $x$ to a constant $c$. Then, we will not be able to find $y^{(1)}$. Of course, this is to be expected because we know that a constant is always independent of other random variables.

**Example 11.13.**

(a) For $Y = AX + b$, where $A$ is a square, invertible matrix, 

$$f_Y(y) = \frac{1}{|\det A|} f_X(A^{-1}(y - b)). \quad (31)$$

(b) Transformation between Cartesian coordinates $(x, y)$ and polar coordinates $(r, \theta)$

- $x = r \cos \theta, y = r \sin \theta, r = \sqrt{x^2 + y^2}, \theta = \tan^{-1}\left(\frac{y}{x}\right)$.
- $$\begin{vmatrix}
  \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\
  \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta}
\end{vmatrix} = \begin{vmatrix}
  \cos \theta & -r \sin \theta \\
  \sin \theta & r \cos \theta
\end{vmatrix} = r. \quad (Recall \ that \ dx dy = rdrd\theta).$$

We have 

$$f_{R,\Theta}(r, \theta) = rf_{X,Y}(r \cos \theta, r \sin \theta), \quad r \geq 0 \text{ and } \theta \in (-\pi, \pi), \quad (32)$$

and 

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{x^2 + y^2}} f_{R,\Theta}\left(\sqrt{x^2 + y^2}, \tan^{-1}\left(\frac{y}{x}\right)\right).$$

If, furthermore, $\Theta$ is uniform on $(0, 2\pi)$ and independent of $R$. Then, 

$$f_{X,Y}(x, y) = \frac{1}{2\pi} \frac{1}{\sqrt{x^2 + y^2}} f_R\left(\sqrt{x^2 + y^2}\right).$$

(c) A related transformation is given by $Z = X^2 + Y^2$ and $\Theta = \tan^{-1}\left(\frac{Y}{X}\right)$. In this case, 

$$X = \sqrt{Z} \cos \Theta, Y = \sqrt{Z} \sin \Theta, \text{ and}$$

$$f_{Z,\Theta}(z, \theta) = \frac{1}{2} f_{X,Y}\left(\sqrt{z} \cos \theta, \sqrt{z} \sin \theta\right) \quad (33)$$

which gives (29).

11.14. Suppose $X, Y$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Then, $R = \sqrt{X^2 + Y^2}$ and $\Theta = \arctan\frac{Y}{X}$ are independent with $R$ being Rayleigh \(f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} U(r)\) and $\Theta$ being uniform on $[0, 2\pi]$. 

102
11.15 (Generation of a random sample of a normally distributed random variable). Let $U_1, U_2$ be i.i.d. $\mathcal{U}(0, 1)$. Then, the random variables

$$
X_1 = \sqrt{-2\ln U_1 \cos(2\pi U_2)} \\
X_2 = \sqrt{-2\ln U_1 \sin(2\pi U_2)}
$$

are i.i.d. $\mathcal{N}(0, 1)$. Moreover,

$$
Z_1 = \sqrt{-2\sigma^2 \ln U_1 \cos(2\pi U_2)} \\
Z_2 = \sqrt{-2\sigma^2 \ln U_1 \sin(2\pi U_2)}
$$

are i.i.d. $\mathcal{N}(0, \sigma^2)$.

- The idea is to generate $R$ and $\Theta$ according to (11.14) first.
- $\det(dx(u)) = -\frac{2\pi}{u_1}, u_1 = e^{-\frac{x_1^2 + x_2^2}{2}}$.

11.16. In (11.11), suppose $\dim(Y) = \dim(g) \leq \dim(X)$. To find the joint pdf of $Y$, we introduce “arbitrary” $Z = h(X)$ so that $\dim\left(\begin{array}{c} Y \\ Z \end{array}\right) = \dim(X)$.

11.4 Order Statistics

Given a sample of $n$ random variables $X_1, \ldots, X_n$, reorder them so that $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. Then, $Y_i$ is called the $i^{th}$ order statistic, sometimes also denoted $X_{i:n}, X^{(i)}, X_{(i)}, X_{n:i}, X_{i:n}$, or $X_{(i)n}$.

In particular

- $Y_1 = X_{1:n} = X_{\min}$ is the first order statistic denoting the smallest of the $X_i$’s,
- $Y_2 = X_{2:n}$ is the second order statistic denoting the second smallest of the $X_i$’s . . . , and
- $Y_n = X_{n:n} = X_{\max}$ is the $n$th order statistic denoting the largest of the $X_i$’s.

In words, the order statistics of a random sample are the sample values placed in ascending order [2, p 226]. Many results in this section can be found in [4].

11.17. Events properties:

$$
[X_{\min} \geq y] = \bigcap_i [X_i \geq y] \\
[X_{\min} > y] = \bigcap_i [X_i > y] \\
[X_{\min} \leq y] = \bigcup_i [X_i \leq y] \\
[X_{\min} < y] = \bigcup_i [X_i < y]
$$

$$
[X_{\max} \geq y] = \bigcup_i [X_i \geq y] \\
[X_{\max} > y] = \bigcup_i [X_i > y] \\
[X_{\max} \leq y] = \bigcap_i [X_i \leq y] \\
[X_{\max} < y] = \bigcap_i [X_i < y]
$$

Let $A_y = [X_{\max} \leq y], B_y = [X_{\min} > y]$. Then, $A_y = [\forall i \ X_i \leq y]$ and $B_y = [\forall i \ X_i > y]$. 

103
11.18 (Densities). Suppose the $X_i$ are absolutely continuous with joint density $f_X$. Let $S_y$ be the set of all $n!$ vector which comes from permuting the coordinates of $y$.

$$f_Y(y) = \sum_{x \in S_y} f_X(x), \quad y_1 \leq y_2 \leq \cdots \leq y_n. \quad (34)$$

To see this, note that $f_Y(y) \left( \prod_j \Delta y_j \right)$ is the probability that $Y_j$ is in the small interval of length $\Delta y_j$ around $y_j$. This probability can be calculated from finding the probability that all $X_k$ fall into the above small regions.

From the joint density, we can find the joint pdf/cdf of $Y_I$ for any $I \subset [n]$. However, in many cases, we can directly reapply the above technique to find the joint pdf of $Y_I$. This is especially useful when the $X_i$ are independent or i.i.d.

(a) The marginal density $f_{Y_k}$ can be found by approximating $f_{Y_k}(y) \Delta y$ with

$$\sum_{j=1}^{n} \sum_{I \in \binom{[n]}{\ell}} \frac{P[X_j \in [y, y + \Delta y) \text{ and } \forall i \in I, X_i \leq y \text{ and } \forall r \in (I \cup \{k\})^c, X_r > y]}{\binom{n}{\ell}}$$

where for any set $A$ and integer $\ell \in |A|$, we define $\binom{A}{\ell}$ to be the set of all $k$-element subsets of $A$. Note also that we assume $(I \cup \{k\})^c = [n] \setminus (I \cup \{k\})$.

To see this, we first choose the $X_j$ that will be $Y_k$ with value around $y$. Then, we must have $k - 1$ of the $X_i$ below $y$ and have the rest of the $X_i$ above $y$.

(b) For integers $r < s$, the joint density $f_{Y_r,Y_s}(y_r, y_s) \Delta y_r \Delta y_s$ can be approximated by the probability that two of the $X_i$ are inside small regions around $y_r$ and $y_s$. To make them $Y_r$ and $Y_s$, for the other $X_i$, $r - 1$ of them before $y_r$, $s - r - 1$ of them between $y_r$ and $y_s$, and $n - s$ of them beyond $y_s$.

- $f_{X_{\text{max}},x_{\text{min}}}(u, v) \Delta u \Delta v$ can be approximated by by

$$\sum_{(j,k) \in S} P[X_j \in [u, u+\Delta u), X_j \in [v, v+\Delta v), \text{ and } \forall i \in [n] \setminus \{j,k\}, v < X_i \leq u, \quad v \leq u,$$

where $S$ is the set of all $n(n-1)$ pairs $(j,k)$ from $[n] \times [n]$ with $j \neq k$. This is simply choosing the $j,k$ so that $X_j$ will be the maximum with value around $u$, and $X_k$ will be the minimum with value around $v$. Of course, the rest of the $X_i$ have to be between the min and max.

- When $n = 2$, we can use $(34)$ to get

$$f_{X_{\text{max}},x_{\text{min}}}(u, v) = f_{X_1,x_2}(u, v) + f_{X_1,x_2}(v, u), \quad v \leq u.$$\[Note that the joint density at point $y_I$ is 0 if the elements in $y_I$ are not arranged in the “right” order.\]

11.19 (Distribution functions). We note again the the cdf may be obtained by integration of the densities in (11.18) as well as by direct arguments valid also in the discrete case.
(a) The marginal cdf is
\[ F_{Y_k}(y) = \sum_{j=k}^{n} \sum_{I \in \binom{[n]}{j}} P[\forall i \in I, X_i \leq y \text{ and } \forall r \in [n] \setminus I, X_r > y]. \]

This is because the event \([Y_k \leq y]\) is the same as event that at least \(k\) of the \(X_i\) are \(\leq y\). In other words, let \(N(a) = \sum_{i=1}^{n} 1[X_i \leq a]\) be the number of \(X_i\) which are \(\leq a\).

\[ [Y_k \leq y] = \left[ \sum_i N(y) \geq k \right] = \bigcup_{j \geq k} \left[ \sum_i N(y) = j \right], \tag{35} \]

where the union is a disjoint union. Hence, we sum the probability that exactly \(j\) of the \(X_i\) are \(\leq y\) for \(j \geq k\). Alternatively, note that the event \([Y_k \leq y]\) can also be expressed as a disjoint union

\[ \bigcup_{j \geq k} [X_i \leq k \text{ and exactly } k - 1 \text{ of the } X_1, \ldots, X_{j-1} \text{ are } \leq y]. \]

This gives

\[ F_{Y_k}(y) = \sum_{j=k}^{n} \sum_{I \in \binom{[n]}{j}} P[X_j \leq y, \forall i \in I, X_i \leq y, \text{ and } \forall r \in [j-1] \setminus I, X_r > y]. \]

(b) For \(r < s\), Because \(Y_r \leq Y_s\), we have

\[ [Y_r \leq y_r] \cap [Y_s \leq y_s] = [Y_s \leq y_s], \quad y_s \leq y_r. \]

By (35), for \(y_r < y_s\),

\[ [Y_r \leq y_r] \cap [Y_s \leq y_s] = \left( \bigcup_{j=r}^{n} [N(y_r) = j] \right) \cap \left( \bigcup_{m=s}^{n} [N(y_s) = m] \right) \]

\[ = \bigcup_{m=s}^{n} \bigcup_{j=r}^{n} [N(y_r) = j \text{ and } N(y_s) = m], \]

where the upper limit of the second union is changed from \(n\) to \(m\) because we must have \(N(y_r) \leq N(y_s)\). Now, to have \(N(y_r) = j\) and \(N(y_s) = m\) for \(m > j\) is to put \(j\) of the \(X_i\) in \((-\infty, y_r]\), \(m - j\) of the \(X_i\) in \((y_r, y_s]\), and \(n - m\) of the \(X_i\) in \((y_s, \infty)\).

(c) Alternatively, for \(X_{\text{max}}, X_{\text{min}}\), we have

\[ F_{X_{\text{max}}, X_{\text{min}}}(u, v) = P(A_u \cap B_v^c) = P(A_u) - P(A_u \cap B_v) \]

\[ = P[\forall i X_i \leq u] - P[\forall i v < X_i \leq u] \]

where the second term is 0 when \(u < v\). So,

\[ F_{X_{\text{max}}, X_{\text{min}}}(u, v) = F_{X_1, \ldots, X_n}(u, \ldots, u) \]
when \( u < v \). When \( v \geq u \), the second term can be found by \([23]\) which gives
\[
F_{X_{\max}, X_{\min}}(u, v) = F_{X_1, \ldots, X_n}(u, \ldots, u) - \sum_{w \in S} (-1)^{i : w_i = v} F_{X_1, \ldots, X_n}(w)
\]
\[
= \sum_{w \in S \setminus \{(u, \ldots, u)\}} (-1)^{i : w_i = v+1} F_{X_1, \ldots, X_n}(w).
\]

where \( S = \{u, v\}^n \) is the set of all \( 2^n \) vertices \( w \) of the “box” \( \times_{i \in [n]} (a_i, b_i) \). The joint density is 0 for \( u < v \).

11.20. For independent \( X_i \)'s,

(a) \( f_Y(y) = \sum_{x \in S_y} \prod_{i=1}^{n} f_{X_i}(x_i) \)

(b) Two forms of marginal cdf:
\[
F_{Y_k}(y) = \sum_{j=k}^{n} \sum_{I \in \binom{[n]}{j}} \left( \prod_{i \in I} F_{X_i}(y) \right) \left( \prod_{r \in [n] \setminus I} (1 - F_{X_r}(y)) \right)
\]
\[
= \sum_{j=k}^{n} \sum_{I \in \binom{[k-1]}{j-1}} F_{X_j}(y) \left( \prod_{i \in I} F_{X_i}(y) \right) \left( \prod_{r \in [j-1] \setminus I} (1 - F_{X_r}(y)) \right)
\]

(c) \( F_{X_{\max}, X_{\min}}(u, v) = \prod_{k \in [n]} F_{X_k}(u) - \left\{ \begin{array}{ll} 0, & u \leq v \\ \prod_{k \in [n]} (F_{X_k}(u) - F_{X_k}(v)), & v < u \end{array} \right\} \)

(d) The marginal cdf is
\[
F_{Y_k}(y) = \sum_{j=k}^{n} \sum_{I \in \binom{[n]}{j}} \left( \prod_{i \in I} F_{X_i}(y) \right) \left( \prod_{r \in [n] \setminus I} (1 - F_{X_r}(y)) \right).
\]

(e) \( F_{X_{\min}}(v) = 1 - \prod_{i} (1 - F_{X_i}(v)). \)

(f) \( F_{X_{\max}}(u) = \prod_{i} F_{X_i}(u). \)

11.21. Suppose \( X_i \overset{i.i.d.}{\sim} X \) with common density \( f \) and distribution function \( F \).

(a) The joint density is given by
\[
f_Y(y) = n! f(y_1) f(y_2) \cdots f(y_n), \quad y_1 \leq y_2 \leq \cdots \leq y_n.
\]
If we define $y_0 = -\infty$, $y_{k+1} = \infty$, $n_0 = 0$, $n_{k+1} = n + 1$, then for $k \in [n]$ and $1 \leq n_1 < \ldots < n_k \leq n$, the joint density $f_{Y_{n_1}, Y_{n_2}, \ldots, Y_{n_k}}(y_{n_1}, y_{n_2}, \ldots, y_{n_k})$ is given by

$$n! \left( \prod_{j=1}^{k} f(y_{n_j}) \right) \prod_{j=1}^{k} \frac{(F(y_{n_{j+1}}) - F(y_{n_j}))^{n_{j+1} - n_j - 1}}{(n_{j+1} - n_j - 1)!}.$$  

(36)

In particular, for $r < s$, the joint density $f_{Y_r, Y_s}(y_r, y_s)$ is given by

$$\frac{n!}{(r - 1)!(s - r - 1)!(n - s)!} f(y_r) f(y_s) F^{r-1}(y_r)(F(y_s) - F(y_r))^{s-r-1}(1 - F(y_s))^{n-s}$$

[2] Theorem 5.4.6 p 230].

