4.3 Expectations of Discrete Random Variables ........................................... 38
4.4 Variance .............................................................................................. 40
4.5 Probability Generating Functions .............................................................. 42
4.6 The Binomial Distribution .................................................................... 43
4.7 The Poisson Distribution ..................................................................... 45
  4.7.1 Poisson Processes ......................................................................... 47
4.8 Linearity of Expectations ..................................................................... 48
4.9 The Geometric Distribution .................................................................. 50
4.10 The Negative Binomial Distribution ...................................................... 51
4.11 The Hypergeometric Distribution .......................................................... 52

5 Continuous Random Variables .................................................................. 54
  5.1 Cumulative Distribution Functions ......................................................... 54
  5.2 Continuous Random Variables ............................................................... 55
  5.3 Expectation ............................................................................................ 56
  5.4 The Uniform Distribution ...................................................................... 58
  5.5 The Normal Distribution ...................................................................... 60
  5.6 The Exponential Distribution and Relatives .......................................... 62
    5.6.1 The Gamma Distribution ................................................................ 64
    5.6.2 The Beta Distribution .................................................................... 65
  5.7 Compositions and Distributions ............................................................... 66

6 Random Vectors ....................................................................................... 70
  6.1 Multivariate Distributions ..................................................................... 70
  6.2 Independent Random Variables ............................................................. 72
  6.3 Sums of Independent Random Variables .............................................. 74
  6.4 Conditional Distributions .................................................................... 76
    6.4.1 Discrete Conditional Distributions ................................................ 76
    6.4.2 Continuous Conditional Distributions .......................................... 78

7 Expectations ............................................................................................. 80
  7.1 Expectations of Sums ........................................................................... 80
  7.2 Variance and Covariance ..................................................................... 83
  7.3 Conditional Expectation ....................................................................... 87
7.3.1 Conditional Variance .................................................. 91
7.4 Moment Generating Functions .......................................... 92

8 Limit Theorems for Sequences of Random Variables ............. 96
8.1 The Weak Law of Large Numbers ..................................... 96
8.2 The Strong Law of Large Numbers ................................... 98
8.3 The Central Limit Theorem .............................................. 99
Chapter 1

Overview and Conceptual Foundations of Probability

1.1 Deterministic and Statistical Regularity

Scientific knowledge is predicated on the belief that there is regularity in nature. We can contrast two kinds of regularity.

**Definition 1.1.** *Deterministic regularity* holds if we can predict how a system will evolve over time from its initial conditions. In particular, in a deterministic world, we should observe the same outcome whenever we repeat an experiment under the same conditions.

The influence of this point of view is reflected in the success of Newtonian mechanics and the use of ordinary and partial differential equations to model physical phenomena, e.g., Newton’s second law (force equals mass times accelerations) can be expressed as a differential equation:

\[ F = ma = m \frac{d^2x}{dt^2}. \]

Nonetheless, there are several limits to this perspective:

- imprecise knowledge of the initial conditions of any system;
- imprecise knowledge of the laws governing the system, e.g., how do we model the dynamics of an entire cell?
- numerical solutions of ODEs and PDEs always have some limited accuracy and this may not be good enough to deal with either chaotic or high-dimensional models;
- quantum mechanics provides a statistical rather than a deterministic description of the world, i.e., solutions to Schrödinger’s equation characterize the probability that a system is in a given state.

Thus, science must be able to cope with a pervasive uncertainty. This appears to be possible in part because the world exhibits some measure of statistical regularity.

**Definition 1.2.** *Statistical regularity* holds if when we repeat an experiment a large number of times, the proportion of outcomes that satisfy a condition of interest tends to a limit.
Example 1.1. Suppose that we repeatedly toss a coin and we let $X_1, X_2, \cdots$ denote the sequence of outcomes, i.e., $X_1$ is either heads or tails depending on what we get on the first toss. Under deterministic regularity, we should be able to predict the exact sequence of heads or tails. Under statistical regularity, this prediction may not be possible, but the following ratio should converge to a limit (that we call the probability of getting heads):

$$P(\text{heads}) = \lim_{n \to \infty} \frac{1}{n} \# \left\{ i : 1 \leq i \leq n : X_i = \text{heads} \right\}.$$  

1.2 Probability Spaces

Although the mathematical study of probability has its roots in the 16’th century, the first comprehensive theories were proposed in the early 20’th century. Two fundamentally different approaches were developed:

- The frequentist theory developed by R. von Mises took sets of infinite sequences exhibiting certain properties of statistical regularity as its basic objects.
- The measure theoretic probability developed by Kolmogorov took measures and $\sigma$-algebras (defined below) as the basic objects.

Although frequentist theories are more closely tied to our intuitive understanding of statistical regularity, they are mathematically unwieldy. Measure theory, in contrast, is initially less intuitive, but leads to a more straightforward mathematical formulation of probability. Furthermore, at the end of the semester, we will see that the frequentist concept of probability can be deduced from the measure theoretic formulation in the guise of the law of large numbers.

To motivate the following definition, imagine that you are conducting an experiment with a random outcome and that you want to describe this mathematically.

Definition 1.3. A probability space is a triple $\{ \Omega, \mathcal{F}, P \}$ where:

- $\Omega$ is the sample space, i.e., the set of all possible outcomes of the experiment.
- $\mathcal{F}$ is a collection of subsets of $\Omega$ which we call events. $\mathcal{F}$ is called a $\sigma$-algebra and is required to satisfy the following conditions:
  1. The empty set and the sample space are both events: $\emptyset, \Omega \in \mathcal{F}$.
  2. If $E$ is an event, then its complement $E^c = \Omega \setminus E$ is also an event.
  3. If $E_1, E_2, \cdots$ are events, then their union $\bigcup_n E_n$ is an event.
- $P$ is a function from $\mathcal{F}$ into $[0, 1]$: if $E$ is an event, then $P(E)$ is the probability of $E$. $P$ is said to be a probability distribution or probability measure on $\mathcal{F}$ and is also required to satisfy several conditions:
  1. $P(\emptyset) = 0; P(\Omega) = 1$.
  2. Countable additivity: If $E_1, E_2, \cdots$ are mutually exclusive events, i.e., $E_i \cap E_j = \emptyset$ whenever $i \neq j$, then
     $$P \left( \bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} P(E_n).$$
Example 1.2. Suppose that the experiment consists of tossing a fair coin twice and recording whether we get heads or tails on each toss. Then

- $\Omega = \{(H,H), (H,T), (T,H), (T,T)\}$;
- $\mathcal{F}$ is the collection of all 16 subsets of $\Omega$;
- if $E$ is a subset of $\Omega$ containing $k$ outcomes, then $P(E) = k/4$.

Example 1.3. Now suppose that the experiment consists of tossing a fair coin twice and recording the total number of heads. Then

- $\Omega = \{(H,H), (H,T), (T,H), (T,T)\}$;
- $\mathcal{F}$ is the collection $\{\emptyset, \{(H,H)\}, \{(H,T), (T,H)\}, \{(T,T)\}, \{(H,H), (H,T), (T,H)\}, \{(H,T), (T,H), (T,T)\}, \{(H,H), (T,T)\}, \Omega\}$;
- $P(0$ heads$) = P(2$ heads$) = 1/4$, $P(1$ heads$) = 1/2$.

Remark 1.1.

- $\Omega$ is required to contain all possible outcomes, but sometimes it is convenient to allow $\Omega$ to also contain outcomes that are not possible, e.g., if the experiment consists of recording the mass of a bird in a field study, we could take $\Omega = \mathbb{R}$ despite the fact that the mass will never be a negative number.
- An event is a set of outcomes that is knowable, e.g., $\{(H,T)\}$ is not an event in Example 1.3 because we cannot know whether the sequence of tosses was heads followed by tails if all we know is the number of tosses that resulted in heads.
- $P(E)$ is only defined if $E \in \mathcal{F}$.
- $P(\emptyset) = 0$ means that the probability that nothing (whatsoever) happens is zero.
- $P(\Omega) = 1$ means that the probability that something (whatever it is) happens is one.
- If $E_1$ and $E_2$ are mutually exclusive events, then $E_1 \cup E_2$ is the event that either $E_1$ or $E_2$ happens and the probability of that is just the sum of the probability that $E_1$ happens and the probability that $E_2$ happens:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

Countable additivity says that this property holds when we have a countable collection of disjoint events.

Why do we need to restrict to countable additivity?

Example 1.4. $\Omega = [0,1]$, $\mathcal{F}$ contains all open and closed intervals, $P([a,b]) = b - a$

- $P(x) = 0$
- $1 = P([0,1]) \neq \sum_x P(x) = 0$. 
The problem is that \([0, 1]\) is uncountable. (Proof by Cantor diagonalization.)

**Lemma 1.1.** The following are properties of countable sets.

- a finite union of countable sets is countable
- a countable union of countable sets is countable
- a finite product of countable sets is countable
- a countably infinite product of finite sets is not countable

In this course, we will mostly not pay too much attention to the \(\sigma\)-algebras defined on the probability spaces that we study. With finite or countably-infinite sets, these usually are taken to be the collection of all subsets of the sample space, so that the probability of every subset is defined. Probability measures on uncountable sets are much more complicated objects and one needs to pay attention to these technical details to avoid contradiction. David Williams’ book ‘Probability with Martingales’ (Cambridge, 1991) provides an excellent introduction to these issues for those who wish to learn the mathematics more rigorously than it is presented in Ross.

### 1.3 Properties of Events

For future reference, we recall some basic facts from set theory.

Operations with events:

- union: \(E \cup F\) is the set of outcomes that are in \(E\) or \(F\) (non-exclusive)
- intersection: \(E \cap F\) is the set of outcomes that are in both \(E\) and \(F\)
  - Ross uses the notation \(EF\)
  - \(E\) and \(F\) are said to be **mutually exclusive** if \(E \cap F = \emptyset\)
- unions and intersections can be extended to more than two sets.
- complements: \(E^c\) is the set of outcomes in \(\Omega\) that are not in \(E\)

Unions and intersections satisfy several important algebraic laws:

- Commutative laws: \(E \cup F = F \cup E, E \cap F = F \cap E\)
- Associative laws: \((E \cup F) \cup G = E \cup (F \cup G), (E \cap F) \cap G = E \cap (F \land G)\)
- Distributive laws: \((E \cup F) \cap G = (E \cap G) \cup (F \cap G), (E \cap F) \cup G = (E \cup G) \cap (F \cup G)\)
Lemma 1.2. DeMorgan’s laws:

\[
\left( \bigcup_{n=1}^{\infty} E_n \right)^c = \bigcap_{n=1}^{\infty} E_n^c \\
\left( \bigcap_{n=1}^{\infty} E_n \right)^c = \bigcup_{n=1}^{\infty} E_n^c
\]

One consequence of DeMorgan’s laws is that $\sigma$-algebras are closed under countable intersections: if $E_1, E_2, \cdots$ are events in a $\sigma$-algebra $\mathcal{F}$, then

\[
\bigcap_{n=1}^{\infty} E_n
\]

is also an event in $\mathcal{F}$.

1.4 Properties of Probabilities

The following lemma lists some useful properties that can be deduced from Definition 1.3.

Lemma 1.3. The following properties hold for any two events $A, B$ in a probability space:

1. Complements: $P(A^c) = 1 - P(A)$.
2. Subsets: if $A \subset B$, then $P(A) \leq P(B)$;
3. Disjoint events: If $A$ and $B$ are mutually exclusive, then

\[
P(A \cap B) = 0.
\]
4. Unions: For any two events $A$ and $B$ (not necessarily mutually exclusive), we have:

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]

Example 1.5. Suppose that the frequency of HIV infection in a community is 0.05, that the frequency of tuberculosis infection is 0.1, and that the frequency of dual infection is 0.01. What proportion of individuals have neither HIV nor tuberculosis?

\[
P(\text{HIV or tuberculosis}) = P(\text{HIV}) + P(\text{tuberculosis}) - P(\text{both}) = 0.05 + 0.1 - 0.01 = 0.14
\]
\[
P(\text{neither}) = 1 - P(\text{HIV or tuberculosis}) = 0.86
\]

The identity that appears in the next theorem is usually known as the inclusion-exclusion formula. Also, if $n = 2$, then this is just the fourth statement of Lemma 1.3.
Theorem 1.1. Let \( E_1, \cdots, E_n \) be a collection of events. Then
\[
\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) = \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{1 \leq i_1 < i_2 \leq n} \mathbb{P}(E_{i_1} \cap E_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1 < i_2 < \cdots < i_r} \mathbb{P}(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_r}) + \cdots + (-1)^{n+1} \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n)
\]

Later in the semester we will give a simple proof of this result using indicator functions. For now, we will use Venn diagrams to prove it in the special case when \( n = 3 \).

1.5 Continuity properties of probability measures

We first recall what it means for a sequence of real numbers to converge to a limit.

**Definition 1.4.** A sequence of real numbers \( x_1, x_2, \cdots \) is said to **converge to a limit** \( x \) if for every positive real number \( \epsilon > 0 \), we can find an integer \( N \) such that the difference \(|x_n - x|\) is less than \( \epsilon \) whenever \( n \geq N \). When this is true, we write \( x_n \to x \) or, more formally,
\[
x = \lim_{n \to \infty} x_n.
\]

The intuition behind this definition is that as \( n \) increases, the terms \( x_n \) in a convergent sequence should approach the limit arbitrarily closely and then remain close to that limit.