(b) The joint cdf $F_{Y_r, Y_s}(y_r, y_s)$ is given by

$$\sum_{m=s}^{n} \sum_{j=r}^{m} \frac{n!}{j!(m-j)!(n-m)!} (F(y_r))^j (F(y_s) - F(y_r))^{m-j} (1 - F(y_s))^{n-m}, \quad y_r < y_s.$$  

(c) $F_{X_{\text{max}}, X_{\text{min}}}(u, v) = (F(u))^n - \left\{ \begin{array}{ll} 0, & u \leq v \\ (F(u) - F(v))^n, & v < u \end{array} \right\}.$

(d) $f_{X_{\text{max}}, X_{\text{min}}}(u, v) = \left\{ \begin{array}{ll} 0, & u \leq v \\ n(n-1) f_X(u) f_X(v) (F(u) - F(v))^{n-2}, & v < u \end{array} \right\}.$

(e) Marginal cdf:

$$F_{Y_k}(y) = \sum_{j=k}^{n} \binom{n}{j} (F(y))^j (1 - F(y))^{n-j}$$

$$= \sum_{j=k}^{n} \binom{j-1}{k-1} (F(y))^k (1 - F(y))^{j-k} = (F(y))^k \sum_{m=0}^{k} \binom{k+m-1}{k-1} (1 - F(y))^m$$

$$= \frac{n!}{(k-1)!(n-k)!} \int_{0}^{F(y)} t^{k-1} (1-t)^{n-k} dt.$$  

Note that $N(y) \sim B(n, F(y))$. The last equality comes from integrating the marginal density $f_{Y_k}$ in (36) with change of variable $t = F(y)$.

(i) $F_{X_{\text{max}}}(y) = (F(y))^n$ and $f_{X_{\text{max}}}(y) = n(F(y))^{n-1} f_X(y)$.

(ii) $F_{X_{\text{min}}}(y) = 1 - (1 - F(y))^n$ and $f_{X_{\text{min}}}(y) = n(1 - F(y))^{n-1} f_X(y)$.

(f) Marginal density:

$$f_{Y_k}(y) = \frac{n!}{(k-1)!(n-k)!} (F(y))^{k-1} (1 - F(y))^{n-k} f_X(y)$$

(36)

[2] Theorem 5.4.4 p 229]

Consider small neighborhood $\Delta_y$ around $y$. To have $Y_k \in \Delta_y$, we must have exactly one of the $X_i$’s in $\Delta_y$, exactly $k-1$ of them less than $y$, and exactly $n-k$ of them greater than $y$. There are $n \binom{n-1}{k-1} = \frac{n!}{(k-1)!(n-k)!} = \frac{1}{B(k,n-k+1)}$ possible setups.
The range $R$ is defined as $R = X_{\text{max}} - X_{\text{min}}$.

(i) For $x > 0$, $f_R(x) = n(n-1) \int (F(u) - F(u-x))^{n-2} f(u-x) f(u) du$.

(ii) For $x \geq 0$, $F_R(x) = n \int (F(u) - F(u-x))^{n-1} f(u) du$.

Both pdf and cdf above are derived by first finding the distribution of the range conditioned on the value of the $X_{\text{max}} = u$.

See also [4, Sec. 2.2] and [2, Sec. 5.4].

11.22. Let $X_1, X_2, \ldots, X_n$ be a random sample from a discrete distribution with pmf $p_X(x_i) = p_i$, where $x_1 < x_2 < \ldots$ are the possible values of $X$ in ascending order. Define $P_0 = 0$ and $P_i = \sum_{k=1}^{i} p_k$, then

$$P[Y_j \leq x_i] = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

and

$$P[Y_j = x_i] = \sum_{k=j}^{n} \binom{n}{k} \left( P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right).$$

Example 11.23. If $U_1, U_2, \ldots, U_k$ are independently uniformly distributed on the interval 0 to $t_0$, then they have joint pdf

$$f_{U_k}^k(u_k) = \begin{cases} \frac{1}{t_0^k}, & 0 \leq u_i \leq t_0 \\ 0, & \text{otherwise} \end{cases}$$

The order statistics $\tau_1, \tau_2, \ldots, \tau_k$ corresponding to $U_1, U_2, \ldots, U_k$ have joint pdf

$$f_{\tau_k}^k(t_k) = \begin{cases} \frac{k!}{t_0^k}, & 0 \leq t_1 \leq t_2 \leq \cdots \leq t_k \leq t_0 \\ 0, & \text{otherwise} \end{cases}$$

Example 11.24 ($n = 2$). Suppose $U = \max(X,Y)$ and $V = \min(X,Y)$ where $X$ and $Y$ have joint cdf $F_{XY}$.

$$F_{U,V}(u,v) = \begin{cases} F_{X,Y}(u,u), & u \leq v, \\ F_{X,Y}(v,u) + F_{X,Y}(u,v) - F_{X,Y}(v,v), & u > v \end{cases},$$

$$F_U(u) = F_{X,Y}(u,u),$$

$$F_V(v) = F_X(v) - F_Y(v) - F_{X,Y}(v,v).$$

[9, Ex 7.5, 7.6]. The joint density is

$$f_{U,V}(u,v) = f_{X,Y}(u,v) + f_{X,Y}(v,u), \quad v < u.$$
The marginal densities is given by

\[ f_U (u) = \frac{\partial}{\partial x} F_{X,Y} (x, y) \bigg|_{x=u, y=u} + \frac{\partial}{\partial y} F_{X,Y} (x, y) \bigg|_{x=u, y=u} \]

\[ = \int_{-\infty}^{u} f_{X,Y} (x, u) dx + \int_{-\infty}^{u} f_{X,Y} (u, y) dy, \]

\[ f_V (v) = f_X (v) + f_Y (v) - f_U (v) \]

\[ = f_X (v) + f_Y (v) - \left( \frac{\partial}{\partial x} F_{X,Y} (x, y) \bigg|_{x=v, y=v} + \frac{\partial}{\partial y} F_{X,Y} (x, y) \bigg|_{x=v, y=v} \right) \]

\[ = f_X (v) + f_Y (v) - \left( \int_{-\infty}^{v} f_{X,Y} (x, u) dx + \int_{-\infty}^{v} f_{X,Y} (u, y) dy \right). \]

\[ = \int_{v}^{\infty} f_{X,Y} (v, y) dy + \int_{v}^{\infty} f_{X,Y} (x, v) dx \]

If, furthermore, \( X \) and \( Y \) are independent, then

\[ F_{U,V} (u, v) = \begin{cases} F_X (u) F_Y (u), & u \leq v \\ F_X (v) F_Y (v) + F_X (u) F_Y (v) - F_X (v) F_Y (v), & u > v, \end{cases} \]

\[ F_U (u) = F_X (u) F_Y (u), \]

\[ F_V (v) = F_X (v) + F_Y (v) - F_X (v) F_Y (v), \]

\[ f_U (u) = f_X (u) F_Y (u) + F_X (u) f_Y (u), \]

\[ f_V (v) = f_X (v) + f_Y (v) - f_X (v) F_Y (v) - F_X (v) f_Y (v). \]

If, furthermore, \( X \) and \( Y \) are i.i.d., then

\[ F_U (u) = F^2 (u), \]

\[ F_V (v) = 2 F (v) - F^2 (v), \]

\[ f_U (u) = 2 f (u) F (u), \]

\[ f_V (v) = 2 f (v) (1 - F (v)). \]

**11.25.** Let the \( X_i \) be i.i.d. with density \( f \) and cdf \( F \). The range \( R \) is defined as \( R = X_{\text{max}} - X_{\text{min}}. \)

\[ F_R (x) = \begin{cases} 0, & x < 0 \\ n \int (F (v) - F (v-x))^{n-1} f (v) dv, & x \geq 0 \end{cases} \]

\[ f_R (x) = \begin{cases} 0, & x < 0 \\ n (n-1) \int (F (v) - F (v-x))^{n-2} f (v-x) f (v) dv, & x > 0. \end{cases} \]

For example, when \( X_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}(0, 1), \)

\[ f_R (x) = n (n-1) x^{n-2} (1-x), \quad 0 \leq x \leq 1 \]

12 Convergences

**Definition 12.1.** A sequence of random variables \( (X_n) \) **converges pointwise** to \( X \) if for all \( \omega \in \Omega \), 
\[
\lim_{n \to \infty} X_n(\omega) = X(\omega)
\]

**Definition 12.2 (Strong Convergence).** The following statements are all equivalent conditions/notations for a sequence of random variables \( (X_n) \) to converge almost surely to a random variable \( X \):

(a) \( X_n \overset{\text{a.s.}}{\to} X \)

(i) \( X_n \to X \) a.s.

(ii) \( X_n \to X \) with probability 1.

(iii) \( X_n \to X \) w.p. 1

(iv) \( \lim_{n \to \infty} X_n = X \) a.s.

(b) \( (X_n - X) \overset{\text{a.s.}}{\to} 0 \)

(c) \( P[X_n \to X] = 1 \)

(i) \( P \left[ \left\{ \omega : \lim_{n \to \infty} X_n(\omega) = X(\omega) \right\} \right] = 1 \)

(ii) \( P \left[ \left\{ \omega : \lim_{n \to \infty} X_n(\omega) = X(\omega) \right\}^c \right] = 0 \)

(iii) \( P \left[ \left\{ \omega : \lim_{n \to \infty} X_n(\omega) \neq X(\omega) \right\} \right] = 0 \)

(iv) \( P[X_n \to X] = 0 \)

(v) \( P \left[ \left\{ \omega : \lim_{n \to \infty} |X_n(\omega) - X(\omega)| = 0 \right\} \right] = 1 \)

(d) \( \forall \varepsilon > 0 \ P \left[ \left\{ \omega : \lim_{n \to \infty} |X_n(\omega) - X(\omega)| < \varepsilon \right\} \right] = 1 \)

12.3. Properties of convergence a.s.

(a) Uniqueness: if \( X_n \overset{\text{a.s.}}{\to} X \) and \( X_n \overset{\text{a.s.}}{\to} Y \), then \( X = Y \) a.s.

(b) If \( X_n \overset{\text{a.s.}}{\to} X \) and \( Y_n \overset{\text{a.s.}}{\to} Y \), then

(i) \( X_n + Y_n \overset{\text{a.s.}}{\to} X + Y \)

(ii) \( X_n Y_n \overset{\text{a.s.}}{\to} XY \)

(c) \( g \) continuous, \( X_n \overset{\text{a.s.}}{\to} X \) \( \Rightarrow \) \( g(X_n) \overset{\text{a.s.}}{\to} g(X) \)

(d) Suppose that \( \forall \varepsilon > 0 \), \( \sum_{n=1}^{\infty} P[|X_n - X| > \varepsilon] < \infty \), then \( X_n \overset{\text{a.s.}}{\to} X \)

(e) Let \( A_1, A_2, \ldots \) be independent. Then, \( 1_{A_n} \overset{\text{a.s.}}{\to} 0 \) if and only if \( \sum_n P(A_n) < \infty \)
**Definition 12.4** (Convergence in probability). The following statements are all equivalent conditions/notations for a sequence of random variables \((X_n)\) to converge in probability to a random variable \(X\)

(a) \(X_n \xrightarrow{p} X\)
   
   (i) \(X_n \xrightarrow{P} X\)
   
   (ii) \(p \lim_{n \to \infty} X_n = X\)

(b) \((X_n - X) \xrightarrow{P} 0\)

(c) \(\forall \varepsilon > 0 \lim_{n \to \infty} P[|X_n - X| < \varepsilon] = 1\)
   
   (i) \(\forall \varepsilon > 0 \lim_{n \to \infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0\)
   
   (ii) \(\forall \varepsilon > 0 \forall \delta > 0 \exists N_\delta \in \mathbb{N} \text{ such that } \forall n \geq N_\delta \ P[|X_n - X| > \varepsilon] < \delta\)
   
   (iii) \(\forall \varepsilon > 0 \lim_{n \to \infty} P[|X_n - X| > \varepsilon] = 0\)
   
   (iv) The strict inequality between \(|X_n - X|\) and \(\varepsilon\) can be replaced by the corresponding “non-strict” version.

**12.5. Properties of convergence in probability**

(a) Uniqueness: If \(X_n \xrightarrow{P} X\) and \(X_n \xrightarrow{P} Y\), then \(X = Y\) a.s.

(b) Suppose \(X_n \xrightarrow{P} X\), \(Y_n \xrightarrow{P} Y\), and \(a_n \to a\), then
   
   (i) \(X_n + Y_n \xrightarrow{P} X + Y\)
   
   (ii) \(a_nX_n \xrightarrow{P} aX\)
   
   (iii) \(X_nY_n \xrightarrow{P} XY\)

(c) Suppose \((X_n)\) i.i.d. with distribution \(\mathcal{U}[0, \theta]\). Let \(Z_n = \max\{X_i : 1 \leq i \leq n\}\). Then, \(Z_n \xrightarrow{P} \theta\)

(d) \(g\) continuous, \(X_n \xrightarrow{P} X \Rightarrow g(X_n) \xrightarrow{P} g(X)\)
   
   (i) Suppose that \(g : \mathbb{R}^d \to \mathbb{R}\) is continuous. Then, \(\forall i \ X_{i,n} \xrightarrow{P_{n \to \infty}} X_i\) implies
   
   \[g(X_{1,n}, \ldots, X_{d,n}) \xrightarrow{P_{n \to \infty}} g(X_1, \ldots, X_d)\]

(e) Let \(g\) be a continuous function at \(c\). Then, \(X_n \xrightarrow{P} c \Rightarrow g(X_n) \xrightarrow{P} g(c)\)

(f) **Fatou’s lemma**: \(0 \leq X_n \xrightarrow{P} X \Rightarrow \liminf_{n \to \infty} \mathbb{E}X_n \geq \mathbb{E}X\)

(g) Suppose \(X_n \xrightarrow{P} X\) and \(|X_n| \leq Y\) with \(\mathbb{E}Y < \infty\), then \(\mathbb{E}X_n \to \mathbb{E}X\)
(h) Let \( A_1, A_2, \ldots \) be independent. Then, \( 1_{A_n} \xrightarrow{P} 0 \) iff \( P(A_n) \to 0 \)

(i) \( X_n \xrightarrow{P} 0 \) iff \( \exists \delta > 0 \) such that \( \forall t \in [-\delta, \delta] \) we have \( \varphi_{X_n}(t) \to 1 \)

**Definition 12.6** (Weak convergence for probability measures). Let \( P_n \) and \( P \) be probability measure on \( \mathbb{R}^d \) \((d \geq 1)\). The sequence \( P_n \) **converges weakly** to \( P \) if the sequence of real numbers \( \int g dP_n \to \int g dP \) for any \( g \) which is real-valued, continuous, and bounded on \( \mathbb{R}^d \)

12.7. Let \((X_n), X\) be \( \mathbb{R}^d \)-valued random variables with distribution functions \((F_n), F\), distributions \((\mu_n), \mu\), and ch.f. \((\varphi_n), \varphi\) respectively.

The following are equivalent conditions for a sequence of random variables \((X_n)\) to converge in distribution to a random variable \(X\)

(a) \((X_n)\) **converges in distribution** (or **in law**) to \(X\)

(i) \( X_n \to X \)
   
   i. \( X_n \overset{L}{\to} X \)
   
   ii. \( X_n \overset{D}{\to} X \)

(ii) \( F_n \to F \)

   i. \( F_{X_n} \to F_X \)

   ii. \( F_n \) converges weakly to \( F \)

   iii. \( \lim_{n \to \infty} P^{X_n}(A) = P^X(A) \) for every \( A \) of the form \( A = (-\infty, x] \) for which \( P^X\{x\} = 0 \)

(iii) \( \mu_n \Rightarrow \mu \)

   i. \( P^{X_n} \) converges weakly to \( P^X \)

(b) **Skorohod’s theorem**: \( \exists \) random variables \( Y_n \) and \( Y \) on a common probability space \((\Omega, \mathcal{F}, P)\) such that \( Y_n \overset{D}{\to} X_n, Y \overset{D}{\to} X, \) and \( Y_n \to Y \) on (the whole) \( \Omega \)

(c) \( \lim_{n \to \infty} F_n = F \) for all continuity points of \( F \)

   i. \( F_{X_n}(x) \to F_X(x) \) \( \forall x \) such that \( P[X = x] = 0 \)

(d) \( \exists \) a (countable) set \( D \) dense in \( \mathbb{R} \) such that \( F_n(x) \to F(x) \) \( \forall x \in D \)

(e) **Continuous Mapping theorem**: \( \lim_{n \to \infty} \mathbb{E}g(X_n) = \mathbb{E}g(X) \) for all \( g \) which is real-valued, continuous, and bounded on \( \mathbb{R}^d \)

   (i) \( \lim_{n \to \infty} \mathbb{E}g(X_n) = \mathbb{E}g(X) \) for all bounded real-valued function \( g \) such that \( P[X \in D_g] = 0 \) where \( D_g \) is the set of points of discontinuity of \( g \)

   (ii) \( \lim_{n \to \infty} \mathbb{E}g(X_n) = \mathbb{E}g(X) \) for all bounded Lipschitz continuous functions \( g \)

   (iii) \( \lim_{n \to \infty} \mathbb{E}g(X_n) = \mathbb{E}g(X) \) for all bounded uniformly continuous functions \( g \)
(iv) \( \lim_{n \to \infty} \mathbb{E}g(X_n) = \mathbb{E}g(X) \) for all complex-valued functions \( g \) whose real and imaginary parts are bounded and continuous

(f) \textbf{Continuity Theorem:} \( \varphi_{X_n} \to \varphi_X \)

(i) For nonnegative random variables: \( \forall s \geq 0 \ L_{X_n}(s) \to L_X(s) \) where \( L_X(s) = \mathbb{E}e^{-sX} \)

Note that there is no requirement that \((X_n)\) and \(X\) be defined on the same probability space \((\Omega, \mathcal{A}, \mathbb{P})\)

12.8. \textbf{Continuity Theorem:} Suppose \( \lim_{n \to \infty} \varphi_{n}(t) \) exists \( \forall t \); call this limit \( \varphi_{\infty} \) Furthermore, suppose \( \varphi_{\infty} \) is continuous at 0. Then there exists \( \exists \) a probability distribution \( \mu_{\infty} \) such that \( \mu_n \Rightarrow \mu_{\infty} \) and \( \varphi_{\infty} \) is the characteristic function of \( \mu_{\infty} \)

12.9. Properties of convergence in distribution

(a) If \( F_n \Rightarrow F \) and \( F_n \Rightarrow G \), then \( F = G \)

(b) Suppose \( X_n \Rightarrow X \)

(i) If \( P[X \text{ is a discontinuity point of } g] = 0 \), then \( g(X_n) \Rightarrow g(X) \)

(ii) \( \mathbb{E}g(X_n) \to \mathbb{E}g(X) \) for every bounded real-valued function \( g \) such that \( P[X \in D_g] = 0 \) where \( D_g \) is the set of points of discontinuity of \( g \)

i. \( g(X_n) \Rightarrow g(X) \) for \( g \) continuous.

(iii) If \( Y_n \overset{P}{\to} 0 \), then \( X_n + Y_n \Rightarrow X \)

(iv) If \( X_n - Y_n \overset{P}{\to} 0 \), then \( Y_n \Rightarrow X \)

(c) If \( X_n \Rightarrow a \) and \( g \text{ is continuous at } a \), then \( g(X_n) \Rightarrow g(a) \)

(d) Suppose \((\mu_n)\) is a sequence of probability measures on \( \mathbb{R} \) that are all point masses with \( \mu_n(\{\alpha_n\}) = 1 \). Then, \( \mu_n \) converges weakly to a limit \( \mu \) iff \( \alpha_n \to \alpha \); and in this case \( \mu \) is a point mass at \( \alpha \)

(e) \textbf{Scheffé’s theorem:}

(i) Suppose \( P^{X_n} \) and \( P^X \) have densities \( \delta_n \) and \( \delta \) w.r.t. the same measure \( \mu \). Then, \( \delta_n \to \delta \) \( \mu \text{-a.e.} \) implies

\begin{itemize}
  \item \( \forall B \in \mathcal{B}_{\mathbb{R}} \ P^{X_n}(E) \to P^X(E) \)
  \item \( F_{X_n} \to F_X \)
  \item \( F_{X_n} \Rightarrow F_X \)
\end{itemize}

(ii) Suppose \( g \) is bounded. Then, \( \int g(x) P^{X_n}(dx) \to \int g(x) P^X(dx) \). In Equivalent, \( \mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)] \) where the \( \mathbb{E} \) is defined with respect to appropriate \( P \).

(ii) \textbf{Remarks:}
i. For absolutely continuous random variables, $\mu$ is the Lebesgue measure, $\delta$ is the probability density function.

ii. For discrete random variables, $\mu$ is the counting measure, $\delta$ is the probability mass function.

(f) Normal r.v.

(i) Let $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$. Suppose $\mu_n \to \mu \in \mathbb{R}$ and $\sigma_n^2 \to \sigma^2 \geq 0$. Then, $X_n \Rightarrow \mathcal{N}(\mu, \sigma^2)$

(ii) Suppose that $X_n$ are normal random variables and let $X_n \Rightarrow X$. Then, (1) the mean and variance of $X_n$ converge to some limit $m$ and $\sigma^2$. (2) $X$ is normal with mean $m$ and variance $\sigma^2$

(g) $X_n \Rightarrow 0$ if and only if $\exists \delta > 0$ such that $\forall t \in [-\delta, \delta]$ we have $\varphi_{X_n}(t) \to 1$

12.10 (Convergence in distribution of products of random variables). Let $(X_{n,k})$ be a triangular array of random variables in $(0, c]$, where $c < 1$, and let $X$ be a nonnegative random variable. Assume

$$\sum_k X_{n,k}^2 \Rightarrow 0 \text{ or } \sup_k X_{n,k} \Rightarrow 0 \text{ as } n \to \infty.$$  

Then as $n \to \infty$,

$$\prod_k (1 - X_{n,k}) \Rightarrow e^{-X} \text{ if and only if } \sum_k X_{n,k} \Rightarrow X$$

[10]. See also A.4

12.11. Relationship between convergences

(a) For discrete probability space, $X_n \overset{a.s.}\to X$ if and only if $X_n \overset{P}\to X$

(b) Suppose $X_n \overset{P}\to X$, $\forall n \ |X_n| \leq Y$, $Y \in L^p$. Then, $X \in L^p$ and $X_n \overset{L^p}\to X$

(c) Suppose $X_n \overset{a.s.}\to X$, $\forall n \ |X_n| \leq Y$, $Y \in L^p$. Then, $X \in L^p$ and $X_n \overset{L^p}\to X$

(d) $X_n \overset{P}\to X \Rightarrow X_n \overset{D}\to X$

(e) If $X_n \overset{D}\to X$ and if $\exists a \in \mathbb{R}X = a$ a.s., then $X_n \overset{P}\to X$

(i) Hence, when $X = a$ a.s., $X_n \overset{P}\to X$ and $X_n \overset{D}\to X$ are equivalent.

See also Figure 22

Example 12.12.

(a) $\Omega = [0, 1]$; $P$ is uniform on $[0, 1]$. $X_n(\omega) = \begin{cases} 0, & \omega \in \left[0, 1 - \frac{1}{n^2}\right] \\ e^n, & \omega \in \left(1 - \frac{1}{n^2}, 1\right] \end{cases}$
(i) $X_n \overset{L^p}{\longrightarrow} 0$
(ii) $X_n \overset{a.s.}{\longrightarrow} 0$
(iii) $X_n \overset{P}{\longrightarrow} 0$

(b) Let $\Omega = [0,1]$ with uniform probability distribution. Define $X_n(\omega) = \omega + 1[\frac{i-1}{k}, \frac{i}{k}]$, where $k = \left[ \frac{1}{2} \left( \sqrt{1+8n} - 1 \right) \right]$ and $j = n - \frac{1}{2}k(k-1)$. The sequence of intervals $[\frac{j-1}{k}, \frac{j}{k}]$ under the indicator function is shown in Figure [23]. Let $X(\omega) = \omega$.

(i) The sequence of real numbers $(X_n(\omega))$ does not converge for any $\omega$
(ii) $X_n \overset{L^p}{\longrightarrow} X$
(iii) $X_n \overset{P}{\longrightarrow} X$

(c) $X_n = \left\{ \begin{array}{ll} \frac{1}{n}, & \text{w.p. } 1 - \frac{1}{n} \\ \frac{1}{n}, & \text{w.p. } \frac{1}{n} \end{array} \right.$

(i) $X_n \overset{P}{\longrightarrow} 0$
(ii) $X_n \overset{L^p}{\not\longrightarrow} 0$
12.1 Summation of random variables

Set $S_n = \sum_{i=1}^{n} X_i$.

12.13 (Markov’s theorem; Chebyshev’s inequality). For finite variance $(X_i)$, if $\frac{1}{n^2} \text{Var} S_n \to 0$, then $\frac{1}{n} S_n - \frac{1}{n} \mathbb{E} S_n \overset{P}{\to} 0$

- If $(X_i)$ are pairwise independent, then $\frac{1}{n^2} \text{Var} S_n = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var} X_i$ See also (12.17).

12.2 Summation of independent random variables

12.14. For independent $X_n$, the probability that $\sum_{i=1}^{n} X_i$ converges is either 0 or 1.

12.15 (Kolmogorov’s SLLN). Consider a sequence $(X_n)$ of independent random variables. If $\sum_{n} \frac{\text{Var} X_n}{n^2} < \infty$, then $\frac{1}{n} S_n - \frac{1}{n} \mathbb{E} S_n \overset{\text{a.s.}}{\to} 0$

- In particular, for independent $X_n$, if $\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} X_n \to a$ or $\mathbb{E} X_n \to a$, then $\frac{1}{n} S_n \overset{\text{a.s.}}{\to} a$

12.16. Suppose that $X_1, X_2, \ldots$ are independent and that the series $X = X_1 + X_2 + \cdots$ converges a.s. Then, $\varphi_X = \varphi_{X_1} \varphi_{X_2} \cdots$

12.17. For pairwise independent $(X_i)$ with finite variances, if $\frac{1}{n^2} \sum_{i=1}^{n} \text{Var} X_i \to 0$, then $\frac{1}{n} S_n - \frac{1}{n} \mathbb{E} S_n \overset{P}{\to} 0$

(a) **Chebyshev’s Theorem**: For pairwise independent $(X_i)$ with uniformly bounded variances, then $\frac{1}{n} S_n - \frac{1}{n} \mathbb{E} S_n \overset{P}{\to} 0$
12.3 Summation of i.i.d. random variable

Let \((X_i)\) be i.i.d. random variables.

12.18 (Chebyshev’s inequality). \(P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_n - \mathbb{E}X_1 \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} \frac{\text{Var}[X_1]}{n} \)

- The \(X_i\)’s don’t have to be independent; they only need to be pairwise uncorrelated.