**Example 1.6.** If \( x_n = \frac{n-1}{n} \), then the sequence \( x_1, x_2, \cdots \) converges to the limit \( x = 1 \). Indeed, if \( \epsilon > 0 \) is a positive real number and we take \( N \) to be any positive integer greater than \( 1/\epsilon \), then for any integer \( n \geq N \), we have
\[
|x - x_n| = \left| 1 - \frac{n-1}{n} \right| = \frac{1}{n} \leq \frac{1}{N} < \epsilon.
\]

**Definition 1.5.** A sequence of nested sets \( \{E_n; n \geq 1\} \) is said to be **increasing** if
\[
E_1 \subset E_2 \subset E_3 \cdots,
\]
in which case we define
\[
\lim_{n \to \infty} E_n = \bigcup_{n \geq 1} E_n.
\]

Similarly, a sequence of sets is said to be **decreasing** if
\[
E_1 \supset E_2 \supset E_3 \cdots,
\]
and then we define
\[
\lim_{n \to \infty} E_n = \bigcap_{n \geq 1} E_n.
\]

Notice that in both cases the sets \( E_n \) are nested.
Theorem 1.2. Suppose that \( \{E_n; n \geq 1\} \) is either an increasing or decreasing sequence of sets. Then,

\[
\lim_{n \to \infty} \mathbb{P}(E_n) = \mathbb{P}(\lim_{n \to \infty} E_n).
\]

Proof. Suppose that \( E_1, E_2, \cdots \) is an increasing sequence of events and observe that this implies that

\[
E_n = \bigcup_{k=1}^{n} E_k \quad (1.1)
\]

for every \( n \geq 1 \). We would like to use countable additivity to prove the result (since this is the only tool currently at our disposal that involves the probability of countably infinitely many events) and to this end we will recursively define a new collection of events:

\[
\begin{align*}
B_1 &= E_1 \\
B_2 &= E_2 \setminus E_1 \\
B_3 &= E_3 \setminus E_2 \\
& \quad \quad \vdots \\
B_n &= E_n \setminus E_{n-1}.
\end{align*}
\]

We first observe that the events \( B_1, B_2, \cdots \) are disjoint. Indeed, given any two positive integers \( k < n \), notice that since \( B_n \) is disjoint from \( E_{n-1} \) and since \( E_k \) is a subset of \( E_{n-1} \), \( B_n \) is disjoint from \( E_k \). Furthermore, since \( B_k \) is itself a subset of \( E_k \) (look at how \( B_k \) is defined), it follows that \( B_n \) is disjoint from \( B_k \), as claimed above.

I also claim that

\[
\bigcup_{k=1}^{n} B_k = \bigcup_{k=1}^{n} E_k \quad (1.2)
\]

for every integer \( n \geq 1 \). We can prove this by induction on \( n \). First, notice that since \( B_1 = E_1 \), this identity is true by definition when \( n = 1 \). Now suppose that it is true for an arbitrary value of \( n \). Then

\[
\begin{align*}
\bigcup_{k=1}^{n+1} B_k &= B_{n+1} \cup \left( \bigcup_{k=1}^{n} B_k \right) \\
&= E_{n+1} \setminus E_n \cup \left( \bigcup_{k=1}^{n} B_k \right) \\
&= E_{n+1} \setminus E_n \cup \left( \bigcup_{k=1}^{n} E_k \right) \\
&= E_{n+1} \cup \left( \bigcup_{k=1}^{n} E_k \right) \\
&= \bigcup_{k=1}^{n+1} E_k,
\end{align*}
\]

which completes the induction step. Notice that it immediately follows that

\[
\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} E_k \quad (1.3)
\]
Consequently,
\[
\lim_{n \to \infty} \mathbb{P}(E_n) = \lim_{n \to \infty} \mathbb{P} \left( \bigcup_{k=1}^{n} E_k \right) \quad \text{by (1.1)}
\]
\[
= \lim_{n \to \infty} \mathbb{P} \left( \bigcup_{k=1}^{n} B_k \right) \quad \text{by (1.2)}
\]
\[
= \lim_{n \to \infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \quad \text{since the } B_k \text{ are disjoint}
\]
\[
= \sum_{k=1}^{\infty} \mathbb{P}(B_k)
\]
\[
= \mathbb{P} \left( \bigcup_{k=1}^{\infty} B_k \right) \quad \text{by countable additivity}
\]
\[
= \mathbb{P} \left( \bigcup_{k=1}^{\infty} E_k \right) \quad \text{by (1.3)},
\]
which completes the proof. DeMorgan’s laws can then be used to deduce the result for a decreasing sequence of events from the increasing case. (To see this, notice that if \( E_1, E_2, \cdots \) is a decreasing sequence of events, then their complements \( E_1^c, E_2^c, \cdots \) form an increasing sequence.) \( \square \)

**Caveat:** In general, the set \( \lim_{n \to \infty} E_n \) is not defined for an arbitrary sequence of events and the sequence \( \mathbb{P}(E_n) \) need not converge.

### 1.6 Littlewood-Ross paradox

Continuity of probability plays an important role in the stochastic variant of the Littlewood-Ross paradox.

**Example 1.7.** Suppose that we have an urn and an infinite collection of balls numbered 1, 2, 3, \cdots.
We consider three different procedures that we can carry out with these objects.

- (i) At 1 minute to noon, we add balls 1 − 10 to the urn and then remove ball number 10. At 30 seconds to noon, we add balls 11 − 20 to the urn and then remove ball number 20. At 15 seconds to noon, we add balls 21 − 30 to the urn and then remove ball number 30. We repeat this procedure until noon, at which point the urn will contain infinitely many balls.

- (ii) Same procedure as in (i), except that now we remove ball number \( n \) at the \( n \)th step. In this case, every ball is removed before noon, so that the urn is ultimately empty.

- (iii) Now consider a stochastic version in which at each step we randomly select one of the balls present in the urn and remove it. It can be shown that with probability 1, every ball is removed by noon so that the urn is eventually empty.
Proof for (iii): Let $E_{1,n}$ be the event that ball number 1 is still in the urn after the first $n$ withdrawals and let $F_1$ be the event that ball number 1 is still in the urn at noon. Notice that the sequence $\{E_{1,n}; n \geq 1\}$ is decreasing and that

$$F_1 = \lim_{n \to \infty} E_{1,n} = \bigcap_{n=1}^{\infty} E_{1,n}. $$

Then

$$P(E_{1,n}) = \frac{9}{10} \cdot \frac{18}{19} \cdot \ldots \cdot \frac{9n}{9n+1},$$

and a direct calculation shows that

$$P(\bigcap_{n=1}^{\infty} E_{1,n}) = \lim_{n \to \infty} P(E_{1,n}) = 0.$$ 

Similarly, it can be shown that for any $j \geq 1$, the probability that ball number $j$ remains in the urn at noon is $P(F_j) = 0$. Therefore, the probability that the urn is not empty at noon is

$$P\left( \bigcup_{n=1}^{\infty} F_n \right) \leq \sum_{n=1}^{\infty} P(F_n) = 0.$$

$\square$
Chapter 2

Combinatorial Analysis

Many problems in probability involve finite sample spaces in which every outcome is equally likely (e.g., a sequence of tosses of a fair coin). In this case, the probability of an event can be calculated by dividing the number of outcomes that belong to that particular event by the total number of outcomes in the sample space. Combinatorial analysis provides us with a set of tools that can be used to efficiently determine the sizes of these sets.

2.1 Ordered Samples

Theorem 2.1. Basic Principle of Counting: Suppose that $r$ experiments are performed such that the first experiment can result in any one of $n_1$ outcomes, the second experiment can result in any one of $n_2$ outcomes, the third experiment can result in any one of $n_3$ outcomes, and so forth. Here the particular outcomes (but not the number thereof) that can occur on the $k$’th experiment may depend on the outcomes of experiments 1 through $k - 1$. Then the total number of possible outcomes for the ordered sequence of all $r$ experiments is $n_1 n_2 \cdots n_r$.

Here are some important special cases:

Ordered samples: If set $E_1$ contains $n_1$ elements, $E_2$ contains $n_2$ elements, ..., and $E_r$ contains $n_r$ elements, then the number of ordered $r$-tuples $(x_1, x_2, \cdots, x_r)$ that can be formed with $x_i \in E_i$ is equal to $n_1 n_2 \cdots n_r$.

Sampling with replacement: If $E_1 = \cdots = E_r = E$, which contains $n$ elements, then the number of ordered $r$-tuples is just $n^r$. We can think of the $r$-tuple $(x_1, x_2, \cdots, x_r)$ as an ordered sample of size $r$ from $E$ where we are sampling with replacement, i.e., the same element can be sampled repeatedly.

Sampling without replacement: If we sample a set containing $n$ elements $r$-times without replacement, then the number of ordered samples is $n(n − 1)(n − 2)\cdots(n − r + 1)$. In this case, each element can be sampled at most once, so that the elements that can be sampled on the $k$’th draw depend on the elements sampled on draws 1, $\cdots$, $k − 1$.

Example 2.1. The number of different subsets of a set $E$ containing $n$ elements is $2^n$. To see this, write $E = \{e_1, \cdots, e_n\}$ and notice that there is a one-to-one correspondence between the subsets of $E$ and binary sequences $(x_1, \cdots, x_n)$ obtained by setting $x_i = 1$ if $e_i$ is in the subset.
and \( x_i = 0 \) otherwise.

**Example 2.2.** Proteins and DNA sequences. Proteins are polymers (sequences) built up from 20 molecules called amino acids. The order in which the amino acids appear in the sequence determines both the structure and the function of the protein. Notice that there are \( 20^L \) distinct proteins that can be made using a sequence of \( L \) amino acids.

DNA is also a polymer, but it is made up 4 basic molecules called nucleotides, denoted A, T, G, and C. Thus there are \( 4^L \) distinct DNA sequences that are \( L \) nucleotides long. Proteins are encoded by DNA sequences, i.e., each protein corresponds to a segment of DNA called a gene, and this encoding is local in the sense that changing one nucleotide in a gene will change at most one amino acid in the corresponding protein. However, more than one nucleotide is needed to specify an amino acid. Indeed, since there are 20 amino acids but only 4 nucleotides, groups of either one nucleotide or two neighboring nucleotides could specify at most 4 and \( 4^2 = 16 \) amino acids. In fact, groups of three neighboring nucleotides, called triplet codons (or codons, for short) are used to specify amino acids. This is possible since there are \( 4^3 = 64 \) such triplet codons, which is more than enough to encode 20 amino acids. Moreover, because there are more than 20 such codons, we can deduce that the genetic code must be degenerate: some amino acids must be encoded by more than one triplet codon. In general, this degeneracy is present at the third nucleotide in the codon, which can often be changed without changing the identity of the corresponding amino acid.

### 2.2 Permutations

Suppose that the set \( E \) contains \( n \) distinct elements. Then the number of ordered arrangements of these elements is equal to

\[
n! = n(n-1)(n-2)\cdots 1,
\]

where the expression \( n! \) is read \( n \) factorial. This is true because there are \( n \) choices of the first element, \( n-1 \) choices of the second element, \( n-3 \) choices of the third element, and so forth. By convention, \( 0! = 1 \) and

\[
1! = 1, 2! = 2, 3! = 6, 4! = 24, 5! = 120, 6! = 720, 8! = 40320, \ldots
\]

**Example 2.3.** The number of distinct shuffles (permutations) of a pack of cards is \( 52! \approx 8.1 \times 10^{67} \). The number of shuffles that move each card from its current place in the deck to a place occupied by another card belonging to the same suit is \((13!)^4 \approx 1.5 \times 10^{39}\).

Now suppose that \( E \) contains \( n = n_1 + n_2 + \cdots + n_r \) elements, \( n_1 \) of which are alike, \( n_2 \) of which are alike, etc. How many ordered arrangements are there of \( E \) in which elements of the same type are not distinguished?

To answer this question, let \( P_{n,r} \) denote the unknown number and notice that for each such arrangement, there are \( n_1! \) ways to permute the type one elements amongst themselves, there are \( n_2! \) ways to permute the type two elements among themselves, and so forth. It follows that for each arrangement of the elements of \( E \) there are \( n_1!n_2!\cdots n_r! \) permutations that shuffle elements of the same type amongst themselves. Since there are \( n! \) ordered arrangements of \( E \) in which we distinguish between all elements (say by adding extra labels to these elements), we see that

\[
n! = P_{n,r}(n_1!n_2!\cdots n_r!),
\]
and therefore the number of ordered arrangements that do not distinguish between elements of the same type is

\[
P_{n,r} = \frac{n!}{n_1!n_2! \cdots n_r!}.
\]

**Example 2.4.** The number of rearrangements of the letters in the word CHACHALACA is:

\[
\frac{10!}{4!3!2!1!} = 12600,
\]

since the letter A is repeated 4 times, C is repeated 3 times, H is repeated 2 times, and L occurs just once.

### 2.3 Combinations

If a set \( E \) contains \( n \) distinct elements, then how many distinct subset of \( r \) objects are contained in \( E \)? Notice that this is equivalent to asking how many different ways we can sample \( r \) objects from \( E \) without replacement and without regard to order.

To solve this, observe that the Basic Principle of Counting implies that there are

\[
n(n - 1)(n - 2) \cdots (n - r + 1) = \frac{n!}{(n - r)!}
\]
different ways to sample \( r \) objects from \( E \) without replacement when the order does matter. Furthermore, there are \( r! \) different permutations (or orders) of each such set of \( r \) objects. Consequently, if \( C_{n,r} \) denotes the number of different subsets containing \( r \) objects in a set of size \( n \), then

\[
\frac{n!}{(n - r)!} = C_{n,r}r!,
\]
since the number of ordered samples (the left-hand side) is equal to the number of unordered samples \( (C_{n,r}) \) times the number of ways of ordering any particular sample containing \( r \) objects \( (r!) \). This shows that

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

where the expression \( \binom{n}{r} \) is read ‘\( n \) choose \( r \)’. Also, observe that

\[
\binom{n}{r} = \binom{n}{n - r},
\]

which follows either by direct calculation or by noticing that each choice of \( r \) elements from \( n \) corresponds to a choice of the \( n - r \) remaining objects.

**Example 2.5.** Suppose that a coin is flipped ten times. Then the number of sequences which contain exactly \( k \) heads is \( \binom{10}{k} \):

\[
\begin{align*}
\binom{10}{0} &= \binom{10}{10} = 1, \\
\binom{10}{1} &= \binom{10}{9} = 10, \\
\binom{10}{2} &= \binom{10}{8} = 45 \\
\binom{10}{3} &= \binom{10}{7} = 120, \\
\binom{10}{4} &= \binom{10}{6} = 210, \\
\binom{10}{5} &= 252
\end{align*}
\]

Notice that there are many more sequences which have roughly equal numbers of heads and tails than there are sequences that have a preponderance of one over the other.
The next theorem explains why the numbers \( \binom{n}{k} \) are known as binomial coefficients.

**Theorem 2.2. Binomial Theorem.** For any non-negative integer \( n \geq 0 \),

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

This can be proved by induction or by observing that when we expand the product \((x + y)(x + y) \cdots (x + y)\), there will be exactly \( \binom{n}{k} \) ways of choosing \( k \) terms to contribute an \( x \) and \( n - k \) terms to contribute a \( y \), with each such choice contributing 1 to the coefficient of \( x^k y^{n-k} \) on the right-hand side.

By taking either \( x = y = 1 \) or \( x = 1, y = -1 \), we obtain the following important identities:

\[
\sum_{k=0}^{n} \binom{n}{k} = (1 + 1)^n = 2^n
\]

\[
\sum_{k=0}^{n} \binom{n}{k} (-1)^k = (-1 + 1)^n = 0.
\]

### 2.4 Multinomial Coefficients

There is an important generalization of the binomial coefficients. Notice that the act of choosing \( k \) elements from a set \( E \) of size \( n \) is equivalent to dividing \( E \) into two disjoint subsets of sizes \( k \) and \( n - k \). More generally, if \( n = n_1 + \cdots + n_r \), where the \( n_i \)'s are positive integers, then the number of ways of dividing \( E \) into \( r \) disjoint subsets of respective sizes \( n_1, n_2, \cdots, n_r \) is equal to

\[
\binom{n}{n_1, n_2, \cdots, n_r} = \frac{n!}{n_1! n_2! \cdots n_r!}.
\]

The quantity \( \binom{n}{n_1, \cdots, n_r} \) is called a multinomial coefficient.

**Example 2.6.** Suppose that a six-sided die is rolled 10 times. Then the number of sequences containing five 1’s, three 3’s and two 6’s is

\[
\binom{10}{5, 0, 3, 0, 2} = \binom{10}{5, 3, 2} = 2520.
\]

**Remark 2.1.** Those \( n_i \) equal to zero can be omitted from the multinomial coefficient.

The binomial expansion also generalizes to expressions involving three or more variables. The following identity is called the multinomial expansion:

\[(x_1 + x_2 + \cdots + x_r)^n = \sum_{n_1 + \cdots + n_r = n} \binom{n}{n_1, n_2, \cdots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}.\]

Here, the sum on the right-hand side is over all nonnegative integer-valued vector \( (n_1, n_2, \cdots, n_r) \) such that \( n_1 + n_2 + \cdots n_r = n \).
2.5 Symmetric Probability Spaces

**Definition 2.1.** Suppose that \( S = \{1, \cdots, N\} \) is finite and let \( F = \mathcal{P}(S) \). If the probability of every element is equal, then \( (S,F,P) \) is said to be a symmetric probability space. In this case, we have

\[
1 = \mathbb{P}(S) = \sum_{i=1}^{N} \mathbb{P}\{\{i\}\} = N \cdot \mathbb{P}(i)
\]

and so \( \mathbb{P}\{\{i\}\} = 1/N \) for each \( i \in S \). Furthermore, if \( E = \{x_1, \cdots, x_k\} \subset S \), then

\[
\mathbb{P}(E) = \sum_{i=1}^{k} \mathbb{P}(\{x_i\}) = \sum_{i=1}^{k} \frac{1}{N} = \frac{k}{n} = \frac{|E|}{|S|},
\]

where \(|E|\) denotes the number of elements in \( E \).

**Remark 2.2.** I am following Ross in using \( S \) rather than \( \Omega \) to denote the sample space of a symmetric probability space.

**Example 2.7.** *(Ross, Example 5a)* If two dice are rolled, what is the probability that the sum of the numbers is equal to 7?

**Solution:** Out of a total of 36 possible outcomes, there are exactly 6 that satisfy the condition that the sum is 7, so the probability is \( 6/36 = 1/6 \).

*(Ross, Example 5a contd.)* If two dice are rolled, what is the probability that the number obtained on the first roll is less than the second number?

**Solution:** Since the first number is either less than, equal to, or greater than the second number and the probability that the first number is less than the second number is equal to the probability that it is greater than it, the probability in question is

\[
\frac{1}{2} \left( 1 - \frac{6}{36} \right) = \frac{5}{12}.
\]

**Example 2.8.** *(Ross, Exercise 49)* If a group of 6 men and 6 women is randomly divided into two groups of size 6 each, what is the probability that both groups have the same number of men?

**Solution:** There are \( \binom{12}{3} \) ways of choosing the two groups. Since both groups can contain the same number of men only if each contains three men, in which case each also contains three women, the number of such choices is equal to \( \binom{6}{3}^2 \), giving \( \binom{6}{3}^2 / \binom{12}{3} = 400/924 \) for the probability.

**Example 2.9.** *(Ross, Example 5b (i))* If 3 balls are sampled without replacement from an urn containing 5 black balls and 6 white balls, what is the probability that the sample contains one white ball and two black balls?

**Solution:** This problem can be solved in two ways. If we take the order of sampling into account, then \( |S| = 11 \times 10 \times 9 = 990 \) and \( |E| = 6 \times 5 \times 4 + 5 \times 6 \times 4 + 5 \times 4 \times 6 = 360 \), so that the probability is \( 360/990 = 4/11 \). Alternatively, if we neglect the order, then \( S = \binom{11}{3} = 165 \) and
E = \binom{6}{1} \cdot \binom{5}{2} = 60, so the probability is 60/165 = 4/11.

**Example 2.10.** *(Ross, Example 5d)* An urn contains \( n - 1 \) red balls and 1 blue ball. If \( k \) of these balls are sampled without replacement, what is the probability that the sample contains the blue ball?

**Solution:** In this case \(|S| = \binom{n}{k}\) and \(|E| = \binom{n-1}{k-1} \cdot \binom{1}{1}\), so the probability is \(\binom{n-1}{k-1}/\binom{n}{k} = k/n\). Alternatively, let \( A_i \) be the event that the \( i \)th ball chosen is blue and notice that the \( A_i \) are mutually exclusive and that \( P(A_i) = 1/n \). Since \( E = A_1 \cup \cdots \cup A_k \), it follows that \( P(E) = k/n \).

**Example 2.11.** *(Ross, Example 5g)* A 5-card poker hand is a full house if it contains 3 cards of one denomination and 2 cards of a second denomination. What is the probability of being dealt a full house?

**Solution:** Note that there are \( \binom{52}{5} \) hands neglecting order. Also, there are \( 13 \cdot 12 \) (ordered) pairs of distinct denominations and, for each such pair, \( \binom{4}{3} \cdot \binom{4}{2} \) ways of being dealt three cards from the first denomination and two cards from the second. Thus the probability is

\[
\frac{13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}}{\binom{52}{5}} \approx 0.0014.
\]

**Example 2.12.** *(Ross, Example 5i)* If \( n \) people are present in a room, what is the probability that no two of them celebrate their birthday on the same day of the year?

**Solution:** The probability is

\[
\frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365-n+1)}{365^n} = \prod_{i=0}^{n-1} (1 - i/365).
\]

The smallest value of \( n \) for which this probability is less than \( 1/2 \) is \( n = 23 \).

**Example 2.13.** *(Ross, Example 5j)* Suppose that a deck of 52 cards is shuffled and that the cards are turned up one at a time until the first ace appears. Is the next card more likely to be an ace of spades or the two of clubs?

**Solution:** In both cases, the probability is \( \frac{1}{52} \).

**Example 2.14.** *(Ross, Example 5m)* Suppose that a deck of \( N \) cards is randomly shuffled. What is the probability that no card is left in place?

**Solution:** Let \( E_i \) be the event that the \( i \)th card is left in place. Using the inclusion-exclusion...
formula, the probability that at least one card is left in place is

\[
\Pr \left( \bigcup_{i=1}^{N} E_i \right) = \sum_{i=1}^{N} \Pr(E_i) - \sum_{i_1 < i_2} \Pr(E_{i_1} \cap E_{i_2}) + \cdots + (-1)^{n+1} \sum_{i_1 < \cdots < i_n} \Pr(\bigcap_k E_{i_k}) \\
+ \cdots + (-1)^{N+1} \Pr(E_1 \cap \cdots \cap E_N).
\]

Observe that

\[
\Pr(E_{i_1} \cap \cdots \cap E_{i_n}) = \frac{(N-n)!}{N!}
\]

since any of the \((N-n)!\) orders of the elements in \(\{1, \cdots, N\} - \{i_1, \cdots, i_n\}\) are possible. Also, as there are \(\binom{N}{n}\) terms in \(\sum_{i_1 < \cdots < i_n} \Pr(E_{i_1} \cap \cdots \cap E_{i_n})\), it follows that

\[
\sum_{i_1 < \cdots < i_n} \Pr(E_{i_1} \cap \cdots \cap E_{i_n}) = \binom{N}{n} \frac{(N-n)!}{N!} = \frac{1}{n!}.
\]

Thus

\[
\Pr \left( \bigcup_{i=1}^{N} E_i \right) = 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{N+1} \frac{1}{N!},
\]

and so the probability that no card is left in place is

\[
1 - 1 + \frac{1}{2!} - \frac{1}{3!} - \cdots + (-1)^N \frac{1}{N!}.
\]

Notice that this quantity is approximately \(e^{-1}\) when \(N\) is large.

**Example 2.15. (Ross, Example 5n)** What is the probability that if 10 married couples are seated at random in a circle, then no couple are seated next to each other?

If \(E_i\) is the event that the \(i\)'th couple are seated next to each other, then the desired probability is \(1 - \Pr(\cup_i E_i)\). As in the previous example, we will use the inclusion-exclusion formula. For this we need to calculate the probability \(\Pr(E_{i_1} \cap \cdots \cap E_{i_n})\). To do so, first notice that there are \(19!\) possible arrangements of the individuals around the table (seats are unlabeled). Next, notice that there are \(2^n \cdot (19-n)!\) arrangements that result in a specified group of \(n\) couples sitting together at the table: there are \(2^n\) orders of the individuals in the specified couples and \((19-n)!\) arrangements of the \(n\) couples and the remaining \(20-2n\) individuals. Consequently,

\[
\Pr(E_{i_1} \cap \cdots \cap E_{i_n}) = \frac{2^n(19-n)!}{19!}
\]

and so

\[
\Pr(\cup_i E_i) = \sum_{n=1}^{10} \binom{10}{n} (-1)^{n+1} \frac{2^n(19-n)!}{19!}.
\]
Chapter 3

Conditional Probabilities

3.1 Motivation

Suppose that two fair dice are rolled and that the sum of the two rolls is even. What is the probability that we have rolled a 1 with the first dice?

Without the benefit of any theory, we could address this question empirically in the following way. Suppose that we perform this experiment $N$ times, with $N$ large, and let $(x_i, y_i)$ be the outcome obtained on the $i$’th trial, where $x_i$ is the number rolled with the first die and $y_i$ is the number rolled with the second die. Then we can estimate the probability of interest (which we will call $P$) by dividing the number of trials for which $x_i = 1$ and $x_i + y_i$ is even by the total number of trials that have $x_i + y_i$ even:

$$P \approx \frac{\# \{i : x_i = 1 \text{ and } x_i + y_i \text{ even} \}}{\# \{i : x_i + y_i \text{ even} \}}$$

where we have divided both the numerator and denominator by $N$ in the second line.

Now, notice that for $N$ large, we expect the following approximations to hold:

$$P\{\text{the first roll equals 1 and the sum is even}\} \approx \frac{\# \{i : x_i = 1 \text{ and } x_i + y_i \text{ even} \}}{\# \{i : x_i + y_i \text{ even} \}}$$

$$P\{\text{the sum is even}\} \approx \frac{\# \{i : x_i + y_i \text{ even} \}}{N}$$

with these becoming identities as $N$ tends to infinity. This suggests that we can evaluate the probability $P$ exactly using the following equation:

$$P = \frac{P\{\text{the first roll equals 1 and the sum is even}\}}{P\{\text{the sum is even}\}}.$$ 

Here $P$ is said to be the conditional probability that the first roll is equal to 1 given that the sum of the two rolls is even. This identity motivates the following definition.

**Definition 3.1.** Let $E$ and $F$ be events and assume that $P(F) > 0$. Then the **conditional probability of $E$ given $F$** is defined by the ratio

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$
$F$ can be thought of as some additional piece of information that we have concerning the outcome of an experiment. The conditional probability $P(E|F)$ then summarizes how this additional information affects our belief that the event $E$ occurs. Notice that the ratio appearing in the definition is only well-defined if the numerator is positive, hence the important requirement that $P(F) > 0$.