12.19 (WLLN). Weak Law of Large Numbers:

(a) \(L^2\) weak law: (finite variance) Let \((X_i)\) be uncorrelated (or pairwise independent) random variables

\[\text{(i)} \quad \text{Var} S_n = \sum_{i=1}^{n} \text{Var} X_i\]

\[\text{(ii)} \quad \text{If } \frac{1}{n^2} \sum_{i=1}^{n} \text{Var} X_i \to 0, \text{ then } \frac{1}{n} S_n - \frac{1}{n} \mathbb{E}S_n \xrightarrow{P} 0\]

\[\text{(iii)} \quad \text{If } \mathbb{E}X_i = \mu \text{ and } \text{Var} X_i \leq C < \infty. \text{ Then, } \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{L^2} \mu\]

(b) Let \((X_i)\) be i.i.d. random variables with \(\mathbb{E}X_i = \mu\) and \(\text{Var} X_i = \sigma^2 < \infty\). Then,

\[\text{(i)} \quad P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_n - \mathbb{E}X_1 \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} \frac{\text{Var}[X_1]}{n}\]

\[\text{(ii)} \quad \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{L^2} \mu\]

(The fact that \(\sigma^2 < \infty\) implies \(\mu < \infty\)).

(c) \(\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{L^2} \mu\) implies \(\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mu\) which in turn imply \(\frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu\)

(d) If \(X_n\) are i.i.d. random variables such that \(\lim_{t \to \infty} tP[|X_1| > t] \to 0\), then \(\frac{1}{n} S_n - \mathbb{E}X_1^{(n)} \xrightarrow{P} 0\)

(e) Khintchine’s theorem: If \(X_n\) are i.i.d. random variables and \(\mathbb{E}|X_1| < \infty\), then \(\mathbb{E}X_1^{(n)} \xrightarrow{a.s.} \mathbb{E}X_1\) and \(\frac{1}{n} S_n \xrightarrow{P} \mathbb{E}X_1\).

- No assumption about the finiteness of variance.

12.20 (SLLN).

(a) Kolmogorov’s SLLN: Consider a sequence \((X_n)\) of independent random variables.

If \(\sum_{n} \frac{\text{Var}X_n}{n^2} < \infty\), then \(\frac{1}{n} S_n - \frac{1}{n} \mathbb{E}S_n \xrightarrow{a.s.} 0\)

- In particular, for independent \(X_n\), if \(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_n \to a\) or \(\mathbb{E}X_n \to a\), then \(\frac{1}{n} S_n \xrightarrow{a.s.} a\)
(b) **Khintchin’s SLLN**: If $X_i$’s are i.i.d. with finite mean $\mu$, then $\frac{1}{n} \sum_{i=1}^{n} X_i \overset{a.s.}{\longrightarrow} \mu$

(c) Consider the sequence $(X_n)$ of i.i.d. random variables. Suppose $\mathbb{E}X_1^+ = \infty$, then $\frac{1}{n} S_n \overset{a.s.}{\longrightarrow} \infty$

- Suppose that $X_n \geq 0$ are i.i.d. random variables and $\mathbb{E}X_n = \infty$. Then, $\frac{1}{n} S_n \overset{a.s.}{\longrightarrow} \infty$

12.21 (Relationship between LLN and the convergence of relative frequency to the probability). Consider i.i.d. $Z_i \sim Z$. Let $X_i = 1_A(Z_i)$. Then, $\frac{1}{n} S_n = \frac{1}{n} \sum_{i=1}^{n} 1_A(Z_i) = r_n(A)$, the relative frequency of an event $A$. Via LLN and appropriate conditions, $r_n(A)$ converges to $\mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^{n} 1_A(Z_i) \right] = P[Z \in A]$.

12.4 Central Limit Theorem (CLT)

Suppose that $(X_k)_{k \geq 1}$ is a sequence of i.i.d. random variables with mean $m$ and variance $0 < \sigma^2 < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$.

12.22 (Lindeberg-Lévy theorem).

(a) $\frac{S_n - mc}{\sigma \sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - m}{\sigma} \Rightarrow \mathcal{N}(0, 1)$.

(b) $\frac{S_n - mc}{\sigma \sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (X_k - m) \Rightarrow \mathcal{N}(0, \sigma)$.

To see this, let $Z_k = \frac{X_k - m}{\sigma}$ iid $\sim Z$ and $Y_n = \sum_{k=1}^{n} Z_k$. Then, $\mathbb{E}Z = 0$, $\text{Var} Z = 1$, and $\varphi_{Y_n}(t) = \left( \varphi_Z \left( \frac{t}{\sqrt{n}} \right) \right)^n$. By approximating $e^x \approx 1 + x + \frac{1}{2} x^2$. We have $\varphi_X(t) \approx 1 +jt\mathbb{E}X - \frac{1}{2} t^2 \mathbb{E}[X^2]$ (see also (24)) and

$$\varphi_{Y_n}(t) = \left( 1 - \frac{t^2}{2n} \right)^n \rightarrow e^{-\frac{t^2}{2}}.$$

- The case of Bernoulli(1/2) was derived by Abraham de Moivre around 1733. The case of Bernoulli($p$) for $0 < p < 1$ was considered by Pierre-Simon Laplace [9, p. 208].

12.23 (Approximation of densities and pmfs using the CLT). Approximate the distribution of $S_n$ by $\mathcal{N}(nm, n\sigma^2)$.

- $F_{S_n}(s) \approx \Phi \left( \frac{s-nm}{\sigma \sqrt{n}} \right)$

- $f_{S_n}(s) \approx \frac{1}{\sqrt{2\pi} \sigma \sqrt{n}} e^{-\frac{1}{2} \left( \frac{s-nm}{\sigma \sqrt{n}} \right)^2}$

- If the $X_i$ are integer-valued, then

$$P[S_n = k] = P \left[ k - \frac{1}{2} < S_n \leq k + \frac{1}{2} \right] \approx \frac{1}{\sqrt{2\pi} \sigma \sqrt{n}} e^{-\frac{1}{2} \left( \frac{k-nm}{\sigma \sqrt{n}} \right)^2}$$

The approximation is best for $s, k$ near $nm$ \[9\] p. 211.

- The approximation $n! \approx \sqrt{2\pi n^{n+\frac{1}{2}}}e^{-n}$ can be derived from approximating the density of $S_n$ when $X_i \sim \mathcal{E}(1)$. We know that $f_{S_n}(s) = \frac{s^{n-1}e^{-s}}{(n-1)!}$. Approximate the density at $s = n$, gives $(n - 1)! \approx \sqrt{2\pi n^{n-\frac{1}{2}}}e^{-n}$. Multiply through by $n$. \[9\] Ex. 5.18, p. 212
- See also normal approximation for the binomial in [14].

## 13 Conditional Probability and Expectation

### 13.1 Conditional Probability

**Definition 13.1.** Suppose conditioned on $X = x$, $Y$ has distribution $Q$. Then, we write $Y|X = x \sim Q \ [23]$ p 40. It might be clearer to write $P^{Y|X=x} = Q$.

**13.2.** Discrete random variables

(a) $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

(b) $p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x)$

(c) The law of total probability: $P[Y \in C] = \sum_x P[Y \in C|X = x]P[X = x]$.

- In particular, $p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x)$.

(d) $p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\sum_{x'} p_{Y|X}(y|x')p_X(x')}$

(e) If $Y = X + Z$ where $X$ and $Z$ are independent, then

- $p_{Y|X}(y|x) = p_Z(y - x)$
- $p_Y(y) = \sum_x p_Z(y - x)p_X(x)$
- $p_{X|Y}(x|y) = \frac{p_Z(y - x)p_X(x)}{\sum_{x'} p_Z(y - x')p_X(x')}.$

(f) The substitution law of conditional probability:

$$P[g(X, Y) = z|X = x] = P[g(x, Y) = z|X = x].$$

- When $X$ and $Y$ are independent, we can “drop the conditioning”.

### 13.3. Absolutely continuous random variables

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

(b) $F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(t|x)dt$

**13.4.** $P[(X, Y) \in A] = E[P[(X, Y) \in A|X]] = \int_A f_{X,Y}(x,y)d(x,y)$

**13.5.** \( \frac{\partial}{\partial z} F_{Y|X}(g(z, x)|x) = \frac{\partial}{\partial z} \int_{-\infty}^{g(z,x)} f_Y(y|x)dy = f_Y|X g(z, x)|x) \frac{\partial}{\partial z} g(z, x) \)
13.6. \( F_{X|Y}(x|y) = P[X \leq x | Y = y] = \lim_{\Delta \to 0} \frac{P[x \leq x, y - \Delta < Y \leq y]}{P[y - \Delta < Y \leq y]} = \mathbb{E}[1_{(-\infty, x]}(X) | Y = y]. \)

13.7. Define \( f_{X|A}(x) \) to be \( \frac{f_{X,A}(x)}{P(A)} \). Then,
\[
P[A | X = x] = \lim_{\Delta \to 0} \frac{\int_{x-\Delta}^{x} f_{X,A}(x') \, dx'}{\int_{x-\Delta}^{x} f_X(x') \, dx'} = \frac{f_{X,A}(x)}{f_X(x)} = \frac{P(A) f_{X|A}(x)}{f_X(x)}.
\]

13.8. For independent \( X_i, P[\forall i \, X_i \in B_i | \forall i \, X_i \in C_i] = \prod_i P[X_i \in B_i | X_i \in C_i]. \)

13.2 Conditional Expectation

Definition 13.9. \( \mathbb{E}[g(Y) | X = x] = \sum_y g(y)p_{Y|X}(y|x). \)

- In particular, \( \mathbb{E}[Y | X = x] = \sum_y y p_{Y|X}(y|x). \)
- Note that \( \mathbb{E}[Y | X = x] \) is a function of \( x. \)

13.10. Properties of conditional expectation:

(a) Substitution law for conditional expectation: \( \mathbb{E}[g(X, Y) | X = x] = \mathbb{E}[g(x, Y) | X = x] \)

(b) \( \mathbb{E}[h(X)g(Y) | X = x] = h(x)\mathbb{E}[g(Y) | X = x]. \)

(c) (The Rule of Iterated Expectations) \( \mathbb{E}Y = \mathbb{E}[\mathbb{E}[Y | X]]. \)

(d) Law of total probability for expectation:
\[
\mathbb{E}[g(X, Y)] = \mathbb{E}[\mathbb{E}[g(X, Y) | X]].
\]

(i) \( \mathbb{E}[g(Y)] = \mathbb{E}[\mathbb{E}[g(Y) | X]]. \)

(ii) \( F_Y(y) = \mathbb{E}[F_{Y|X}(y|x)]. \) (Take \( g(z) = 1_{(-\infty, y]}(z) \).)

(e) \( \mathbb{E}[g(X) h(Y) | X = x] = g(x) \mathbb{E}[h(Y) | X = x] \)

(f) \( \mathbb{E}[g(X) h(Y)] = \mathbb{E}[g(X) \mathbb{E}[h(Y) | X]] \)

(g) \( \mathbb{E}[X + Y | Z = z] = \mathbb{E}[X | Z = z] + \mathbb{E}[Y | Z = z] \)

(h) \( \mathbb{E}[AX | z] = A\mathbb{E}[X | z]. \)

(i) \( \mathbb{E}[(X - \mathbb{E}[X | Y])Y] = 0 \) and \( \mathbb{E}[X - \mathbb{E}[X | Y]] = 0 \)

(j) \( \min_{g(x)} \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(Y - \mathbb{E}[Y | X])^2] \) where \( g \) ranges over all functions. Hence, \( \mathbb{E}[Y | X] \) is sometimes called the regression of \( Y \) on \( X \), the “best” predictor of \( Y \) conditional on \( X. \)
Definition 13.11. The **conditional variance** is defined as
\[
\text{Var}[Y|X = x] = \int (y - m(x))^2 f(y|x) \, dy
\]
where \( m(x) = \mathbb{E}[Y|X = x] \).


(a) \( \text{Var} Y = \mathbb{E}[\text{Var}[Y|X]] + \text{Var}[\mathbb{E}[Y|X]] \).

In other words, suppose given \( X = x \), the mean and variance of \( Y \) is \( m(x) \), \( v(x) \). Then, the variance of \( Y \) is \( \text{Var} Y = \mathbb{E}[v(X)] + \text{Var}[m(X)] \). Recall that for any function \( g \), we have \( \mathbb{E}[g(Y)] = \mathbb{E}[\mathbb{E}[g(Y)|X]] \). Because \( \mathbb{E}Y \) is just a constant, say \( \mu \), we can define \( g(y) = (y - \mu)^2 \) which then implies \( \text{Var} Y = \mathbb{E}[\mathbb{E}[g(Y)|X]] \). Note, however, that \( \mathbb{E}[g(Y)|X] \) and \( \text{Var}[Y|X] \) are not the same. Suppose conditioned on \( X \), \( Y \) has distribution \( Q \) with mean \( m(x) \) and variance \( v(x) \). Then, \( v(x) = \int (y - m(x))^2 Q(dy) \). However, \( \mathbb{E}[g(Y)|X = x] = \int (y - \mu)^2 Q(dy) \); note the use of \( \mu \) in stead of \( m(x) \). Therefore, in general, \( \text{Var} Y \neq \mathbb{E}[\text{Var}[Y|X]] \).

- All three terms in the expression are nonnegative. \( \text{Var} Y \) is an upper bound for each of the terms on the RHS.

(b) Suppose \( N \perp (X, Z) \), then \( \text{Var}[X + N|Z] = \text{Var}[X|Z] + \text{Var} N \).

(c) \( \text{Var}[AX|z] = A \text{Var}[X|z]A^H \).

13.13. Suppose \( \mathbb{E}[Y|X] = X \). Then, \( \text{Cov} [X,Y] = \text{Var} X \). See also [5.20]. This is also true for \( Y = X + N \) with \( X \perp N \) and \( N \) is zero-mean noise.

Definition 13.14. \( \mu_n [Y|X] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^n | X] \)

13.15. Properties

(a) \( \mu_3 [Y] = \mathbb{E}[\mu_3 [Y|X]] + \mu_3 [\mathbb{E}[Y|X]] \)
\( \mu_4 [Y] = \mathbb{E}[\mu_4 [Y|X]] + 6 \mathbb{E} \text{Var}[Y|X] \text{Var} [\mathbb{E}[Y|X]] + \mu_4 [\mathbb{E}[Y|X]] \)

13.3 Conditional Independence

13.16. The following statements are equivalent: conditions for \( X_1, X_2, \ldots, X_n \) to be mutually independent conditioning on \( Y \) (a.s.).

(a) \( p(x^n_1|y) = \prod_{i=1}^n p(x_i|y) \).

(b) \( \forall i \in [n] \setminus \{1\} \ p(x_i | x_{i-1}^i, y) = p(x_i | y) \).

(c) \( \forall i \in [n] \ X_i \) and the vector \( (X_j)_{[n] \setminus \{i\}} \) are independent conditioning on \( Y \).

Example 13.17. Suppose \( X \) and \( Y \) are independent. Conditioned on another random variable \( Z \), it is not true in general that \( X \) and \( Y \) are still independent. See example [4.46]. Recall that \( Z = X \oplus Y \) which can be rewritten as \( Y = X \oplus Z \). Hence, when \( Z = 0 \), we must have \( Y = X \).
13.18. Suppose we know that \( f_{X|Y,Z}(x|y,z) = g(x,y) \); that is \( f_{X|Y,Z} \) does not depend on \( z \). Then, conditioned on \( Y, X \) and \( Z \) are independent. In which case,

\[
 f_{X|Y,Z}(x|y,z) = f_{X|Y}(x|y) = g(x,y).
\]

13.19. Suppose we know that \( f_{Z|V,U_1,U_2}(z|v,u_1,u_2) = f_{Z|V}(z|v) \) for all \( z, v, u_1, u_2 \), then conditioned on \( V \), we can conclude that \( Z \) and \( (U_1,U_2) \) are independent. This further implies \( Z \) and \( U_i \) are independent. Moreover,

\[
 f_{Z|V,U_1,U_2}(z|v,u_1,u_2) = f_{Z|V,U_1}(z|v,u_1) = f_{Z|V,U_2}(z|v,u_2) = f_{Z|V}(z|v).
\]

14 Real-valued Jointly Gaussian

**Definition 14.1.** Random vector \( \mathbb{R}^d \) is **jointly Gaussian** or **jointly normal** if and only if \( \forall v \in \mathbb{R}^d \), the random variable \( v^T X \) is Gaussian.

- In order for this definition to make sense when \( v = 0 \) or when \( X \) has a singular covariance matrix, we agree that any constant random variable is considered to be Gaussian.
- Of course, the mean and variance are \( v^T \mathbb{E}X \) and \( v^T \Lambda X v \), respectively.

If \( X \) is a Gaussian random vector with mean vector \( m \) and covariance matrix \( \Lambda \), we write \( X \sim \mathcal{N}(m, \Lambda) \).

14.2. Properties of jointly Gaussian random vector \( X \sim \mathcal{N}(m, \Lambda) \)

(a) \( m = \mathbb{E}X, \Lambda = \text{Cov}[X] = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T] \).

(b) \( f_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\det(\Lambda)}} e^{-\frac{1}{2} (x-m)^T \Lambda^{-1} (x-m)} \).

- To remember the form of the above formulas, both exponents have to be scalar. So, we better have \( (x-m)^T \Lambda^{-1} (x-m) \) instead of having the transpose on the last term. To make this more clear, set \( \Lambda = I \), then we must have a dot product. Note also that \( v^T A v = \sum_k \sum_\ell v_k A_{k\ell} v_\ell \).
- The above formula can be derived by starting form a random vector \( Z \) whose components are i.i.d. \( \mathcal{N}(0,1) \). Let \( X = \sum_i X_i Z + m \). Use (31) and (37).
- For \( X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(m_i, \sigma_i^2) \),

\[
 f_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}} \prod_i \sigma_i} e^{-\frac{1}{2} \sum_i \left( \frac{x - m_i}{\sigma_i} \right)^2}.
\]

In particular, if \( X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \),

\[
 f_X(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2} x^T x} \quad (37)
\]
• $(2\pi)^{n/2} \sqrt{\det(\Lambda)} = \sqrt{\det(2\pi\Lambda)}$
• The Gaussian density is constant on the “ellipsoids” centered at $m$,
  \[ \{ x \in \mathbb{R}^n : (x - m)^T C^{-1}_X (x - m) = \text{constant} \} . \]
  \[ x \in \mathbb{R}^n : (x - m)^T C^{-1}_X (x - m) = \text{constant} \}

(c) \[ \varphi_X(v) = e^{j\mu^T m - \frac{1}{2}v^T \Lambda v} = e^{j\left(\sum_i v_i E X_i \right) - \frac{1}{2} \left(\sum_i v_i \text{Cov}[X_i, X_j] \right)} \]
  • This can be derived from definition (14.1) by noting that
  \[ \varphi_X(v) = \mathbb{E} \left[ e^{jv^T X} \right] = \mathbb{E} \left[ e^{jv^T X} \right] \]

  is simply $\varphi_Y(1)$ where $Y = v^T X$ which by definition is normal.

(d) Random vector $\mathbb{R}^d$ is Jointly Gaussian if and only if $\forall v \in \mathbb{R}^d$, the random variable $v^T X$ is Gaussian.
  • Independent Gaussian random variables are jointly Gaussian.

(e) Joint normality is preserved under linear transformation: suppose $Y = AX + b$, then $Y \sim \mathcal{N}(Am + b, A\Lambda A^T)$.

(f) If $(X, Y)$ jointly Gaussian, then $X$ and $Y$ are independent if and only if $\text{Cov}[X, Y] = 0$. Hence, uncorrelated jointly Gaussian random variables are independent.

(g) Note that the joint density does not exists when the covariance matrix $\Lambda$ is singular.

(h) For i.i.d. $\mathcal{N}(\mu, \sigma^2)$:
  \[ f_{X_i^n}(x_i^n) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \| x - \mu \|^2 \right\} \]
  \[ = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu^2}{\sigma^2} \sum_{i=1}^n x_i - \frac{\mu^2}{2\sigma^2} \right\} \]

(i) Third order joint Gaussian moments are 0:
  \[ \mathbb{E} \left[ (X_i - \mathbb{E}X_i) (X_j - \mathbb{E}X_j) (X_k - \mathbb{E}X_k) \right] = 0 \quad \forall \ i, j, k \text{ not necessarily distinct.} \]
  In particular, $\mathbb{E} \left[ (X - \mathbb{E}X)^3 \right] = 0$.

(j) **Isserlis’s Theorem:** Any forth-order central moment of jointly Gaussian r.v. is expressible as the a sum of all possible products of pairs of their covariances:
  \[ \mathbb{E} \left[ (X_i - \mathbb{E}X_i) (X_j - \mathbb{E}X_j) (X_k - \mathbb{E}X_k) (X_\ell - \mathbb{E}X_\ell) \right] \]
  \[ = \text{Cov}[X_i, X_j] \text{Cov}[X_k, X_\ell] + \text{Cov}[X_i, X_k] \text{Cov}[X_j, X_\ell] + \text{Cov}[X_i, X_\ell] \text{Cov}[X_j, X_k] . \]
  Note that $\frac{1}{2} \binom{4}{2} = 3$. 

123
• In particular, \( \mathbb{E} [(X - \mathbb{E}X)^4] = 3\sigma^4 \).

(k) To generate \( \mathcal{N}(m, \Lambda) \). First, by spectral theorem, \( \Lambda = VDV^T \) where \( V \) is orthogonal matrix whose columns are eigenvectors of \( \Lambda \) and \( D \) is diagonal matrix with the eigenvalues of \( \Lambda \). The random variable we want is \( VX + m \) where \( X \sim \mathcal{N}(0, D) \).

14.3. For bivariate normal, \( f_{X,Y}(x, y) \) is

\[
\frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \exp \left\{ - \frac{\left( \frac{x - \mathbb{E}X}{\sigma_X} \right)^2 - 2\rho \frac{x - \mathbb{E}X}{\sigma_X} \frac{y - \mathbb{E}Y}{\sigma_Y} + \left( \frac{y - \mathbb{E}Y}{\sigma_Y} \right)^2}{2(1 - \rho^2)} \right\},
\]

where \( \rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1, 1] \). Here, \( x, y \in \mathbb{R} \).

- \( f_{X,Y}(x, y) = \frac{1}{\sigma_X \sigma_Y} \psi_{\rho} \left( \frac{x - m_X}{\sigma_X}, \frac{y - m_Y}{\sigma_Y} \right) \)
- \( f_{X,Y} \) is constant on ellipses of the form \( \left( \frac{x}{\sigma_X} \right)^2 + \left( \frac{y}{\sigma_Y} \right)^2 = r^2 \).
- \( \Lambda = \begin{pmatrix} \sigma_X^2 & \text{Cov}[X,Y] \\ \text{Cov}[X,Y] & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho \sigma_X \sigma_Y \\ \rho \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix} \).
- The following are equivalent:
  1. \( \rho = 0 \)
  2. \( \text{Cov}[X,Y] = 0 \)
  3. \( X \) and \( Y \) are independent.
- \( |\rho| = 1 \) if and only if \( (X - \mathbb{E}X) = k (Y - \mathbb{E}Y) \). In which case
  - \( \rho = \frac{k}{|k|} = \text{sign}(k) \)
  - \( |k| = \frac{\sigma_X}{\sigma_Y} \)
- Suppose \( f_{X,Y}(x, y) \) only depends on \( \sqrt{x^2 + y^2} \) and \( X \) and \( Y \) are independent, then \( X \) and \( Y \) are normal with zero mean and equal variance.
  - \( X|Y \sim \mathcal{N} \left( \rho \frac{\sigma_X}{\sigma_Y} (Y - m_Y) + m_X, \sigma_X^2 (1 - \rho^2) \right) \)
  - \( Y|X \sim \mathcal{N} \left( \rho \frac{\sigma_Y}{\sigma_X} (X - m_X) + m_Y, \sigma_Y^2 (1 - \rho^2) \right) \)
- The standard bivariate density is defined as
  \[
  \psi_{\rho}(u, v) = \frac{1}{2\pi \sqrt{1 - \rho^2}} e^{-\frac{1}{2(1-\rho^2)}(u^2 - 2\rho uv + v^2)}
  \]
  \[
  = \psi(u) \frac{1}{\sqrt{1 - \rho^2}} \psi \left( \frac{v - \rho u}{\sqrt{1 - \rho^2}} \right) = \frac{1}{\sqrt{1 - \rho^2}} \psi \left( \frac{u - \rho v}{\sqrt{1 - \rho^2}} \right)
  \]

124
This is the joint density of $U,V$ where $U,V \sim \mathcal{N}(0,1)$ with $\text{Cov}[U,V] = \mathbb{E}[UV] = \rho$.

- The general bivariate Gaussian pair is obtained from the transformation
  \[
  \begin{pmatrix}
  X \\
  Y
  \end{pmatrix} = \begin{pmatrix}
  \sigma_X U + m_X \\
  \sigma_Y V + m_Y
  \end{pmatrix} = \begin{pmatrix}
  \sigma_X & 0 \\
  0 & \sigma_Y
  \end{pmatrix} \begin{pmatrix}
  U \\
  V
  \end{pmatrix} + \begin{pmatrix}
  m_X \\
  m_Y
  \end{pmatrix}.
  \]

- $f_{U\mid V}(u \mid v)$ is $\mathcal{N}(\rho v, 1 - \rho^2)$. In other words, $U \mid V \sim \mathcal{N}(\rho V, 1 - \rho^2)$.

14.4 (Conditional Gaussian).

(a) Suppose $(X,Y)$ are jointly Gaussian; that is $(X,Y) \sim \mathcal{N}(\mu_X, \Lambda_X)$. Then,
\[
  f_{X \mid Y}(x \mid y) \sim \mathcal{N}\left(\frac{\mu_X + \Lambda_{XY}^{-1}(y - \mu_Y)}{\sigma_X^2 + \Lambda_{XY}^{-1}}, \Lambda_{X \mid Y}\right).
\]

(b) Suppose $(X,Y,W)$ are jointly Gaussian with $W \perp (X,Y)$. Set $V = BX + W$. Then,
\[
  V \mid y \sim \mathcal{N}(\mathbb{E}[V \mid y], \Lambda_{V \mid y}) \text{ where } \mathbb{E}[V \mid y] = B\mathbb{E}[X \mid y] + \mathbb{E}W \text{ and } \Lambda_{V \mid y} = B\Lambda_{X \mid y}B^T + \Lambda_W.
\]

15 Bayesian Detection and Estimation

Consider a pair of random vectors $\Theta$ and $Y$, where $\Theta$ is not observed, but $Y$ is observed. We know the joint distribution of the pair $(\Theta, Y)$ which is usually given in the form of the prior distribution $p_\Theta(\theta)$ and the conditional distribution $p_{Y\mid\Theta}(y \mid \theta)$. By an estimator of $\Theta$ based on $Y$, we mean a function $g$ such that $\hat{\Theta}(Y) = g(Y)$ is our estimate or “guess” of the value of $\Theta$.

15.1 (Orthogonality Principle). Let $\mathcal{D}$ be a collection of random vectors with the same dimension as $\Theta$. For a random vector $Z$, suppose that
\[
  \forall X \in \mathcal{D}, \quad Z - X \in \mathcal{D}.
\]
If
\[
  \mathbb{E}[X^T(\Theta - Z)] = 0, \quad \forall X \in \mathcal{D},
\]
then
\[
  \mathbb{E}[|\Theta - X|^2] = \mathbb{E}[|\Theta - Z|^2] + \mathbb{E}[|Z - X|^2] + 2 \mathbb{E}\left[\sum_{X \in \mathcal{D}} (Z - X)^T(\Theta - Z), \left|\overset{\rightarrow}{\in \mathcal{D}}\right| = 0\right] = 0
\]
which implies
\[
  \mathbb{E}[|\Theta - Z|^2] \leq \mathbb{E}[|\Theta - X|^2] \quad \forall X \in \mathcal{D}.
\]
• If $D$ is a subspace and $Z \in D$, then (38) is automatically satisfied.
• (39) says that the vector $\Theta - Z$ is orthogonal to all vectors in $D$.

**Example 15.2.** Suppose $\Theta$ and $N$ are independent Poisson random variables with respective parameters $\lambda$ and $\mu$. Let $Z = \Theta + N$.

- $Y$ is $P(\lambda + \mu)$
- Conditioned on $Y = y$, $\Theta$ is $B\left(y, \frac{\lambda}{\lambda + \mu}\right)$.
- $\hat{\Theta}_{\text{MMSE}}(Y) = \mathbb{E}[\Theta|Y] = \frac{\lambda Y}{\lambda + \mu}$.
- $\text{Var}[\Theta|Y] = Y \frac{\lambda \mu}{(\lambda + \mu)^2}$, and $\text{MSE} = \mathbb{E}[\text{Var}[\Theta|Y]] = \frac{\lambda \mu}{\lambda + \mu} < \text{Var}Y = \lambda + \mu$.