**Example 3.1. (Ross, Example 2a)** A student is taking an exam. Suppose that the probability that the student finishes the exam in less than $x$ hours is $x/2$ for $x \in [0,2]$. Show that the conditional probability that the student does not finish the exam in one hour given that they are still working after 45 minutes is 0.8.

**Example 3.2. (Ross, Example 2b)** A fair coin is flipped twice. Show that the conditional probability that both flips lands on heads given that the first lands on heads is $1/2$. Show that the conditional probability that both flips land on heads given that at least one flip lands on heads is $1/3$.

The next result is a simple, but very useful consequence of the definition of conditional probability.

**Proposition 3.1. (Multiplication Rule)** Suppose that $E$ and $F$ are events and that $P(F) > 0$. Then

$$P(E \cap F) = P(E|F) \cdot P(F).$$

(3.1)

Similarly, if $E_1, \ldots, E_k$ are events and $P(E_1 \cap \cdots \cap E_{k-1}) > 0$, then recursive application of equation (3.1) shows that

$$P(E_1 \cap \cdots \cap E_k) = P(E_1) P(E_2|E_1) P(E_3|E_1 \cap E_2) \cdots P(E_k|E_1 \cap \cdots \cap E_{k-1}).$$

**Example 3.3. (Ross, Example 2d)** Suppose that $n$ balls are chosen without replacement from an urn that initially contains $r$ red balls and $b$ blue balls. Given that $k$ of the $n$ balls are blue, show that the conditional probability that the first ball chosen is blue is $k/n$.

**Solution:** Let $B$ be the event that the first ball chosen is blue and let $B_k$ be the event that $k$ of the $n$ balls chosen are blue. Then we need to calculate:

$$P(B|B_k) = \frac{P(B \cap B_k)}{P(B_k)}.$$

To evaluate the numerator, write

$$P(B_k \cap B) = P(B_k|B)P(B),$$

and notice that $P(B) = b/(b + r)$. Also, $P(B_k|B)$ is just the probability that a set of $n - 1$ balls chosen from an urn containing $r$ red balls and $b-1$ blue balls contains $k-1$ blue balls, which is equal to

$$P(B_k|B) = \frac{\binom{b-1}{k-1} \binom{r}{n-k}}{\binom{r+b-1}{n-1}}.$$

Similarly,

$$P(B_k) = \frac{\binom{b}{k} \binom{r}{n-k}}{\binom{r+b}{n}}.$$
It follows that
\[ P(B|B_k) = \left( \frac{b}{r+b} \right) \binom{b-1}{k-1} \binom{r}{n-k} \frac{(r+b)}{(r+b-1)} \binom{b}{n-k} \binom{r}{n} = \frac{k}{n}. \]

Example 3.4. (Ross, Example 2g) Suppose that a deck of \( N \) cards is randomly shuffled. In Example 2.14 of the previous chapter, we showed that the probability that no card is left in its original position is
\[ P_N = \sum_{n=0}^{N} (-1)^n \frac{1}{n!}. \]
Calculate the probability that exactly \( k \) cards are left in their original positions.

Solution: We first fix our attention on a particular set of \( k \) cards and let \( E \) be the event that only these cards are left in position and \( G \) be the event that none of the other \( N-k \) cards are left in position. Then
\[ P(E \cap G) = P(G|E)P(E). \]
To calculate \( P(E) \), let \( F_i \) be the event that the \( i \)’th member of this set of \( k \) cards is left in its original position and observe that
\[ P(E) = P(F_1 \cap F_2 \cap \cdots \cap F_k) = P(F_1)P(F_2|F_1)P(F_3|F_1 \cap F_2) \cdots P(F_k|F_1 \cap \cdots \cap F_{k-1}) \]
\[ = \frac{1}{N} \frac{1}{N-1} \frac{1}{N-2} \cdots \frac{1}{N-k+1} \]
\[ = \frac{(N-k)!}{N!}. \]
Given that those \( k \) cards are left in position, the probability that none of the remaining cards are left in position is just
\[ P(G|E) = P_{N-k} = \sum_{n=0}^{N-k} (-1)^n \frac{1}{n!} \]
and so
\[ P(E \cap G) = \frac{(N-k)!}{N!} P_{N-k}. \]
Now, to answer the original question, we need only note that there are \( \binom{N}{k} \) possible choices of \( k \) cards, and that the events that exactly those cards belonging to any one of these sets are the ones left in position are disjoint. It follows that
\[ P\{\text{exactly } k \text{ cards are fixed}\} = \binom{N}{k} \frac{(N-k)!}{N!} P_{N-k} \]
\[ = \frac{1}{k!} P_{N-k} \]
\[ \approx e^{-1} \frac{1}{k!}. \]
3.2 The Law of Total Probability

Let $E$ and $F$ be two events and notice that we can write $E$ as the following disjoint union:

$$E = (E \cap F) \cup (E \cap F^c).$$

Then, using the additivity of probabilities and the definition of conditional probabilities, we obtain:

$$P(E) = P(E \cap F) + P(E \cap F^c)$$

$$= P(E|F)P(F) + P(E|F^c)P(F^c)$$

$$= P(E|F)P(F) + P(E|F^c)(1 - P(F)).$$

(3.2)

This formula can sometimes be used to calculate the probability of a complicated event by exploiting the additional information provided by knowing whether or not $F$ has also occurred. The hard part is knowing how to choose $F$ so that the conditional probabilities $P(E|F)$ and $P(E|F^c)$ are both easy to calculate.

Example 3.5. (Ross, Example 3c) Suppose that there are two classes of people, those that are accident prone and those that are not, and that the risk of having an accident within a one year period is either 0.4 or 0.2 depending on which class a person belongs to. If the proportion of accident-prone people in the population is 0.3, calculate the probability that a randomly selected person has an accident in a given year (0.26). What is the conditional probability that an individual is accident-prone given that they have had an accident in the last year? (6/13).

Example 3.6. (Ross, Example 3d) Suppose that a blood test has a 95% chance of detecting a disease when it is present, and that it has a 1% rate of false positives. If 0.5% of the population has the disease, then the probability that a person has the disease given that their test result is positive is only 0.323.

Equation (3.2) has the following far-reaching generalization.

Proposition 3.2. (Law of Total Probability) Suppose that $F_1, \cdots, F_n$ are mutually exclusive events such that $P(F_i) > 0$ for each $i = 1, \cdots, n$ and $E \subset F_1 \cup \cdots \cup F_n$. Then

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i)$$

$$= \sum_{i=1}^{n} P(E|F_i) \cdot P(F_i).$$

Proof. The result follows from the countable additivity of $P$ and the fact that $E$ is the disjoint union of the sets $E \cap F_i$ for $i = 1, \cdots, n$. \qed

3.3 Bayes’ Formula

Bayes’ formula describes the relationship between the two conditional probabilities $P(E|F)$ and $P(F|E)$. 
3.3. **Bayes’ Formula**

**Theorem 3.1. (Bayes’ Formula)** Suppose that $E$ and $F$ are events such that $\mathbb{P}(E) > 0$ and $\mathbb{P}(F) > 0$. Then

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F) \cdot \mathbb{P}(F)}{\mathbb{P}(E)}.$$  \hspace{1cm} (3.3)

Similarly, if $F_1, \cdots, F_n$ are events such that $\mathbb{P}(F_i) > 0$ for $i = 1, \cdots, n$, then

$$\mathbb{P}(F_j|E) = \frac{\mathbb{P}(E|F_j) \cdot \mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E|F_i) \cdot \mathbb{P}(F_i)}.$$  \hspace{1cm} (3.4)

**Example 3.7. (Ross, Example 3l)** Suppose that we have three cards, and that the first card has two red sides, the second card has two black sides, and the third card has one red side and one black side. One of these cards is chosen at random and placed on the ground. If the upper side of this card is red, show that the probability that the other side is black is $1/3$.

**Example 3.8. (Ross, Example 3n)** A bin contains three types of flashlights (types 1, 2 or 3) in proportions 0.2, 0.3 and 0.5, respectively. Suppose that the probability that a flashlight will give over 100 hours of light is 0.7, 0.4, and 0.3 for type 1, type 2 or type 3.

a) Show that the probability that a randomly chosen flashlight gives over 100 hours of light is 0.41.

b) Given that a flashlight gives over 100 hours of light, show that the probability that it is a type $j$ flashlight is $\frac{14}{41}$ for type 1, $\frac{12}{41}$ for type 2 and $\frac{15}{41}$ for type 3.

**Example 3.9. (Ross, Exercise 3.26)** Suppose that 5% of men and 0.25% of women are color blind. A color-blind person is chosen at random. Assuming that there are an equal number of males and females in the population, what is the probability that the chosen individual is male?

**Solution:** Let $M$ and $F$ be the event that the chosen individual is a male or female, respectively, and let $C$ denote the event that a randomly chosen individual is color blind. Then the probability of interest is

$$\mathbb{P}(M|C) = \frac{\mathbb{P}(C|M) \mathbb{P}(M)}{\mathbb{P}(C|M) \mathbb{P}(M) + \mathbb{P}(C|F) \mathbb{P}(F)}$$

$$= \frac{0.05 \cdot 0.5}{0.05 \cdot 0.5 + 0.0025 \cdot 0.5}$$

$$= 0.9524.$$

### 3.3.1 Bayesian Statistics

Bayes’ formula is often applied to **problems of inference** in the following way. Suppose that our goal is to decide among a set of competing hypotheses, $H_1, \cdots, H_n$. Let $p_1, \cdots, p_n$ be a probability distribution on the hypotheses, i.e., $p_1 + \cdots + p_n = 1$, where $p_i$ is a measure of the strength of our belief in hypothesis $H_i$ prior to performing the experiment. In the jargon of Bayesian statistics, $(p_1, \cdots, p_n)$ is said to be the **prior distribution** on the space of hypotheses.

Now suppose that we perform an experiment to help decide between the hypotheses, and let the outcome of this experiment be the event $E$. Bayes’ formula can be used to update our belief in the relative probabilities of the different hypotheses in light of the new data. The **posterior probability**, $p_i^*$, of hypothesis $H_i$ is defined to be the conditional probability of $H_i$ given $E$:

$$p_i^* = \mathbb{P}(H_i|E) = \frac{\mathbb{P}(E|H_i) \mathbb{P}(H_i)}{\sum_{j=1}^{n} \mathbb{P}(E|H_j) \mathbb{P}(H_j)} = \frac{p_i \cdot \mathbb{P}(E|H_i)}{\sum_{j=1}^{n} p_j \cdot \mathbb{P}(E|H_j)}.$$
Notice that \((p_1^*, \ldots , p_n^*)\) is a probability distribution on the space of hypotheses that is called the **posterior distribution**. It depends both on the data collected in the experiment as well as on our prior belief about the likelihood of the different hypotheses.

**Example 3.10.** Suppose that we are studying a population containing \(n\) individuals and that our goal is to estimate the incidence of tuberculosis in the population. Let \(H_k\) be the hypothesis that \(k\) individuals are infected, and suppose that our prior distribution on \((H_0, \ldots , H_n)\) is given by

\[
p_k = \frac{1}{n+1}, \quad k = 0, 1, \ldots , n.
\]

Such a prior is said to be **uninformative** and means that we have no prior information concerning the incidence of tuberculosis in the population.

Now, suppose that the experiment consists of sampling one member of the population and testing them for tuberculosis infection. Let \(E\) be the event that that individual is infected and observe that

\[
P(E|H_k) = \frac{k}{n}.
\]

Using Bayes’ formula, the posterior probability of hypothesis \(H_i\) is

\[
p_i^* = \frac{p_i \cdot P(E|H_i)}{1/2} = \frac{1}{n+1} \cdot 2i = \frac{2i}{n(n+1)}
\]

for \(i = 0, \ldots , n\). Notice that \(p_i^* \geq p_i\) for \(i \geq n/2\), i.e., observing an infected individual in our sample increases the posterior probability that the incidence of tuberculosis is greater than 1/2.

### 3.3.2 Mendelian Genetics

Gregor Mendel’s 1866 paper proposing a mathematical model for heredity was one of the first applications of probability to the natural sciences. Although we now know that the genetic architecture of some traits is much more complicated than Mendel imagined, Mendelian genetics still provides a useful framework for understanding how many traits of interest, including thousands connected with hereditary human diseases, are inherited. In fact, these models play an important role both in genetic counseling and in forensic DNA analysis.

We begin with a list of important facts and concepts:

- Humans have two copies of most genes (except for those on the sex chromosomes).
- A child independently inherits one gene from each parent, i.e., one from the mother and one from the father.
- Each of the two gene copies in a parent is equally likely to be inherited.
- Different types of a gene are called **alleles**. These are often denoted by capital and lowercase letters, e.g., \(B\) and \(b\).
- A **genotype** is a list of the types of the two genes possessed by an individual.
• If there are two alleles, then there are three possible genotypes: $BB$, $bb$, and $Bb$. Notice that we usually ignore the order of alleles when writing the genotype, i.e., $Bb$ and $bB$ are the same genotype.

• The types and frequencies of the genotypes inherited by offspring can be calculated using a **Punnett square**. With two alleles, $B$ and $b$, the possible crosses are:
  - $BB \times BB$ produces only $BB$
  - $bb \times bb$ produces only $bb$
  - $BB \times Bb$ produces $\frac{1}{4}BB + \frac{1}{2}Bb$
  - $bb \times Bb$ produces $\frac{1}{4}BB + \frac{1}{2}Bb + \frac{1}{4}bb$
  - $BB \times bb$ produces only $Bb$.

**Example 3.11.** Assume that eye color is determined by a single pair of genes, which can be either $B$ or $b$. If a person has one or more copies of $B$, then they will have brown eyes. If a person has two copies of $b$, then they will have blue eyes. Suppose that Smith and his parents have brown eyes, and that Smith’s sister has blue eyes.

a) What is the probability that Smith carries a blue-eyed gene?
b) Suppose that Smith’s partner has blue eyes. What is the probability that their first child will have blue eyes?
c) Suppose that their first child has brown eyes. What is the probability that their second child will also have brown eyes?

**Solutions:**

a) Because Smith’s sister has blue eyes, her genotype is $bb$. Consequently, both of Smith’s parents (who are brown-eyed) must have genotypes $Bb$, since the sister must inherit a $b$ gene from each parent. However, Smith himself could have genotype $BB$ or $Bb$. Let $E_1$ and $E_2$ be the events that Smith has genotype $BB$ or $Bb$, respectively, and let $F$ be the event that Smith has brown eyes. Then the probability that Smith has genotype $Bb$ is

$$P(E_2|F) = \frac{P(F|E_2)P(E_2)}{P(F)} = \frac{1 \cdot 1/2}{3/4} = \frac{2}{3},$$

while the probability that Smith has genotype $BB$ is $P(E_1|F) = 1 - P(E_2|F) = 1/3$.

b) If Smith’s partner has blue eyes, then her genotype is $bb$. Now if Smith has genotype $BB$, then his children can only have genotype $Bb$ and therefore have zero probability of having blue eyes. On the other hand, if Smith has genotype $Bb$, then his first child will have blue eyes with probability $1/2$ (i.e., with the probability that the child inherits a $b$ gene from Smith). Thus, the total probability that the first child has blue eyes is:

$$1/3 \cdot 0 + 2/3 \cdot 1/2 = 1/3.$$

c) Knowing that their first child has brown eyes provides us with additional information about the genotypes of the parents. Let $G$ be the event that the first child has brown eyes and let $E_1$
and $E_2$ be the events that Smith has genotype $BB$ or $Bb$, respectively. From (a) we know that $\Pr(E_1) = 1/3$ and $\Pr(E_2) = 2/3$, while (b) tells us that $\Pr(G) = 1 - 1/3 = 2/3$. Consequently,

$$\Pr(E_1|G) = \frac{\Pr(G|E_1)\Pr(E_1)}{\Pr(G)} = \frac{1 \cdot 1/3}{2/3} = 1/2,$$

$$\Pr(E_2|G) = 1 - \Pr(E_1|G) = 1/2.$$

Let $H$ be the event that the second child has brown eyes and notice that

$$\Pr(H|E_1, G) = 1$$
$$\Pr(H|E_2, G) = 1/2.$$

Then,

$$\Pr(H|G) = \Pr(H \cap E_1|G) + \Pr(H \cap E_2|G)$$
$$= \Pr(H|E_1, G)\Pr(E_1|G) + \Pr(H|E_2, G)\Pr(E_2|G)$$
$$= 1 \cdot 1/2 + 1/2 \cdot 1/2 = 3/4.$$

### 3.4 Independence

We begin with the definition of pairwise independence, i.e., of what it means for two events to be independent. In the sequel, we will extend this to more than two sets.

**Definition 3.2.** Two events $E$ and $F$ are said to be **independent** if

$$\Pr(E \cap F) = \Pr(E)\Pr(F).$$

Intuitively, $E$ and $F$ are independent if knowing that $F$ has occurred provides no information about the likelihood that $E$ has occurred and *vice versa*:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E)$$
$$\Pr(F|E) = \frac{\Pr(E \cap F)}{\Pr(E)} = \frac{\Pr(E)\Pr(F)}{\Pr(E)} = \Pr(F).$$

**Remark 3.1.** Independence is a symmetric relation: to say that $E$ is independent of $F$ also means that $F$ is independent of $E$.

**Example 3.12.** *(Ross, Example 4b)* Suppose that two coins are tossed and that all four outcomes are equally likely. Let $E$ be the event that the first coin lands on heads and let $F$ be the event that the second coin lands on heads. Then, as expected, $E$ and $F$ are independent since

$$\Pr(E \cap F) = \frac{1}{4} = \Pr(E)\Pr(F).$$

**Example 3.13.** *(Ross, Example 4c)* Suppose that we toss two fair dice. Let $E_1$ be the event that the sum of the dice is 6 and let $F$ be the event that the first die equals 4. Then,

$$\Pr(E_1 \cap F) = \frac{1}{36} \neq \frac{5}{36} \cdot \frac{1}{6} = \Pr(E_1)\Pr(F),$$
and thus $E_1$ and $F$ are not independent. On the other hand, if we let $E_2$ be the event that the sum of the dice is 7, then
\[ P(E_2 \cap F) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = P(E_2)P(F), \]
and so $E_2$ and $F$ are independent.

**Proposition 3.3.** If $E$ and $F$ are independent, then so are each of the following three pairs: $E$ and $F^c$, $E^c$ and $F$, and $E^c$ and $F^c$.

**Proof.** This follows by direct calculation of the joint probabilities. For example,
\[
P(E \cap F^c) = P(E) - P(E \cap F)
= P(E) - P(E) \cdot P(F)
= P(E) \cdot P(F^c),
\]
which shows that $E$ and $F^c$ are independent. \(\square\)

As promised, we can extend the definition of independence to collections of more than two events. Furthermore, in the example that follows the definition, we show that independence of a collection of objects can be a strictly stronger condition than mere independence of each pair of objects.

**Definition 3.3.** A countable collection of events, $E_1, E_2, \ldots$, is said to be independent if for every finite sub-collection $E_{i_1}, E_{i_2}, E_{i_3}, \ldots, E_{i_n}$, we have:
\[
P \left( \bigcap_{r=1}^{n} E_{i_r} \right) = \prod_{r=1}^{n} P(E_{i_r}).
\]

**Example 3.14.** **Pairwise independence does not imply independence:** Suppose that two fair dice are tossed and let $E_1$ be the event that the first die equals 4, let $E_2$ be the event that the second die equals 4, and let $E_3$ be the event that the sum of the dice is 7. As above, we know that each pair of events is independent. However, it is not the case that all three events are independent since
\[
P(E_1 \cap E_2 \cap E_3) = 0 \neq \frac{1 \cdot 1}{6 \cdot 6} = P(E_1)P(E_2)P(E_3).
\]

**Example 3.15.** (Ross, Example 4f) Suppose that an infinite sequence of independent trials is performed such that each trial results in a success with probability $p$ and a failure with probability $q = 1 - p$. Calculate the probability that:

- (a) at least one success occurs in the first $n$ trials;
- (b) exactly $k$ successes occur in the first $n$ trials;
- (c) all trials result in successes.
3.4.1 The Gambler’s Ruin Problem

The following is a classic problem of probability that dates back to the 17’th century.

Two gamblers, A and B, bet on the outcomes of successive flips of a coin. On each flip, A wins a dollar from B if the coin comes up heads, while B wins a dollar from A if the coin comes up tails. The game continues until one of the players runs out of money. Suppose that the coin tosses are mutually independent and that each toss turns up heads with probability \( p \). If A starts with \( i \) dollars and B starts with \( N-i \) dollars, what is the probability that A ends up with all of the money?

**Solution:** Let \( E \) be the event that A eventually wins all of the money and let 
\[
P_i = \mathbb{P}(E|A \text{ starts with } i \text{ dollars})
\]
be the conditional probability of \( E \) given that A starts with \( i \) dollars. We will use the independence of the successive coin tosses to obtain a system of linear equations for the probabilities \( P_0, P_1, \ldots, P_N \). To this end, let \( H \) be the event that the first toss lands on heads and observe that
\[
P_i = \mathbb{P}(E|H)\mathbb{P}(H) + \mathbb{P}(E|H^c)\mathbb{P}(H^c) = p\mathbb{P}(E|H) + (1-p)\mathbb{P}(E|H^c) = pP_{i+1} + (1-p)P_{i-1},
\]
where \( P_0 = 0 \) (A loses if she has no money to wager). Thus,
\[
P_2 - P_1 = \alpha (P_1 - P_0) = \alpha P_1
\]
\[
P_3 - P_2 = \alpha (P_2 - P_1) = \alpha^2 P_1
\]
\[
\ldots
\]
\[
P_i - P_{i-1} = \alpha (P_{i-1} - P_{i-2}) = \alpha^{i-1} P_1
\]
\[
\ldots
\]
\[
P_N - P_{N-1} = \alpha (P_{N-1} - P_{N-2}) = \alpha^{N-1} P_1.
\]
Adding the first \( i - 1 \) equations gives
\[
P_i - P_1 = P_1 \sum_{k=1}^{i-1} \alpha^k,
\]
and therefore
\[
P_i = \begin{cases} 
  P_1 \left( \frac{1 - \alpha^i}{1 - \alpha} \right) & \text{if } p \neq 1/2 \ (\alpha \neq 1) \\
  P_1 i & \text{if } p = 1/2 \ (\alpha = 1).
\end{cases}
\]

Then, using the fact that \( P_N = 1 \) (\( A \) wins if she begins with all of the money), we can solve for \( P_1 \):
\[
P_1 = \begin{cases} 
  \frac{1 - \alpha}{1 - \alpha^N} & \text{if } p \neq 1/2 \\
  \frac{1}{N} & \text{if } p = 1/2.
\end{cases}
\]

Substituting these results back into the expression for \( P_i \), we obtain
\[
P_i = \begin{cases} 
  \frac{1 - (q/p)^i}{1 - (q/p)^N} & \text{if } p \neq 1/2 \\
  \frac{i}{N} & \text{if } p = 1/2.
\end{cases}
\]

If we let \( Q_i \) denote the probability that \( B \) ends up with all the money when the players have \( i \) and \( N - i \) dollars at the start of the game, then by the symmetry of the problem, we see that
\[
Q_i = \begin{cases} 
  \frac{1 - (p/q)^{N-i}}{1 - (p/q)^N} & \text{if } p \neq 1/2 \\
  \frac{N-i}{N} & \text{if } p = 1/2.
\end{cases}
\]

In particular, a simple calculation shows that \( P_i + Q_i = 1 \), i.e., with probability 1, either \( A \) or \( B \) eventually wins all of the money. In other words, the probability that the game continues indefinitely with neither \( A \) nor \( B \) winning is 0.

**Remark 3.2.** If we let \( X_n \) denote \( A \)'s holdings after the \( n \)th coin toss, then the random sequence \((X_n : n \geq 0)\) is an example of a Markov chain and, in the case \( p = q = 1/2 \), of a martingale.

### 3.5 Conditional Probabilities are Probabilities

**Theorem 3.2.** Let \( F \) be an event in a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) such that \( \mathbb{P}(F) > 0 \). Then the conditional probability \( \mathbb{P}(\cdot|F) \) is a probability measure on \((\Omega, \mathcal{F})\).

**Proof.** We need to show that \( \mathbb{P}(\cdot|F) \) satisfies the three axioms of a probability measure. The first axiom requires that for any event \( E \),
\[
0 \leq \mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} \leq 1,
\]
which follows because \( \mathbb{P}(E \cap F), \mathbb{P}(F) > 0 \) and because \( \mathbb{P}(E \cap F) \leq \mathbb{P}(F) \) (since \( E \cap F \subset F \)).
The second axiom follows by a simple calculation:
\[ P(\Omega | F) = \frac{P(\Omega \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1. \]

Lastly, let \( E_1, E_2, \cdots \) be a countable collection of disjoint events and let \( E = \bigcup_{n=1}^{\infty} E_n \). Then
\[
P(E | F) = \frac{P(E \cap F)}{P(F)}
= \frac{1}{P(F)} P\left( \bigcup_{n=1}^{\infty} (E_n \cap F) \right)
= \frac{1}{P(F)} \sum_{n=1}^{\infty} P(E_n \cap F)
= \sum_{n=1}^{\infty} P(E_n | F),
\]
where in passing from the second line to the third we have made use of the countable additivity of \( P(\cdot) \) and of the fact that the events \( E_n \cap F, n \geq 1 \) are disjoint. This verifies that \( P(\cdot | F) \) is countably additive, which completes the proof.

One useful consequence of Proposition 5.1 is that we can condition on a series of events, e.g., if \( E, F, \) and \( G \) are events and we define
\[ Q(E) = P(E | F), \]
then the conditional probability \( Q(E | G) \) is well-defined. In fact,
\[
Q(E | G) = Q(E \cap G) = \frac{Q(E \cap G)}{Q(G)}
= \frac{P(E \cap G | F)}{P(G | F)}
= \left( \frac{P(E \cap G \cap F)}{P(F)} \right) \cdot \left( \frac{P(F)}{P(G \cap F)} \right)
= \frac{P(E \cap F \cap G)}{P(F \cap G)}
= \frac{P(E \cap F \cap G)}{P(E | F \cap G)},
\]
which shows that conditioning first on \( F \) and then on \( G \) is equivalent to conditioning on the intersection of \( F \) and \( G \). Furthermore, this correspondence allows us to rewrite the identity
\[ Q(E) = Q(E | G)Q(G) + Q(E | G^c)Q(G^c) \]
in the useful form
\[ P(E | F) = P(E | F \cap G)P(G | F) + P(E | F \cap G^c)P(G^c | F). \]

**Example 3.16.** (Ross, Example 5b) (See the primer on Mendelian genetics in Section 3.3 for definitions.) A female chimp gives birth to a single offspring, but it is unknown which of two male chimps is the father. Before any genetic data is collected, it is believed that the probability
that male number 1 is the father is \( p \) and that the probability that male number 2 is the father is \( 1 - p \). (Recall that \((p, 1 - p)\) is the prior distribution on the two hypotheses concerning paternity.) DNA is collected from all four individuals and sequenced at a single site on the genome, showing that the maternal genotype is AA, the genotype of male 1 is aa, the genotype of male 2 is Aa, and that the genotype of the offspring is Aa. What is the posterior probability that male 1 is the father?