See also [7, Q 15.17].

15.3 (Weighted Error). Suppose we define the error by $E = (\Theta - \hat{\Theta}(Y))^T W (\Theta - \hat{\Theta}(Y))$ for some positive definite matrix $W$. (Note that the usual MSE use $W = I$.) The MSE $\mathbb{E}E$ is uniquely minimized by the MMSE estimator $\hat{\Theta}(Y) = \mathbb{E}[\Theta|Y]$. The resulting MSE is $\mathbb{E}[(\Theta - \mathbb{E}[\Theta|Y])^T W (\Theta - \mathbb{E}[\Theta|Y])]$.

In fact, for any function, $g(Y)$, the conditional weight error $\mathbb{E}\left[(\Theta - g(Y))^T W (\Theta - g(Y))\right|Y]$ is given by

$$\mathbb{E}\left[(\Theta - \mathbb{E}[\Theta|Y])^T W (\Theta - \mathbb{E}[\Theta|Y])\right|Y] + (\mathbb{E}[\Theta|Y] - g(Y))^T W (\mathbb{E}[\Theta|Y] - g(Y)).$$

Hence, for each $Y$, it is minimized by having $g(Y) = \mathbb{E}[\Theta|Y]$.

15.4 (Linear minimum mean-squared-error estimator). A linear MMSE estimator $\hat{\Theta}_{\text{LMMSE}} = g_{\text{LMMSE}}(Y)$ minimizes the MSE $\mathbb{E}[(\Theta - g(Y))^2]$ among all affine estimators of the form $g(y) = Ay + b$.

(a) It is sometimes called Wiener filters.

(b) The scalar linear (affine) MMSE estimator is given by

$$\hat{\Theta}_{\text{LMMSE}}(Y) = \mathbb{E}\Theta + \frac{\text{Cov}[Y, \Theta]}{\text{Var}Y} (Y - \mathbb{E}Y).$$

• To see this in Hilbert space, note that we want the orthogonal projection of $\Theta$ onto the subspace spanned by two elements: $Y$ and $1$. The orthogonal basis of the subspace is $\{1, Y - \mathbb{E}Y\}$. Hence, the orthogonal projection is

$$\frac{\langle \Theta, 1 \rangle}{\langle 1, 1 \rangle} + \frac{\langle \Theta, Y - \mathbb{E}Y \rangle}{\langle Y - \mathbb{E}Y, Y - \mathbb{E}Y \rangle} (Y - \mathbb{E}Y).$$

• The above discussion suggest alternative ways of arriving at the LMMSE by finding $a, b$ in $\hat{\Theta}(Y) = aY + b$ such that the error $E = \Theta - \hat{\Theta}(Y)$ is orthogonal to both $1$ and $Y - \mathbb{E}Y$. The condition $\langle E, 1 \rangle = 0$ requires $\hat{\Theta}(Y)$ to be unbiased. The condition $\langle E, Y - \mathbb{E}Y \rangle = 0$ gives $a = \frac{\text{Cov}[Y, \Theta]}{\text{Var}Y}$. 

126
(c) The vector linear (affine) MMSE estimator is given by
\[ \hat{\Theta}_{\text{LMMSE}}(Y) = E\Theta + \Sigma_{\Theta Y} \Sigma_Y^{-1} (Y - EY) \]
and
\[ \text{MMSE} = \text{Cov} \left[ \Theta - \hat{\Theta}_{\text{LMMSE}}(Y) \right] = \Sigma_{\Theta} - \Sigma_{\Theta Y} \Sigma_Y^{-1} \Sigma_{Y\Theta}. \]
In fact, the optimal choice of \( A \) is any solution of
\[ A \Sigma_Y = \Sigma_{\Theta Y}. \]
In which case,
\[ \text{Cov} \left[ \Theta - \hat{\Theta}_{\text{LMMSE}}(Y) \right] = \Sigma_{\Theta} - A \Sigma_{Y\Theta} - A \Sigma_{Y} A^T + A \Sigma_{Y} A^T = \Sigma_{\Theta} - A \Sigma_{Y} A^T. \]
When \( \Sigma_Y \) is invertible, \( A = \Sigma_{\Theta Y} \Sigma_Y^{-1} \). When \( \Sigma_Y \) is singular, see [9, Q.8.38, p. 359].

- The MSE can be rewrite as
  \[ E \left[ (\Theta - E\Theta - A(Y - EY))^2 \right] + |E\Theta - A\Theta - b|^2, \]
  which show that the optimal choice of \( b \) is \( b = E\Theta - A\Theta \). This is the \( b \) which makes the estimator unbiased.

- Fix \( A \). Let \( \tilde{\Theta} = \Theta - E\Theta \) and \( \tilde{Y} = Y - EY \). Suppose for any matrix \( B \), we have
  \[ E \left[ (B\tilde{Y})^T (\tilde{\Theta} - A\tilde{Y}) \right] = 0. \]
  if and only if, for all matrix \( B \),
  \[ E \left[ |\tilde{\Theta} - A\tilde{Y}|^2 \right] \leq E \left[ |\tilde{\Theta} - B\tilde{Y}|^2 \right]. \]
In which case,
\[ E \left[ |\tilde{\Theta} - B\tilde{Y}|^2 \right] = E \left[ |\tilde{\Theta} - A\tilde{Y}|^2 \right] + E \left[ |(A - B)\tilde{Y}|^2 \right]. \]

- Additive Noise in 1-D: \( Y = \Theta + N \) where \( \text{Cov} [\Theta, N] = 0. \)
\[ \hat{\Theta}_{\text{LMMSE}}(Y) = E\Theta + \frac{\text{Cov} [\Theta, Y]}{\text{Var} Y} (Y - EY) \]
\[ = E\Theta + \frac{\text{Var} \Theta}{\text{Var} \Theta + \text{Var} N} (Y - (E\Theta + E\Theta)) \]
\[ = E\Theta + \frac{\text{SNR}}{1 + \text{SNR}} (Y - (E\Theta + E\Theta)). \]
and
\[ \text{MMSE} = \text{Var} \Theta - \frac{\text{Cov}^2 [\Theta, Y]}{\text{Var} Y} = \frac{\text{Var} \Theta \text{Var} N}{\text{Var} \Theta + \text{Var} N}. \]
A Math Review

A.1 Inequalities

A.1. By definition,
\[ \sum_{n=1}^{\infty} a_n = \sum_{n \in \mathbb{N}} a_n = \lim_{N \to \infty} \sum_{n=1}^{N} a_n \]
\[ \prod_{n=1}^{\infty} a_n = \prod_{n \in \mathbb{N}} a_n = \lim_{N \to \infty} \prod_{n=1}^{N} a_n. \]

A.2. Inequalities involving exponential and logarithm.

(a) For any \( x \),
\[ e^x \leq 1 + x \]
with equality if and only if \( x = 0 \).

(b) If we consider \( x > -1 \), then we have \( \ln(x+1) \leq x \). If we replace \( x+1 \) by \( x \), then we have \( \ln(x) \leq x - 1 \) for \( x > 0 \). If we replace \( x \) by \( \frac{1}{x} \), we have \( \ln(x) \geq 1 - \frac{1}{x} \). This gives the fundamental inequalities of information theory:
\[ 1 - \frac{1}{x} \leq \ln(x) \leq x - 1 \text{ for } x > 0 \]
with equality if and only if \( x = 1 \). Alternative forms are listed below.

(i) For \( x > -1 \), \( \frac{x}{1+x} \leq \ln(1+x) < x \) with equality if and only if \( x = 0 \).

(ii) For \( x < 1 \), \( x \leq -\ln(1-x) \leq \frac{x}{1-x} \) with equality if and only if \( x = 0 \).

A.3. For \( |x| \leq 0.5 \), we have
\[ e^{x-x^2} \leq 1 + x \leq e^x. \]
This is because
\[ x - x^2 \leq \ln(1+x) \leq x, \]
which is semi-proved by the plot in Figure 24.

A.4. Consider a triangular array of real numbers \((x_{n,k})\). Suppose (i) \( \sum_{k=1}^{r_n} x_{n,k} \to x \) and (ii) \( \sum_{k=1}^{r_n} x_{n,k}^2 \to 0 \). Then,
\[ \prod_{k=1}^{r_n} (1 + x_{n,k}) \to e^x. \]
Moreover, suppose the sum \( \sum_{k=1}^{r_n} |x_{n,k}| \) converges as \( n \to \infty \) (which automatically implies that condition (i) is true for some \( x \)). Then, condition (ii) is equivalent to condition (iii) where condition (iii) is the requirement that \( \max_{k \in [r_n]} |x_{k,n}| \to 0 \) as \( n \to \infty \).
Proof. When $n$ is large enough, conditions (ii) and (iii) each implies that $|x_{n,k}| \leq 0.5$. (For (ii), note that we can find $n$ large enough such that $|x_{n,k}|^2 \leq \sum_k x_{n,k}^2 \leq 0.5^2$.) Hence, we can apply (A.3) and get

$$\sum_{k=1}^{\frac{r}{e}} x_{n,k} - \sum_{k=1}^{\frac{r}{e}} x_{n,k}^2 \leq \prod_{k=1}^{\frac{r}{e}} (1 + x_{n,k}) \leq e^{\sum_{k=1}^{\frac{r}{e}} x_{n,k}}.$$ (42)

Suppose $\sum_{k=1}^{\frac{r}{e}} |x_{n,k}| \to x_0$. To show that (iii) implies (ii), let $a_n = \max_{k \in [r_n]} |x_{k,n}|$. Then,

$$0 \leq \sum_{k=1}^{\frac{r}{e}} x_{n,k}^2 \leq a_n \sum_{k=1}^{\frac{r}{e}} |x_{n,k}| \to 0 \times x_0 = 0.$$

On the other hand, suppose we have (ii). Given any $\varepsilon > 0$, by (ii), $\exists n_0$ such that $\forall n \geq n_0$, $\sum_{k=1}^{\frac{r}{e}} x_{n,k}^2 \leq \varepsilon^2$. Hence, for any $k$, $x_{n,k} \leq \sum_{k=1}^{\frac{r}{e}} x_{n,k}^2 \leq \varepsilon^2$ and hence $|x_{n,k}| \leq \varepsilon$ which implies $a_n \leq \varepsilon$.

Note that when the $x_{k,n}$ are non-negative, condition (i) already implies that the sum $\sum_{k=1}^{\frac{r}{e}} |x_{n,k}|$ converges as $n \to \infty$. Alternative versions of A.4 are as followed.

(a) Suppose (ii) $\sum_{k=1}^{\frac{r}{e}} x_{n,k}^2 \to 0$ as $n \to \infty$. Then, as $n \to \infty$ we have

$$\prod_{k=1}^{\frac{r}{e}} (1 + x_{n,k}) \to e^x \quad \text{if and only if} \quad \sum_{k=1}^{\frac{r}{e}} x_{n,k} \to x.$$ (43)

Proof. We already know from A.4 that the RHS of (43) implies the LHS. Also, condition (ii) allows the use of (A.3) which implies

$$\prod_{k=1}^{\frac{r}{e}} (1 + x_{n,k}) \leq e^{\sum_{k=1}^{\frac{r}{e}} x_{n,k}} \leq e^{\sum_{k=1}^{\frac{r}{e}} x_{n,k}^2} \prod_{k=1}^{\frac{r}{e}} (1 + x_{n,k}).$$ (42b)
(b) Suppose the \( x_{n,k} \) are nonnegative and (iii) \( a_n \to 0 \) as \( n \to \infty \). Then, as \( n \to \infty \) we have

\[
\prod_{k=1}^{r_n} (1 - x_{n,k}) \to e^{-x} \quad \text{if and only if} \quad \sum_{k=1}^{r_n} x_{n,k} \to x.
\]  

(44)

Proof. We already know from A.4 that the RHS of (43) implies the LHS. Also, condition (iii) allows the use of A.3 which implies

\[
\prod_{k=1}^{r_n} (1 - x_{n,k}) \leq e^{-r_n \sum_{k=1}^{r_n} x_{n,k}^2} \prod_{k=1}^{r_n} (1 - x_{n,k}).
\]

(42c)

Furthermore, by (41), we have

\[
\sum_{k=1}^{r_n} x_{n,k}^2 \leq a_n \left( -\sum_{k=1}^{r_n} \ln (1 - x_{n,k}) \right) \to 0 \times x = 0.
\]

A.5. Let \( \alpha_i \) and \( \beta_i \) be complex numbers with \( |\alpha_i| \leq 1 \) and \( |\beta_i| \leq 1 \). Then,

\[
\left| \prod_{i=1}^{m} \alpha_i - \prod_{i=1}^{m} \beta_i \right| \leq \sum_{i=1}^{m} |\alpha_i - \beta_i|.
\]

In particular, \( |\alpha^m - \beta^m| \leq m |\alpha - \beta| \).

A.6. Suppose \( \lim_{n \to \infty} a_n = a \). Then \( \lim_{n \to \infty} (1 - \frac{a_n}{n})^n = e^{-a} \) [9, p 584].

Proof. Use [A.4] with \( r_n = n \), \( x_{n,k} = -\frac{a_n}{n} \). Then, \( \sum_{k=1}^{n} x_{n,k} = -a_n \to -a \) and \( \sum_{k=1}^{n} x_{n,k}^2 = a_n \frac{1}{n} \to a \cdot 0 = 0 \).