**Solution:** Let all probabilities be conditional on the genotypes of the adults and define \( M_i, i = 1, 2 \) to be the event that male \( i \) is the father and \( G \) to be the event that the offspring genotype is Aa. Then

\[
\Pr(M_1|G) = \frac{\Pr(M_1 \cap G)}{\Pr(G)} = \frac{\Pr(G|M_1)\Pr(M_1)}{\Pr(G|M_1)\Pr(M_1) + \Pr(G|M_2)\Pr(M_2)} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{2} \cdot (1 - p)} = \frac{2p}{1 + p}
\]

Notice that \( \Pr(M_1|G) > p = \Pr(M_1) \) whenever \( p < 1 \), i.e., the observed genotypes increase the likelihood that male 1 is the father.

**Example 3.17.** *(Ross, Example 5c)* Suppose that a series of independent trials is conducted and that each trial results in a success with probability \( p \) and a failure with probability \( q = 1 - p \). What is the probability that a run of \( n \) consecutive successes occurs before a run of \( m \) consecutive failures?

**Solution:** Let us introduce the following events

\[
E = \{ \text{a run of } n \text{ successes happens before a run of } m \text{ failures} \}, \\
S = \{ \text{the first trial results in a success} \}, \\
S_{n-1} = \{ \text{trials 2 through } n \text{ all result in successes} \}, \\
F_{m-1} = \{ \text{trials 2 through } m \text{ all result in failures} \}.
\]

By the law of total probability, we have

\[
\Pr(E) = \Pr(E|S)\Pr(S) + \Pr(E|S^c)\Pr(S^c) = p \cdot \Pr(E|S) + q \cdot \Pr(E|S^c).
\]

The conditional probabilities in this last expression can be evaluated in the following way. First observe that

\[
\Pr(E|S) = \Pr(E|S_{n-1} \cap S)\Pr(S_{n-1}|S) + \Pr(E|S_{n-1}^c \cap S)\Pr(S_{n-1}^c|S) = p^{n-1} + (1 - p^{n-1})\Pr(E|S_{n-1}^c \cap S),
\]

since \( S_{n-1} \) and \( S \) are independent and \( \Pr(E|S_{n-1} \cap S) = 1 \) (\( n \) consecutive successes occur on the first \( n \) trials if \( S \) and \( S_{n-1} \) are both true). However, since the event \( S_{n-1}^c \cap S \) means that a failure has occurred during at least one of the trials 2 through \( n \) and since whether \( E \) occurs or not does not depend on any event happening before this first failure, it follows that

\[
\Pr(E|S_{n-1}^c \cap S) = \Pr(E|S^c),
\]
i.e., it is as if the failure occurred on the very first trial. Thus
\[ P(E|S) = p^{n-1} + (1 - p^{n-1})P(E|S^c). \]

Similar reasoning shows that
\[ P(E|S^c) = P(E|F_{m-1} \cap S^c)P(F_{m-1}^c|S^c) + P(E|F_{m-1}^c \cap S^c)P(E|F_{m-1}^c|S^c) \]
\[ = 0 + (1 - q^{m-1})P(E|F_{m-1} \cap S) \]
\[ = (1 - q^{m-1})P(E|S), \]
where we have used the fact that \( P(E|F_{m-1} \cap S^c) = 0 \) (since then \( m \) consecutive failures occur during the first \( m \) trials) and that \( P(E|F_{m-1}^c \cap S) = P(E|S) \).

By combining the equations for \( P(E|S) \) and \( P(E|S^c) \), we obtain
\[ P(E|S) = p^{n-1} + (1 - p^{n-1})(1 - q^{m-1})P(E|S) \]
which then implies that
\[ P(E|S) = \frac{p^{n-1}}{1 - (1 - p^{n-1})(1 - q^{m-1})} \]
\[ P(E|S^c) = \frac{p^{n-1}(1 - q^{m-1})}{1 - (1 - p^{n-1})(1 - q^{m-1})}. \]

Finally, by substituting these conditional probabilities into the initial equation derived for \( P(E) \), we arrive at the answer to our problem:
\[ P(E) = \frac{p^n + p^{n-1}(q - q^m)}{1 - (1 - p^{n-1})(1 - q^{m-1})} = \frac{p^{n-1}(1 - q^m)}{1 - (1 - p^{n-1})(1 - q^{m-1})}. \]
Chapter 4

Discrete Random Variables

4.1 Introduction

Example 4.1. Suppose that we have a biased coin that lands on heads with probability $p$ and that we toss this coin a thousand times. If $X$ is the total number of heads that appear, then $X$ is a random number between 0 and 1000 that depends on the outcome of the thousand tosses and

$$
\mathbb{P}(X = n) = \binom{1000}{n} p^n (1 - p)^{1000-n}.
$$

Example 4.2. Suppose that we are charged with developing a management plan for saguaro cacti in South Mountain Park. Because there are many factors (e.g., climate, topography, other plant and animal species) that may influence the population size and distribution of these cacti, the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ representing our uncertainty about the system will be extremely large and complicated. On the other hand, if $X$ denotes the number of saguaros present in our study area, then $X$ is a random non-negative integer that depends on many other unobserved processes, but which we can determine by conducting a survey in the study area. Similarly, if $Y$ denotes the average height of the individuals in the sample, then $Y$ is a random non-negative real number.

These examples illustrate a common situation: often, what we know about the state of a complex system is limited to the indirect knowledge provided by experimentation. This can be described mathematically by supposing that there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that captures all of the uncertainty that we have concerning the state of the system and then introducing functions $X, Y, \cdots$ that represent the outcomes of the experiments that we perform on the system. Although these outcomes depend on the actual state of the system, which is some point $\omega \in \Omega$, what we observe is that the first experiment results in measurement $X(\omega)$, the second experiment results in measurement $Y(\omega)$, etc. The functions $X, Y, \cdots$ are said to be random variables and are defined as follows.

Definition 4.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and suppose that $E$ is a set and that $\mathcal{E}$ is a collection of subsets of $E$. An $E$-valued random variable is a function $X : \Omega \to E$ with the property that whenever $A \in \mathcal{E}$, then

$$
\{X \in A\} \equiv X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}.
$$

The distribution of $X$, which we denote $\mathbb{P}_X$, is the probability distribution on $(E, \mathcal{E})$ defined by

$$
\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in A\}\right), \quad A \in \mathcal{E}.
$$
The assertion that $\mathbb{P}_X$ is a probability distribution on $(E, \mathcal{E})$ can be proved as follows. First, since $\mathbb{P}_X$ is defined in terms of $\mathbb{P}$, it is clear that $\mathbb{P}_X(A) \in [0, 1]$ for any event $A \subset E$. Secondly,

$\mathbb{P}_X(E) = \mathbb{P}(X \in E) = \mathbb{P}(\Omega) = 1$.

Lastly, suppose that $A_1, A_2, \cdots$ is a countable collection of mutually exclusive events in $E$ and let $A = \bigcup_{n=1}^{\infty} A_n$. Then, $\{X \in A_1\}, \{X \in A_2\}, \cdots$ is a countable collection of mutually exclusive events in $\Omega$, and so countable additivity of $\mathbb{P}$ implies that

\[
\mathbb{P}_X(A) = \mathbb{P}_X \left( \bigcup_{n=1}^{\infty} A_n \right) \\
= \mathbb{P} \left( X \in \bigcup_{n=1}^{\infty} A_n \right) \\
= \mathbb{P} \left( \bigcup_{n=1}^{\infty} \{X \in A_n\} \right) \\
= \sum_{n=1}^{\infty} \mathbb{P}(X \in A_n) \\
= \sum_{n=1}^{\infty} \mathbb{P}_X(A_n).
\]

This shows that $\mathbb{P}_X$ is countably additive and completes the proof that $\mathbb{P}_X$ is a probability distribution.

**Remark 4.1.** Although any random variable $X$ is a function defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, in applications we can often ignore the sample space altogether and focus exclusively on $X$ and its distribution $\mathbb{P}_X$. This has the advantage of freeing us from having to worry about the usually complicated and high-dimensional structure of the sample space. In particular, we will usually just write $X$ and suppress the fact that the value that $X$ takes in any experiment depends on $\omega$.

Although Definition 4.1 encompasses random variables that take values in any ‘reasonable’ set $E$ (e.g., there are matrix-valued random variables, function-valued random variables, set-valued random variables, etc.), in this course we will mainly study three classes of random variables: those that take values in a countable set $E$ (discrete random variables), those that take values in a subset of the real numbers (real-valued random variables), and those that take values in a subset of $n$-dimensional Euclidean space $\mathbb{R}^n$ (random vectors). For the most part, we will leave the collection of events $\mathcal{E}$ unspecified, but we will require that any ‘reasonable’ subset of $E$ is an event. In particular, if $E = \mathbb{R}$, we will require that all of the following sets are included in $\mathcal{E}$: $(a, b)$, $[a, b)$, $(a, b]$, and $[a, b]$, where $-\infty \leq a \leq b \leq \infty$, as well as the complements and countable unions and countable intersections of such sets. If $X$ is a real-valued random variable, then this will allow us to talk about the probability of the event $X \leq b$, for example, since

$\{X \leq b\} = \{X \in (-\infty, b]\}$,

and $(-\infty, b] \in \mathcal{E}$ by assumption. In fact, the probabilities of these kinds of events are at the heart of the next definition.

**Definition 4.2.** The cumulative distribution function (or CDF) of a real-valued random variable $X$ is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$F_X(x) = \mathbb{P}(X \leq x).$
4.2. DISCRETE RANDOM VARIABLES

The most important fact about the CDF \( F \) of a real-valued random variables \( X \) is that it uniquely determines the distribution of \( X \), i.e., if \( Y \) is another real-valued random variable with CDF \( F \), then \( Y \) and \( X \) have the same distribution. However, this does not necessarily means that \( X = Y \), since \( X \) and \( Y \) could be defined on different probability spaces.

4.2 Discrete Random Variables

**Definition 4.3.** \( X \) is said to be a **discrete random variable** if \( X \) takes values in a countable set \( E \). \( E \) can either be finite, in which case \( E = \{x_1, \cdots, x_n\} \), or countably infinite, in which case \( E = \{x_1, x_2, \cdots\} \). In either case, we will implicitly take \( E \) to be the power set of \( E \), i.e., \( E \) is the collection of all subsets of \( E \).

The distribution of a discrete random variable \( X \) is uniquely determined by its **probability mass function** (abbreviated p.m.f.), which is the function \( p_X : E \to [0, 1] \) defined by

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

To see that \( p_X \) determines the distribution of \( X \), let \( A = \{x_j : j \in J\} \) be a subset of \( E \) and observe that

\[
P(X \in A) = P(X \in \bigcup_{j \in J} \{x_j\}) = \sum_{j \in J} P(X = x_j) = \sum_{j \in J} p_X(x_j).
\]

In other words, to calculate the probability that a discrete random variable takes values in a set \( A \), we only need to sum the values of the probability mass function evaluated on each element of \( A \). In particular, by taking \( A = E \) in the above identities, we see that

\[
1 = P(X \in E) = \sum_{x \in E} p_X(x),
\]

i.e., the values of the probability mass function sum to 1. Furthermore, if \( X \) is a discrete real-valued random variables, i.e., \( E = \{x_1, x_2, \cdots\} \subset \mathbb{R} \), then the cumulative distribution function of \( X \) can be calculated from the probability mass function using the identity

\[
F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p_X(X_i).
\]

This shows that \( F_X(x) \) is a **step function** with a discontinuity of size \( p(x_i) \) at each point \( x_i \in E \).

**Example 4.3.** Suppose that \( X \) takes values in \( E = \{0, 1\} \) and that \( p_X(1) = p \) and \( p_X(0) = 1 - p \). Then \( X \) is said to be a **Bernoulli** random variable with parameter \( p \) and its c.d.f. is

\[
F_X(x) = \begin{cases} 
0 & \text{if } x < 0 \\
1 - p & \text{if } 0 < x < 1 \\
1 & \text{if } x \geq 1.
\end{cases}
\]
Example 4.4. If $A$ is an event in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the indicator function of $A$ is the function $1_A : \Omega \to \{0, 1\}$ defined by

$$1_A(\omega) = \begin{cases} 0 & \text{if } \omega \notin A \\ 1 & \text{if } \omega \in A. \end{cases}$$

In other words, the indicator function of an event $A$ ‘indicates’ whether that event does or does not occur when a random point is chosen in the sample space. Furthermore, $1_A$ is a Bernoulli random variable with parameter

$$p = \mathbb{P}(1_A = 1) = \mathbb{P}(\omega \in A) = \mathbb{P}(A).$$

Example 4.5. Suppose that $X$ is a discrete random variable with values in the non-negative integers $\mathbb{N} = \{0, 1, \cdots\}$ and probability mass function

$$p_X(n) = C \cdot p^n,$$

where $C$ is a constant to be determined and $p \in (0, 1)$. Using the fact that $\mathbb{P}(X \in \mathbb{N})$, we obtain

$$1 = \sum_{n=0}^{\infty} C \cdot p^n = \frac{C}{1-p},$$

which shows that $C = (1-p)$. $X$ is said to have a geometric distribution and belongs to one of the classes of discrete random variables that we will study later in the chapter. The probability that $X$ is even is equal to

$$\mathbb{P}(X \text{ is even}) = \sum_{n=0}^{\infty} p_X(2n) = \sum_{n=0}^{\infty} (1-p) \cdot p^{2n} = \frac{1-p}{1-p^2} = \frac{1}{1+p}.$$

### 4.3 Expectations of Discrete Random Variables

**Definition 4.4.** Let $X$ be a discrete random variable with values in the set $E = \{x_1, x_2, \cdots\}$ and probability mass function $p_X(x)$. The expected value (also called the expectation or the mean) of $X$ is defined to be the weighted average of the values that $X$ can assume:

$$\mathbb{E}[X] = \sum_{x_i} p_X(x_i) \cdot x_i.$$ 

**Example 4.6.** Let $X$ be a Bernoulli random variable with parameter $p$ (cf. Example 4.3). Then the expected value of $X$ is

$$\mathbb{E}[X] = (1-p) \cdot 0 + p \cdot 1 = p.$$ 

**Example 4.7.** Let $X$ be the outcome when we roll a fair die. Then

$$\mathbb{E}[X] = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{3}{2}.$$
Notice that in both of these examples, the random variable $X$ cannot actually be equal to its expected value. In this sense, the name ‘expected value’ is somewhat misleading and it may be more accurate to refer to $\mathbb{E}[X]$ as the mean of $X$. In fact, towards the end of the course we will show that if $X_1, X_2, \cdots$ are i.i.d. real-valued random variables with the same distribution as $X$, then

$$\frac{1}{N} \sum_{i=1}^{N} X_i \to \mathbb{E}[X]$$

as $N \to \infty$, i.e., the sample mean tends to the expected value as the size of the sample increases.

One way to obtain new random variables is to compose a real-valued function with an existing random variable. Let $X$ be a discrete random variable that takes values in the set $E = \{x_1, x_2, \ldots\}$ with probabilities $p_X(x_i)$, and let $g : E \to \mathbb{R}$ be a real-valued function. Then $Y = g(X)$ is a discrete random variable with values in the countable set $g(E) = \{y_1, y_2, \cdots\}$ and the probability mass function of $Y$ is

$$p_Y(y_i) \equiv \mathbb{P}(Y = y_i) = \sum_{x_j : g(x_j) = y_i} p_X(x_j).$$

Notice that more than one value $x_i$ may contribute to the probability of each point $y_j$ if $g$ is not one-to-one.

**Example 4.8.** Suppose that $X$ is a discrete random variable that takes values in the set $\{-1, 0, 1\}$ with probabilities

$$p_X(-1) = \frac{1}{4}, \quad p_X(0) = \frac{1}{2}, \quad p_X(1) = \frac{1}{4}.$$

Then $Y = X^2$ is a discrete random variable that takes values in the set $\{0, 1\}$ with probabilities

$$p_Y(0) = p_X(0) = \frac{1}{2},$$

$$p_Y(1) = p_X(-1) + p_X(1) = \frac{1}{2},$$

and so

$$\mathbb{E}[X^2] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}.$$

The following proposition gives us a way to compute $\mathbb{E}[g(X)]$ without having to calculate the p.m.f. of $g(X)$.

**Proposition 4.1.** Let $X$ be a discrete random variable with probability mass function $p_X(x_i)$ on $E$. Then, for any real-valued function $g : E \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \sum_{i} p_X(x_i) \cdot g(x_i).$$

**Proof.** Suppose that $Y = g(X)$ takes on the distinct values $y_1, y_2, \cdots$ with probabilities $p_Y(y_j)$ and let

$$E_j = g^{-1}(y_j) = \{x_i \in E : g(x_i) = y_j\}.$$
Then the sets $E_1, E_2, \cdots$ are disjoint with union $E = E_1 \cup E_2 \cup \cdots$, and

$$
\mathbb{E}[g(X)] = \sum_j p_Y(y_j) \cdot y_j
$$

$$
= \sum_j \left( \sum_{x_i \in E_j} p_X(x_i) \right) \cdot y_j
$$

$$
= \sum_j \sum_{x_i \in E_j} p_X(x_i) \cdot y_j
$$

$$
= \sum_{x_i \in E} p_X(x_i) \cdot y_j
$$

$$
= \sum_{x_i \in E} p_X(x_i) \cdot g(x_i).
$$

An important consequence of Proposition 4.1 is that expectations are linear.

**Corollary 4.1.** If $a$ and $b$ are constants, then

$$
\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.
$$

**Proof.** A simple calculation shows that

$$
\mathbb{E}[aX + b] = \sum_{x_i} p_X(x_i) \cdot (ax_i + b)
$$

$$
= a \sum_{x_i} p_X(x_i) \cdot x_i + b \sum_{x_i} p_X(x_i)
$$

$$
= a\mathbb{E}[X] + b
$$

**Remark 4.2.** For non-linear functions $g$, it is usually true that

$$
\mathbb{E}[g(X)] \neq g(\mathbb{E}[X]),
$$

i.e., we cannot interchange the composition of functions with the taking of expectations.

**Example 4.9.** Let $X$ be a random variable with $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Then

$$
\mathbb{E}[X^n] = \begin{cases} 
1 & \text{if } n \text{ is even} \\
0 & \text{if } n \text{ is odd}.
\end{cases}
$$

However, $\mathbb{E}[X]^n = 0$ for all $n \geq 1$.

### 4.4 Variance

The variance of a random variable is one measure of the dispersion or spread of that variable about its mean.
Definition 4.5. Let $X$ be a random variable with mean $\mu = \mathbb{E}[X]$. Then the variance of $X$ is the quantity

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

In other words, the variance of $X$ is just the mean squared deviation of $X$ from its expectation. If the variance of a random variable is very small, then this indicates that the random variable usually takes on values that are very close to this expectation. On the other hand, if the variance is large, then this indicates that, at least sometimes, the variable may take on values that are very different from the expectation.

Although the formula given in Definition 4.5 can be used to calculate the variance of a random variable, the following expression is often easier to evaluate:

$$\mathbb{E}[(X - \mu)^2] = \sum_i p_X(x_i)(x_i - \mu)^2$$
$$= \sum_i p_X(x_i)(x_i^2 - 2\mu x_i + \mu^2)$$
$$= \sum_i p_X(x_i) \cdot x_i^2 - 2\mu \cdot \sum_i p_X(x_i) \cdot x_i + \mu^2$$
$$= \mathbb{E}[X^2] - 2\mu^2 + \mu^2$$
$$= \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

In other words, the variance of $X$ is equal to the expected value of $X^2$ minus the square of the expected value of $X$.

Example 4.10. Let $X$ be a Bernoulli random variable with parameter $p$. In Example 4.6 we showed that the expected value of $X$ is $p$. Since $X$ takes values in the set $\{0, 1\}$, $X^2 = X$ and so the variance of $X$ is

$$\text{Var}(X) = \mathbb{E}[X^2] - p^2 = p - p^2 = p(1 - p).$$

Notice that the variance is greatest when $p = 1/2$ and is equal to 0 when $p = 0$ or 1. In the latter two cases, we either always have $X = 0$ or $X = 1$ and so $X$ is effectively non-random.

Example 4.11. Let $X$ be the outcome when a fair die is rolled. Then

$$\text{Var}(X) = \mathbb{E}[X^2] - (7/2)^2$$
$$= \frac{1}{6} \sum_{i=1}^6 i^2 - \frac{49}{4}$$
$$= \frac{35}{12}.$$

The counterpart to Corollary 4.1 for the variance is:

Corollary 4.2. If $a$ and $b$ are constants, then

$$\text{Var}(aX + b) = a^2 \text{Var}(X).$$
Proof. To prove this, let \( Y = aX + b \) and notice that Corollary 4.1 tells us that \( \mathbb{E}[Y] = a\mathbb{E}[X] + b \). Consequently,

\[
\begin{align*}
\text{Var}(Y) & = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \\
& = \mathbb{E}[(aX + b - (a\mathbb{E}[X] + b))^2] \\
& = \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \\
& = a^2\text{Var}(X).
\end{align*}
\]

In other words, the variance of a random variable depends on the scale of the units in which we measure the variable, but not the location of the units.

4.5 Probability Generating Functions

The next definition may seem mysterious at first, but we will see a number of uses for probability generating functions during the semester.

Definition 4.6. Let \( X \) be a random variable with values in the non-negative integers and probability mass function \( p(k) = \mathbb{P}(X = k) \). The probability generating function (PGF) of \( X \) is the function \( \psi_X : [0, 1] \rightarrow [0, 1] \) defined by

\[
\psi_X(s) \equiv \mathbb{E}[s^X] = \sum_{k=0}^{\infty} p(k) \cdot s^k.
\]

Example 4.12. If \( X \) is a Bernoulli random variable with parameter \( p \), then the PGF of \( X \) is the function

\[
\psi_X(s) = 1 - p + ps.
\]

Probability generating functions are a special case of a much more general idea. Given any sequence of numbers \( c_0, c_1, \cdots \), we can define a function

\[
\Psi(t) \equiv \sum_{n=0}^{\infty} c_n t^n,
\]

which we call the generating function of the sequence. This is useful because we can sometimes use analytical properties of the function \( \Psi \) to learn about the original sequence.

Several noteworthy properties of probability generating functions are listed below:

\begin{itemize}
  \item \( \psi_X(1) = \sum_{k=0}^{\infty} p(k) = 1 \).
  \item \( \psi_X(0) = p(0) = \mathbb{P}(X = 0) \).
  \item We can recover the probability mass function of \( X \) by repeatedly differentiating its PGF and evaluating these derivatives at \( s = 0 \):
    \[
p(k) = \frac{1}{k!} \cdot \psi_X^{(k)}(0),
    \]
\end{itemize}
where $\psi_X^{(k)}(s)$ is the $k$'th derivative of the function $\psi_X(s)$ with respect to the variable $s$. This shows that the distribution of the variable is uniquely determined by its PGF.

- We can also use the PGF to calculate the mean and the variance of $X$. By differentiating $\psi_X(s)$ with respect to $s$ and then evaluating at $s = 1$, we obtain

$$\psi_X'(1) = \sum_{k=0}^{\infty} p(k)k = \mathbb{E}[X]$$

$$\psi_X''(1) = \sum_{k=0}^{\infty} p(k)k(k - 1) = \mathbb{E}[X^2] - \mathbb{E}[X],$$

from which it follows that

$$\mathbb{E}[X] = \psi_X'(1)$$

$$\text{Var}(X) = \psi_X''(1) + \mathbb{E}[X] - \mathbb{E}[X]^2.$$  \hspace{1cm} (4.1)  \hspace{1cm} (4.2)

We will use Equations (4.1) and (4.2) in the next couple sections to calculate the mean and the variance of several important classes of discrete distributions.

### 4.6 The Binomial Distribution

**Definition 4.7.** A random variable $X$ is said to have the **binomial distribution with parameters** $n$ and $p$ if it takes values in the set $\{0, 1, \ldots, n\}$ with probability mass function

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

The name of this distribution stems from the connection between the probabilities $p(k)$ and the coefficients in a binomial expansion:

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

Notice that a Bernoulli random variable with parameter $p$ (cf. Example 4.3) is also a binomial random variable with parameters $n = 1$ and $p$. In fact, the next theorem shows that there is a close relationship between Bernoulli distributions and binomial distributions even when $n \geq 1$.

**Theorem 4.1.** Let $X_1, \cdots, X_n$ be independent Bernoulli random variables, each with the same parameter $p$. Then the sum $X = X_1 + \cdots + X_n$ is a binomial random variable with parameters $n$ and $p$.

**Proof.** The random variable $X$ counts the number of Bernoulli variables $X_1, \cdots, X_n$ that are equal to 1, i.e., the number of successes in the $n$ independent trials. Clearly $X$ takes values in the set $\{0, \cdots, n\}$. To calculate the probability that $X = k$, let $E$ be the event that $X_{i_1} = X_{i_2} = \cdots = X_{i_k} = 1$ and $X_j = 0$ for all $j \notin \{i_1, \cdots, i_k\}$. Then, because the Bernoulli variables are independent, we know that

$$\mathbb{P}(E) = p^k(1-p)^{n-k}.$$
However, there are \( \binom{n}{k} \) such sets of indices \( i_1, \cdots, i_k \), these corresponding to mutually exclusive events, and so
\[
\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.
\]

**Example 4.13.** Suppose that a fair coin is tossed \( n \) times. Then the number of heads that appear is a binomial random variable with parameters \( n \) and \( p = 1/2 \).

**Example 4.14.** (Ross, Example 6d) Suppose that eye color in humans is under the control of a single gene. Individuals with genotypes BB or Bb have brown eyes, while individuals with genotype bb have blue eyes. Suppose that two brown-eyed individuals with genotypes Bb and Bb have a total of 4 children. Assuming Mendelian inheritance, what is the distribution of the number of blue-eyed offspring?

**Answer:** Let \( X_i \) equal 1 if the \( i \)'th child has blue eyes and 0 otherwise. Then \( X_i \) is a Bernoulli random variable with parameter \( p = 1/4 \), since a child will be blue-eyed only if they inherit a blue allele from each parent, which happens with probability \( 1/4 \). Since the genotypes of the different offspring are independent, the total number of blue-eyed children is binomial with parameters \( n = 4 \) and \( p = 1/4 \).