Alternatively, from L'Hôpital's rule, \( \lim_{n \to \infty} (1 - \frac{a_n}{n})^n = e^{-a} \) (See also [19, Theorem 3.31, p 64]). This gives a direct proof for the case when \( a > 0 \). For \( n \) large enough, note that both \( |1 - \frac{a_n}{n}| \) and \( |1 - \frac{a}{n}| \) are \( \leq 1 \) where we need \( a > 0 \) here. Applying (A.5), we get

\[
|(1 - \frac{a_n}{n})^n - (1 - \frac{a}{n})^n| \leq |a_n - a| \to 0.
\]

For \( a < 0 \), we use the fact that, for \( b_n \to b > 0 \), (1) \( (1 + \frac{b_n}{n})^{-1} = ((1 + \frac{b}{n})^n)^{-1} \to e^{-b} \) and (2) for \( n \) large enough, both \( |(1 + \frac{b}{n})^{-1}| \) and \( |(1 + \frac{b_n}{n})^{-1}| \) are \( \leq 1 \) and hence

\[
\left| \left( 1 + \frac{b_n}{n} \right)^{-1} \right|^n - \left( 1 + \frac{b}{n} \right)^{-1} \right|^n \leq \frac{|b_n - b|}{(1 + \frac{b_n}{n}) (1 + \frac{b}{n})} \to 0.
\]
A.2 Summations

A.7. Basic formulas:

(a) \( \sum_{k=0}^{n} k = \frac{n(n+1)}{2} \)

(b) \( \sum_{k=0}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} = \frac{1}{6} (2n^3 + 3n^2 + n) \)

(c) \( \sum_{k=0}^{n} k^3 = \left( \sum_{k=0}^{n} k \right)^2 = \frac{1}{4} n^2 (n+1)^2 = \frac{1}{4} (n^4 + 2n^3 + n^2) \)

A nicer formula is given by

\[
\sum_{k=1}^{n} k (k+1) \cdots (k+d) = \frac{1}{d+2} n (n+1) \cdots (n+d+1) \quad (45)
\]

A.8. Let \( g(n) = \sum_{k=0}^{n} h(k) \) where \( h \) is a polynomial of degree \( d \). Then, \( g \) is a polynomial of degree \( d+1 \); that is \( g(n) = \sum_{m=1}^{d+1} a_m x^m \).

- To find the coefficients \( a_m \), evaluate \( g(n) \) for \( n = 1, 2, \ldots, d+1 \). Note that the case when \( n = 0 \) gives \( a_0 = 0 \) and hence the sum starts with \( m = 1 \).

- Alternative, first express \( h(k) \) in terms of summation of polynomials:

\[
h(k) = \left( \sum_{i=0}^{d-1} b_i k (k+1) \cdots (k+i) \right) + c. \quad (46)
\]

To do this, substitute \( k = 0, -1, -2, \ldots, -(d-1) \).

- \( k^3 = k (k+1) (k+2) - 3k (k+1) + k \)

Then, to get \( g(n) \), use (45).

A.9. Geometric Sums:

(a) \( \sum_{i=0}^{\infty} \rho^i = \frac{1}{1-\rho} \) for \( |\rho| < 1 \)

(b) \( \sum_{i=k}^{\infty} \rho^i = \frac{\rho^k}{1-\rho} \)

(c) \( \sum_{i=a}^{b} \rho^i = \frac{\rho^a - \rho^{b+1}}{1-\rho} \)

(d) \( \sum_{i=0}^{\infty} i \rho^i = \frac{\rho}{(1-\rho)^2} \)

131
(e) \( \sum_{i=a}^{b} i \rho^i = \frac{\rho^{b+1} (b \rho - b - 1) - \rho^n (a \rho - a - \rho)}{1 - \rho} \)

(f) \( \sum_{i=k}^{\infty} i \rho^i = \frac{k \rho^k}{1 - \rho} + \frac{\rho^{k+1}}{1 - \rho} \)

(g) \( \sum_{i=0}^{\infty} i^2 \rho^i = \frac{\rho + \rho^2}{(1 - \rho)^3} \)

A.10. Double Sums:

(a) \( \left( \sum_{i=1}^{n} a_i \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \)

(b) \( \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} f(i, j) = \sum_{i=1}^{\infty} \sum_{j=1}^{i} f(i, j) = \sum_{(i,j)} 1 \{ i \geq j \} f(i, j) \)

A.11. Exponential Sums:

- \( e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots \)

- \( \lambda e^\lambda + e^\lambda = 1 + 2\lambda + 3\frac{\lambda^2}{2!} + 4\frac{\lambda^3}{3!} + \ldots = \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} \)

A.12. Suppose \( h \) is a polynomial of degree \( d \), then

\[ \sum_{k=0}^{\infty} h(k) \frac{\lambda^k}{k!} = g(\lambda) e^\lambda, \]

where \( g \) is another polynomial of the same degree. For example,

\[ \sum_{k=0}^{\infty} k^3 \frac{\lambda^k}{k!} = \left( \lambda^3 + 3\lambda^2 + \lambda \right) e^\lambda. \quad (47) \]

This result can be obtained by several techniques.

(a) Start with \( e^\lambda = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \). Then, we have

\[ \sum_{k=0}^{\infty} k^3 \frac{\lambda^k}{k!} = \lambda \frac{d}{d\lambda} \left( \lambda \frac{d}{d\lambda} \left( \lambda \frac{d}{d\lambda} e^\lambda \right) \right). \]

(b) We can expand

\[ k^3 = k (k - 1) (k - 2) + 3k (k - 1) + k. \quad (48) \]

similar to (46). Now note that

\[ \sum_{k=0}^{\infty} k (k - 1) \cdots (k - (\ell - 1)) \frac{\lambda^k}{k!} = \lambda^\ell e^\lambda. \]

Therefore, the coefficients of the terms in (48) directly becomes the coefficients in (47).
A.13. **Zeta function** $\xi (s)$ is defined for any complex number $s$ with $\text{Re} \{s\} > 1$ by the Dirichlet series: $\xi (s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$

- For real-valued nonnegative $x$
  
  (a) $\xi (x)$ converges for $x > 1$
  (b) $\xi (x)$ diverges for $0 < x \leq 1$

[Q2.48 p 105].

- $\xi (1) = \infty$ corresponds to harmonic series.

A.14. **Abel’s theorem**: Let $a = (a_i : i \in \mathbb{N})$ be any sequence of real or complex numbers and let

$$G_a(z) = \sum_{i=0}^{\infty} a_i z^i,$$

be the power series with coefficients $a$. Suppose that the series $\sum_{i=0}^{\infty} a_i$ converges. Then,

$$\lim_{z \to 1^{-}} G_a(z) = \sum_{i=0}^{\infty} a_i. \quad (49)$$

In the special case where all the coefficients $a_i$ are nonnegative real numbers, then the above formula (49) holds also when the series $\sum_{i=0}^{\infty} a_i$ does not converge. I.e. in that case both sides of the formula equal $+\infty$.

A.3 Calculus

A.3.1 Derivatives

A.15. Basic Formulas

- (a) $\frac{d}{dx} a^u = a^u \ln a \frac{du}{dx}$
- (b) $\frac{d}{dx} \log_a u = \frac{\log_a e}{u} \frac{du}{dx}, \quad a \neq 0, 1$
- (c) Derivatives of the products: Suppose $f(x) = g(x) h(x)$, then

$$f^{(n)} (x) = \sum_{k=0}^{n} \binom{n}{k} g^{(n-k)} (x) h^{(k)} (x).$$

In fact,

$$\frac{d^n}{dt^n} \prod_{i=1}^{r} f_i(t) = \sum_{n_1 + \ldots + n_r = n} \frac{n!}{n_1! n_2! \ldots n_r!} \prod_{i=1}^{r} \frac{d^{n_i}}{dt^{n_i}} f_i(t).$$
Definition A.16 (Jacobian). In vector calculus, the **Jacobian** is shorthand for either the **Jacobian matrix** or its determinant, the **Jacobian determinant**. Let \( g \) be a function from a subset \( D \) of \( \mathbb{R}^n \) to \( \mathbb{R}^m \). If \( g \) is differentiable at \( z \in D \), then all partial derivatives exist at \( z \) and the Jacobian matrix of \( g \) at a point \( z \in D \) is

\[
dg (z) = \begin{pmatrix}
\frac{\partial g_1}{\partial x_1} (z) & \cdots & \frac{\partial g_1}{\partial x_n} (z) \\
\vdots & \ddots & \vdots \\
\frac{\partial g_m}{\partial x_1} (z) & \cdots & \frac{\partial g_m}{\partial x_n} (z)
\end{pmatrix} = \left( \frac{\partial g}{\partial x_1} (z), \ldots, \frac{\partial g}{\partial x_n} (z) \right).
\]

Alternative notations for the Jacobian matrix are \( J \), \( \frac{\partial (g_1, \ldots, g_n)}{\partial (x_1, \ldots, x_n)} \) [7, p 242], and \( J_g (x) \) where it is assumed that the Jacobian matrix is evaluated at \( z = x = (x_1, \ldots, x_n) \).

- **Linear approximation** around \( z \):

\[
g(x) \approx dg(z)(x-z) + g(z). \tag{50}
\]

The function \( \ell \) is the **linearization** [21] of \( g \) at a point \( z \).

- Let \( g : D \to \mathbb{R}^m \) with open \( D \subset \mathbb{R}^n \), \( y \in D \). If \( \forall k \forall j \) partial derivatives \( \frac{\partial g_k}{\partial x_j} \) exists in a neighborhood of \( z \) and continuous at \( z \), then \( g \) is differentiable at \( z \). And \( dg \) is continuous at \( z \).

- Let \( A \) be an \( n \)-dimensional “box” defined by the corners \( x \) and \( x + \Delta x \). The “volume” of the image \( g(A) \) is \( (\prod_i \Delta x_i) | \det dg(x) | \). Hence, the magnitude of the Jacobian determinant gives the ratios (scaling factor) of \( n \)-dimensional volumes (contents). In other words,

\[
dy_1 \cdots dy_n = \left| \frac{\partial (y_1, \ldots, y_n)}{\partial (x_1, \ldots, x_n)} \right| dx_1 \cdots dx_n.
\]

- \( d(g^{-1}(y)) \) is the Jacobian of the inverse transformation.

- In MATLAB, use `jacobian`.

- Change of variable: Let \( g \) be a continuous differentiable map of the open set \( U \) onto \( V \). Suppose that \( g \) is one-to-one and that \( \det (dg(x)) \neq 0 \) for all \( x \).

\[
\int_U h(g(x)) |\det (dg(x))| \, dx = \int_V h(y) \, dy.
\]

A.17. When \( m = 1 \), we have a scalar function, and we can talk about the gradient vector. The **gradient** (or gradient vector field) of a scalar function \( f(x) \) with respect to a vector variable \( x = (x_1, \ldots, x_n) \) is

\[
\nabla_x f (z) = \begin{pmatrix}
\frac{\partial f}{\partial x_1} (z) \\
\vdots \\
\frac{\partial f}{\partial x_n} (z)
\end{pmatrix} = (df(z))^T.
\]
(a) The RHS of the linear approximation \([50]\) characterizes the **tangent hyperplane**

\[ L(x) = dg(z)(x - z) + g(z) \]

at the point \(z\). The **level surface** that pass through the point \(z\) is given by

\[ \{ x : g(x) = g(z) \}. \]

Using the linear approximation \([50]\), we see that around the point \(z\), the point on the level surface must satisfy

\[ dg(z)(x - z) \approx 0. \]

Hence, the **tangent plane** at \(z\) for the level surface is given by

\[ \{ x : dg(z)(x - z) = 0 \}. \]

Note also that the gradient vector \(\nabla g(z) = (dg(z))^T\) is perpendicular to the tangent plane through \(z\): for any two point \(x^{(1)}, x^{(2)}\) on the tangent plane though \(z\),

\[ dg(z)(x^{(2)} - x^{(1)}) = 0. \]

Hence, the gradient vector at \(z\) is perpendicular to the level surface at \(z\).

(b) When \(n = 2\), given a point \(P = (x_0, y_0)\) we have a **tangent plane**

\[ L(x, y) = \left(\frac{\partial g}{\partial x} (P)(x - x_0) + \frac{\partial g}{\partial y} (P)(y - y_0)\right) + g(P) \]

through \(P\). The **level curve** that pass through the point \(P\) is given by

\[ \{ (x, y) : g(x, y) = g(P) \}. \]

Approximating \(g\) by \(L\), we find that the **tangent line** at \(P\) for the level curve is given by

\[ \left\{ (x, y) : \frac{\partial g}{\partial x} (P)(x - x_0) + \frac{\partial g}{\partial y} (P)(y - y_0) = 0 \right\}. \]

This line is perpendicular to the gradient vector.

(c) When \(n = 3\), given a point \(P = (x_0, y_0, z_0)\), we think about the **level surface**

\( S = \{ (x, y, z) : f(x, y, z) = c \} \) where \(c = f(P)\). The **tangent plane** at the point \(P\) on the level surface \(S\) is given by

\[ \left\{ (x, y, z) : \frac{\partial g}{\partial x} (P)(x - x_0) + \frac{\partial g}{\partial y} (P)(y - y_0) + \frac{\partial g}{\partial z} (P)(z - z_0) = 0 \right\}. \]

This plane is perpendicular to the gradient vector.

A.18. For gradient, if the argument is row vector, then,

\[ \nabla_{\theta} \left( f^T (\theta) \right) = \begin{pmatrix} \frac{\partial f_1}{\partial \theta_1} & \frac{\partial f_2}{\partial \theta_1} & \cdots & \frac{\partial f_m}{\partial \theta_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial \theta_n} & \frac{\partial f_2}{\partial \theta_n} & \cdots & \frac{\partial f_m}{\partial \theta_n} \end{pmatrix}. \]
Definition A.19. Given a scalar-valued function $f$, the **Hessian matrix** is the square matrix of second partial derivatives

$$
\nabla_{\theta}^2 f (\theta) = \nabla_{\theta} \left( \nabla_{\theta} f (\theta) \right)^T = \begin{bmatrix}
\frac{\partial^2 f}{\partial \theta_1^2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial \theta_n \partial \theta_1} & \cdots & \frac{\partial^2 f}{\partial \theta_n^2}
\end{bmatrix}.
$$

It is symmetric for nice function.

A.20. \( \nabla_x \left( f^T (x) \right) = (df (x))^T \).

A.21. Let \( f, g : \Omega \rightarrow \mathbb{R}^m, \ \Omega \subset \mathbb{R}^n \). \( h(x) = \langle f(x), g(x) \rangle : \Omega \rightarrow \mathbb{R} \). Then

\[
\frac{dh(x)}{dx} = \left( \frac{df}{dx} \right)^T \frac{dg}{dx} + \left( \frac{dg}{dx} \right)^T \frac{df}{dx}.
\]

- For an \( n \times n \) matrix \( A \), let \( f(x) = \langle Ax, x \rangle \). Then \( df(x) = (Ax)^T I + x^T A = x^T A^T + x^T A \).
  - If \( A \) is symmetric, then \( df(x) = 2x^T A \). So, \( \frac{\partial}{\partial x_j} \langle Ax, x \rangle = 2 (Ax)_j \).

A.22. **Chain rule**: If \( f \) is differentiable at \( y \) and \( g \) is differentiable at \( z = f(y) \), then \( g \circ f \) is differentiable at \( y \) and

\[
\frac{d (g \circ f)}{dx} (y) = \frac{dg}{dz} (f(y)) \frac{df}{dx} (y)
\]

(matrix multiplication).

- In particular,

\[
\frac{d}{dt} g(x(t), y(t), z(t)) = \left( \frac{\partial}{\partial x} g(x, y, z) \right) \left( \frac{d}{dt} x(t) \right) + \left( \frac{\partial}{\partial y} g(x, y, z) \right) \left( \frac{d}{dt} y(t) \right) + \left( \frac{\partial}{\partial z} g(x, y, z) \right) \left( \frac{d}{dt} z(t) \right) \bigg|_{(x,y,z)=(x(t),y(t),z(t))}.
\]

A.23. Let \( f : D \rightarrow \mathbb{R}^m \) where \( D \subset \mathbb{R}^n \) is open and connected (so arcwise-connected). Then, \( df(x) = 0 \ \forall x \in D \Rightarrow f \) is constant.

A.24. If \( f \) is differentiable at \( y \), then all partial and directional derivative exists at \( y \)

\[
df(y) = \left( \begin{array}{c}
\nabla f_1 (y) \\
\vdots \\
\nabla f_m (y)
\end{array} \right) = \left( \begin{array}{c}
\frac{\partial f_1}{\partial x_1} (y) & \cdots & \frac{\partial f_1}{\partial x_n} (y) \\
\vdots & \ddots & \vdots \\
\frac{\partial f_m}{\partial x_1} (y) & \cdots & \frac{\partial f_m}{\partial x_n} (y)
\end{array} \right) = \left( \begin{array}{c}
\frac{\partial f}{\partial x_1} (y), \ldots, \frac{\partial f}{\partial x_n} (y)
\end{array} \right).
\]

A.25. **Inversion (Mapping) Theorem**: Let open \( \Omega \subset \mathbb{R}^n \), \( f : \Omega \rightarrow \mathbb{R}^n \) is \( C^1 \), \( c \in \Omega \). \( df(c) \) is bijective, then \( \exists U \) open neighborhood of \( c \) such that
(a) $V = f(U)$ is an open neighborhood of $f(c)$.

(b) $f|_U : U \to V$ is bijective.

(c) $g = (f|_U)^{-1} : V \to U$ is $C^1$.

(d) $\forall y \in V \; dg(y) = [df(g(y))]^{-1}$.

\[
\begin{array}{ll}
d(x) = I & \nabla_x (x^T) = I \\
d(\|x\|^2) = 2x^T & \nabla_x \|x\|^2 = 2x \\
d(Ax + b) = A & \nabla_x (Ax + b)^T = A^T \\
d(a^T x) = a^T & \nabla_x (a^T x) = a \\
\end{array}
\]

\[
\begin{array}{ll}
d(f^T(x) g(x)) & = f^T(x) dg(x) + g^T(x) df(x) \\
 & \nabla_x (f^T(x) g(x)) \\
 & = (dg(x))^T f(x) + (df(x))^T g(x) \\
 & = \nabla_x (g^T(x)) f(x) + \nabla_x (f^T(x)) g(x) \\
\end{array}
\]

For symmetric $Q$, $\nabla_x (Qf(x)) = 2(df(x))^T Qf(x)$.

\[
\begin{array}{ll}
d(f^T(x) Qf(x)) & = f^T(x) Qdf(x) + f^T(x) Q^T df(x) \\
 & = \nabla_x (f^T(x) Qf(x)) \\
 & = 2(df(x))^T Qf(x) \\
\end{array}
\]

\[
\begin{array}{ll}
d(x^T Qx) = 2x^T Q & \nabla_x (x^T Qx) = 2Qx. \\
d(\|f(x)\|^2) = 2f^T(x) df(x) & \nabla_x (\|f(x)\|^2) = 2\nabla_x (f^T(x)) f(x) \\
\end{array}
\]

A.3.2 Integration

A.26. Basic Formulas

(a) $\int a^x du = \frac{a^u}{\ln a}, \; a > 0, \; a \neq 1$.

(b) $\int_0^1 t^\alpha dt = \left\{ \begin{array}{ll} \frac{1}{\alpha+1}, & \alpha > -1 \\
\infty, & \alpha \leq -1 \end{array} \right.$ and $\int_1^\infty t^\alpha dt = \left\{ \begin{array}{ll} \frac{1}{\alpha+1}, & \alpha < -1 \\
\infty, & \alpha \geq -1 \end{array} \right.$ So, the integration of the function $\frac{1}{t}$ is the test case. In fact, $\int_0^1 \frac{1}{t} dt = \int_1^\infty \frac{1}{t} dt = \infty$.

(c) $\int x^m \ln x dx = \left\{ \begin{array}{ll} \frac{x^{m+1}}{m+1} \left( \ln x - \frac{1}{m+1} \right), & m \neq -1 \\
\frac{1}{2} \ln^2 x, & m = -1 \end{array} \right.$

A.27. Integration by Parts:

$$\int uv = uv - \int v du$$
(a) Basic idea: Start with an integral of the form $\int f(x) g(x) dx$.

Match this with an integral of the form $\int udv$ by choosing $dv$ to be part of the integrand including $dx$ and possibly $f(x)$ or $g(x)$.

(b) In particular, repeated application of integration by parts gives

$$\int f(x) g(x) dx = f(x) G_1(x) + \sum_{i=1}^{n-1} (-1)^i f^{(i)}(x) G_{i+1}(x) + (-1)^n \int f^{(n)}(x) G_n(x) dx$$

where $f^{(i)}(x) = \frac{d^i}{dx^i} f(x)$, $G_1(x) = \int g(x) dx$, and $G_{i+1}(x) = \int G_i(x) dx$. Figure 26 can be used to derive (51).

To see this, note that

$$\int f(x) g(x) dx = f(x) G_1(x) - \int f'(x) G_1(x) dx,$$

and

$$\int f^{(n)}(x) G_n(x) dx = f^{(n)}(x) G_{n+1}(x) - \int f^{(n+1)}(x) G_{n+1}(x) dx.$$
Figure 27: Examples of Integration by Parts using figure 26.

(c) \( \int x^n e^{ax} \, dx = \frac{x^n e^{ax}}{a} - \frac{n}{a} \int x^{n-1} e^{ax} \, dx \)

(d) \( \int f(x) g'(x) \, dx = f(x) g(x) - \int f'(x) g(x) \, dx \).

To see this, start with the product rule: \((f(x)g(x))' = f(x)g'(x) + f'(x)g(x)\). Then, integrate both sides.

(e) \( \int_a^b f(x) g'(x) \, dx = f(x) g(x) \bigg|_a^b - \int_a^b f'(x) g(x) \, dx \)

A.28. If \( n \) is a positive integer,

\[
\int x^n e^{ax} \, dx = \frac{e^{ax}}{a} \sum_{k=0}^{n} \frac{(-1)^k n!}{a^k (n-k)!} x^{n-k}.
\]

(a) \( n = 1: \frac{e^{ax}}{a} (x - \frac{1}{a}) \)

(b) \( n = 2: \frac{e^{ax}}{a} (x^2 - \frac{2}{a}x + \frac{2}{a^2}) \)

(c) \( \int_0^1 x^n e^{ax} \, dx = \frac{e^{at}}{a} \sum_{k=0}^{n} \frac{(-1)^k n!}{a^k (n-k)!} t^{n-k} - \frac{(-1)^n n!}{a^{n+1}} = \frac{e^{at}}{a} \sum_{k=0}^{n} \frac{(-1)^k n!}{a^k (n-k)!} t^{n-k} + \frac{n!}{(-a)^{n+1}} \)

(d) \( \int_t^\infty x^n e^{-ax} \, dx = \frac{e^{-at}}{a} \sum_{k=0}^{n} \frac{n!}{a^k (n-k)!} t^{n-k} = \frac{e^{-at}}{a} \sum_{j=0}^{n} \frac{n!}{a^{n-k} j!} t^j \)

(e) \( \int_0^\infty x^n e^{ax} \, dx = \frac{n!}{(-a)^{n+1}}, \quad a < 0. \) (See also Gamma function)

- \( n! = \int_0^\infty e^{-t^2} \, dt. \)
- In MATLAB, consider using \texttt{gamma(n+1)} in stead of \texttt{factorial(n)}. Note also that \texttt{gamma()} allows vector input.

(f) \( \int_0^1 x^\beta e^{-x} \, dx \) is finite if and only if \( \beta > -1 \).

Note that \( \frac{1}{e} \int_0^1 x^\beta dx \leq \int_0^1 x^\beta e^{-x} dx \leq \int_0^1 x^\beta dx. \)
(g) \( \forall \beta \in \mathbb{R}, \int_{1}^{\infty} x^\beta e^{-x} dx < \infty. \)

For \( \beta \leq 0, \int_{1}^{\infty} x^\beta e^{-x} dx \leq \int_{1}^{\infty} e^{-x} dx < \int_{0}^{\infty} e^{-x} dx = 1. \)

For \( \beta > 0, \int_{1}^{\infty} x^\beta e^{-x} dx \leq \int_{1}^{\infty} x^{\lceil \beta \rceil} e^{-x} dx \leq \int_{0}^{\infty} x^{\lceil \beta \rceil} e^{-x} dx = \lceil \beta \rceil! \)

(h) \( \int_{0}^{\infty} x^\beta e^{-x} dx \) is finite if and only if \( \beta > -1. \)

A.29 (Differential of integral). **Leibniz’s Rule**: Let \( g : \mathbb{R}^2 \to \mathbb{R}, a : \mathbb{R} \to \mathbb{R}, \) and \( b : \mathbb{R} \to \mathbb{R} \) be \( C^1. \) Then \( f (x) = \int_{a(x)}^{b(x)} g (x, y) dy \) is \( C^1 \) and

\[
    f' (x) = b' (x) g (x, b (x)) - a' (x) g (x, a (x)) + \int_{a(x)}^{b(x)} \frac{\partial g}{\partial x} (x, y) dy. \tag{52}
\]

In particular, we have

\[
    \frac{d}{dx} \int_{a}^{x} f (t) dt = f (x), \tag{53}
\]

\[
    \frac{d}{dx} \int_{a}^{v (x)} f (t) dt = \frac{dv}{dx} \frac{d}{dv} \int_{a}^{v (x)} f (t) dt = f (v (x)) v' (x), \tag{54}
\]

\[
    \frac{d}{dx} \int_{u (x)}^{v (x)} f (t) dt = \frac{d}{dx} \left( \int_{a}^{v (x)} f (t) dt - \int_{a}^{u (x)} f (t) dt \right) = f (v (x)) v' (x) - f (u (x)) u' (x). \tag{55}
\]

Note that (52) can be derived from (A.22) by considering \( f (x) = h (a (x), b (x), x) \) where \( h (a, b, c) = \int_{a}^{c} g (c, y) dy. \) \([9, p 318–319].\)

A.4 Gamma and Beta functions

A.30. Gamma function:

(a) \( \Gamma (q) = \int_{0}^{\infty} x^{q-1} e^{-x} dx. \) ; \( q > 0. \)

(b) \( \Gamma (n) = (n - 1)! \) for \( n \in \mathbb{N}. \)

\( \Gamma (n + 1) = n! \) if \( n \in \mathbb{N} \cup \{0\}. \)

(c) \( 0! = 1. \)
(d) \( \Gamma \left( \frac{1}{2} \right) = \sqrt{\pi} \).

(e) \( \Gamma (x + 1) = x \Gamma (x) \) (Integration by parts).
   
   • This relationship is used to define the gamma function for negative numbers.

(f) \( \frac{\Gamma(q)}{\alpha^q} = \int_0^\infty x^{q-1} e^{-\alpha x} \, dx, \alpha > 0. \)

\[ \begin{array}{c}
\text{Figure 28: Plot of gamma function.}
\end{array} \]

(g) By limiting argument (as \( q \to 0^+ \)), we have \( \Gamma (0) = \infty \) which implies \( \Gamma (-n) = \infty \), for \( n \in \mathbb{N} \).

A.31. The **incomplete beta function** is defined as

\[ B(x; a, b) = \int_0^x t^{a-1} (1 - t)^{b-1} \, dt. \]

For \( x = 1 \), the incomplete beta function coincides with the (complete) beta function.

The **regularized incomplete beta function** (or **regularized beta function** for short) is defined in terms of the incomplete beta function and the (complete) beta function:

\[ I_x(a, b) = \frac{B(x; a, b)}{B(a, b)}. \]

• For integers \( m, k \),

\[ I_x(m, k) = \sum_{j=m}^{m+k-1} \frac{(m + k - 1)!}{j!(m + k - 1 - j)!} x^j (1 - x)^{m+k-1-j}. \]

• \( I_0(a, b) = 0, I_1(a, b) = 1 \)

• \( I_x(a, b) = 1 - I_{1-x}(b, a) \)
References


Index

Bayes Theorem, 22
Binomial theorem, 12
Birthday Paradox, 24
Cauchy-Bunyakovskii-Schwartz Inequality, 81
Chevalier de Mere’s Scandal of Arithmetic, 23
Coefficient of Variation (CV), 74
Delta function, 19
Dice, 23
Die, 23
Dirac delta function, see Delta function
Event Algebra, 25
False Positives on Diagnostic Tests, 24
Fano Factor, 74
gradient, 134
gradient vector field, 134
Hölder’s Inequality, 80
Hessian matrix, 136
Integration by Parts, 137
Isserlis’s Theorem, 123
Jacobian, 100, 134
Jacobian formulas, 101
Jensen’s Inequality, 81
Leibniz’s Rule, 140
Lyapounov’s Inequality, 81
Markov’s Inequality, 78
Minkowski’s Inequality, 81
Monte Hall’s Game, 24
multinomial coefficient, 13
Multinomial Counting, 13
Multinomial Theorem, 15
Order Statistics, 103
Probability of coincidence birthday, 24
Standard Deviation, 74
Total Probability Theorem, 22
uncorrelated not independent, 77, 95
Zeta function, 133
Zipf or zeta random variable, 60