The mean and the variance of a binomial random variable \( X \) with parameters \( n \) and \( p \) can be calculated with the help of the probability generating function, which is
\[
\psi_X(s) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} s^k = (1 - ps + (1 - p)s)^n.
\]

Then, by using equations (4.1) and (4.2), we obtain
\[
\mathbb{E}[X] = np(1 - p + p \cdot 1)^{n-1} = np.
\]
\[
\text{Var}(X) = n(n-1)p^2 + np - n^2p^2 = np(1 - p).
\]

In particular, if \( n = 1 \), then \( X \) is a Bernoulli random variable and these expressions are just \( p \) and \( p(1 - p) \), in agreement with the mean and the variance calculated in Examples 4.6 and 4.10. Later we will see that the mean and the variance of the binomial distribution can also be easily calculated from the results for the Bernoulli distribution with the help of Theorem 4.1

The final result of this section asserts that the probability mass function of a binomial random variable is always peaked around its mean.

**Proposition 4.2.** Let \( X \) be a binomial random variable with parameters \( (n, p) \). Then \( \mathbb{P}(X = k) \) is a unimodal function of \( k \) with its maximum at the largest integral value of \( k \) less than or equal to \((n + 1)p\).

**Proof.** A simple calculation shows that
\[
\frac{\mathbb{P}(X = k)}{\mathbb{P}(X = k - 1)} = \frac{(n - k + 1)p}{k(1 - p)},
\]
which is greater than or equal to 1 if and only if
\[
k \leq (n + 1)p.
\]
4.7 The Poisson Distribution

Definition 4.8. A random variable $X$ is said to have the **Poisson distribution with parameter** $\lambda \geq 0$ if it takes values in the non-negative integers with probability mass function

$$ p(n) = \mathbb{P}(X = n) = e^{-\lambda} \frac{\lambda^n}{n!}. $$

That this formula defines a probability distribution follows from the calculation

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

Similarly, the probability generating function of $X$ is

$$ \phi_X(s) = \mathbb{E}[s^X] = \sum_{n \geq 0} e^{-\lambda} \frac{\lambda^n}{n!} s^n = e^{-\lambda} e^{\lambda s} = e^{-\lambda(1-s)}, $$

while the mean and the variance are both equal to the parameter $\lambda$:

$$ \mathbb{E}[X] = \phi'_X(1) = \lambda $$

$$ \text{Var}(X) = \phi''_X(1) + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. $$

It has long been known that the Poisson distribution provides a surprisingly accurate description of the statistics of a very large number of seeming unrelated phenomena. Some examples include:

- the number of misprints per page of a book;
- the number of wrong telephone numbers dialed in a day;
- the number of customers entering a post office on a given day;
- the number of vacancies (per year) on the US Supreme Court (Table 4.1);
- the number of $\alpha$-particles discharged per day from a C-14 source;
- the number of major earthquakes per year.

Table 4.1: Data from Cole (2010) compared with a Poisson distribution with $\lambda = 0.5$.

<table>
<thead>
<tr>
<th>Number of vacancies (x)</th>
<th>Probability</th>
<th>Years with x vacancies</th>
<th>1837-1932</th>
<th>1933-2007</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Observed</td>
<td>Expected</td>
<td>Observed</td>
</tr>
<tr>
<td>0</td>
<td>0.6065</td>
<td>59</td>
<td>58.2</td>
<td>47</td>
</tr>
<tr>
<td>1</td>
<td>0.3033</td>
<td>27</td>
<td>29.1</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>0.0758</td>
<td>9</td>
<td>7.3</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>0.0126</td>
<td>1</td>
<td>1.2</td>
<td>0</td>
</tr>
<tr>
<td>$\geq 3$</td>
<td>0.0018</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
</tr>
</tbody>
</table>

Since these phenomena are generated by very different physical processes, the fact that they have similar statistical properties must be due to a shared mathematical structure rather than the details of the underlying mechanisms. Some insight into what this structure might be is provided by the following theorem.
**Theorem 4.2.** For each \( n \geq 1 \), let \( X_n \) be a Binomial random variable with parameters \( n \) and \( p_n = \lambda/n \). Then, for any \( k \geq 0 \),

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

**Proof.** Begin by writing the binomial probability as:

\[
P(X_n = k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k}
\]

\[
= \frac{n!}{(n-k)!k!} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k}
\]

\[
= \frac{n!}{n^k (n-k)!} \left( \frac{\lambda^k}{k!} \right) \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k}
\]

\[
= \frac{n(n-1)(n-2) \cdots (n-k+1)}{n^k} \left( \frac{\lambda^k}{k!} \right) \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k}.
\]

The theorem is then a consequence of the following three limits:

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

\[
\lim_{n \to \infty} \left( 1 - \frac{\lambda}{n} \right)^{-k} = 1
\]

\[
\lim_{n \to \infty} \left( \frac{n(n-1)(n-2) \cdots (n-k+1)}{n^k} \right) = 1.
\]

**Interpretation:** The theorem asserts that if we consider a large number \( n \) of independent trials, each of which has a small success probability \( \lambda/n \), then the total number of events observed is approximately Poisson distributed with parameter \( \lambda = np \). This is known as the law of rare events. Of course, the assumption that each trial has the same probability of success is unlikely to be satisfied in most settings, but this is dispensed with in the following more general theorem which we state without proof.

**Theorem 4.3.** For each \( n \geq 1 \), let \( X_{n,1}, \ldots, X_{n,n} \) be a family of independent Bernoulli random variables with parameters \( p_{n,1}, \ldots, p_{n,n} \). Suppose that

\[
\lim_{n \to \infty} \max_{1 \leq i \leq n} p_{n,i} = 0
\]

\[
\lim_{n \to \infty} \sum_{i=1}^{n} p_{n,i} = \lambda.
\]

Then,

\[
\lim_{n \to \infty} P(X_{n,1} + \cdots + X_{n,n} = k) = e^{-\lambda} \frac{\lambda^k}{k!}.
\]

**Remark:** The first condition guarantees that each event is individually unlikely to occur, while the second condition implies that the mean number of events that do occur is approximately \( \lambda \).
when \( n \) is large.

Poisson approximation can even be accurate when the trials are not independent, provided the dependence is sufficiently weak. This is illustrated in the following example.

**Example 4.15.** In Example 3.4, we showed that if a deck of \( n \) cards is shuffled at random (i.e., each of the \( n! \) possible permutations is equally likely), then the probability that exactly \( k \) cards are left in their original position is:

\[
\mathbb{P}\{\text{exactly } k \text{ cards are left in position}\} = \frac{1}{k!} \sum_{i=0}^{n} (-1)^i \frac{1}{i!} \approx e^{-1} \frac{1}{k!}.
\]

In other words, if \( X_n \) denotes the number of cards left in position, then for large \( n \), \( X_n \) is approximately Poisson-distributed with parameter \( \lambda = 1 \). To understand why, let \( E_i \) denote the event that \( i \)’th card is left in position, let \( Y_{n,i} \) be a Bernoulli random variable which is equal to 1 when \( E_i \) is true (i.e., \( Y_{n,i} \) is the indicator function of \( E_i \)), and notice that \( X_n = Y_{n,1} + \cdots + Y_{n,n} \).

Also, observe that for all \( i \neq j \),

\[
\mathbb{P}(Y_{n,i} = 1) = \mathbb{P}(E_i) = \frac{1}{n},
\]

\[
\mathbb{P}(E_i|E_j) = \frac{\mathbb{P}(E_i \cap E_j)}{\mathbb{P}(E_j)} = \frac{1}{n-1} \approx \frac{1}{n}.
\]

Thus, when \( n \) is large, the probability of each event is small (i.e., these are rare events), while the pairwise dependence of the events is also small, since the conditional probability of \( E_i \) given \( E_j \) is nearly equal to the probability of \( E_i \).

### 4.7.1 Poisson Processes

Poisson distributions can also be used to model scenarios in which events occur at a ‘constant rate’ through space or time. Consider a sequence of events that occur repeatedly through time (such as earthquakes or mutations to DNA) and assume that the following properties are satisfied:

- **Property 1:** The probability that exactly one event occurs in a given interval \([t, t+h]\) is equal to \( \lambda h + o(h) \), where \( o(h) \) stands for a function \( f(h) \) for which \( \lim_{h \to 0} f(h)/h = 0 \).

- **Property 2:** The probability that two or more events occur in an interval of length \( h \) is \( o(h) \).

- **Property 3:** For any set of non-overlapping intervals \([s_1, t_1], [s_2, t_2], \ldots, [s_r, t_r]\), the numbers of events occurring in these intervals are independent.

Let \( N(t) \) denote the number of events that occur in the interval \([0, t]\). Our goal is to calculate the distribution of \( N(t) \), which we can do by breaking the interval \([0, t]\) into \( n \) non-overlapping subintervals, each of length \( t/n \). Then,

\[
\mathbb{P}(N(t) = k) = \mathbb{P}(A_n) + \mathbb{P}(B_n)
\]

where

\[
A_n = \{k \text{ subintervals each contain exactly one event, while the others contain no events}\}
\]

\[
B_n = \{N(t) = k \text{ and at least one subinterval contains } \geq 2 \text{ events}\}.
\]
(Notice that $A_n$ and $B_n$ are mutually exclusive.)

We first show that $\mathbb{P}(B_n)$ can be made arbitrarily small by taking $n$ sufficiently large. Indeed,

$$\mathbb{P}(B_n) \leq \mathbb{P}\{\text{at least one subinterval contains two or more events}\}$$

$$\leq \sum_{k=1}^{n} \mathbb{P}\{\text{the } k\text{'th subinterval contains two or more events}\}$$

$$\leq \sum_{k=1}^{n} o\left(\frac{t}{n}\right) = n \cdot o\left(\frac{t}{n}\right) = t \left[ o\left(\frac{t/n}{t/n}\right) \right].$$

However, since $t/n \to 0$ as $n \to \infty$, it follows from the definition of $o(h)$ that this last expression vanishes as $n \to \infty$, which then implies that

$$\lim_{n \to \infty} \mathbb{P}(B_n) = 0.$$

Next, using Properties 1 and 2, observe that

$$\mathbb{P}\{\text{an interval of length } t/n \text{ contains one event}\} = \frac{\lambda t}{n} + o\left(\frac{t}{n}\right)$$

$$\mathbb{P}\{\text{an interval of length } t/n \text{ contains no events}\} = 1 - \frac{\lambda t}{n} + o\left(\frac{t}{n}\right).$$

Also, using Property 3, we have

$$\mathbb{P}(A_n) = \left(\begin{array}{c} n \\ k \end{array}\right) \left[ \frac{\lambda t}{n} + o\left(\frac{t}{n}\right) \right]^k \left[ 1 - \frac{\lambda t}{n} + o\left(\frac{t}{n}\right) \right]^{n-k}$$

$$\to e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad \text{as } n \to \infty,$$

since there are $\left(\begin{array}{c} n \\ k \end{array}\right)$ ways for the $k$ events to be distributed across the $n$ subintervals such that each subinterval contains either 0 or 1 event.

Putting these results together shows that

$$\mathbb{P}(N(t) = k) = \lim_{n \to \infty} (\mathbb{P}(A_n) + \mathbb{P}(B_n))$$

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

In other words, for each $t \geq 0$, the number of events $N(t)$ to have occurred by time $t$ is a Poisson random variable with parameter $\lambda t$. The collection of random variables $(N(t) : t \geq 0)$ is said to be a **Poisson process with intensity** $\lambda$. Such processes are used to model a wide variety of phenomena such as disease transmissions, earthquakes, and arrival of traffic at stoplights.

### 4.8 Linearity of Expectations

Later in the course we will prove the Law of Large Numbers, which asserts that the average of a large collection of independent, identically-distributed random variables tends to the expected value of the distribution. A related but simpler result is described in the following proposition.
Proposition 4.3. Suppose that \((S, \mathcal{F}, P)\) is a probability space with a countable sample space \(S\) and let \(X : S \rightarrow \mathbb{R}\) be a random variable. Then

\[
\mathbb{E}[X] = \sum_{s \in S} P(s) \cdot X(s).
\]

Proof. Since \(S\) is countable, it is clear that \(X\) can take on only countably many values, which we label \(x_i, i \geq 1\), and so \(X\) is a discrete random variable. Furthermore, if we let \(S_i = \{s : X(s) = x_i\} = X^{-1}(x_i)\), then \(S\) is a disjoint union of the sets \(S_1, S_2, \cdots\) and

\[
P(S_i) = \sum_{s \in S_i} P(s).
\]

Consequently,

\[
\mathbb{E}[X] = \sum_i P(X = x_i) \cdot x_i = \sum_i P(S_i) \cdot x_i = \sum_i \left( \sum_{s \in S_i} P(s) \right) \cdot x_i
\]

\[
= \sum_i \sum_{s \in S_i} P(s) \cdot x_i
\]

\[
= \sum_i \sum_{s \in S_i} P(s) \cdot X(s)
\]

\[
= \sum_{s \in S} P(s) \cdot X(s).
\]

We can use this proposition to show that expectation is a linear functional: the expected value of a sum of random variables is equal to the sum of their expectations. Although we will only prove this for discrete random variables, the result is true in general. Also, it should be noted that the result holds even when the random variables inside the expectation are not independent.

Theorem 4.4. For random variables \(X_1, X_2, \cdots, X_n\),

\[
\mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{E} [X_i].
\]

Proof. Assume that \(X_1, \cdots, X_n\) are discrete random variables, all defined on the same sample space \(S\) with probability distribution \(P\), and let \(Z = \sum_{i=1}^{n} X_i\). By Proposition 4.3,

\[
\mathbb{E}[Z] = \sum_{s \in S} P(s) Z(s)
\]

\[
= \sum_{s \in S} P(s) \left( X_1(s) + \cdots + X_n(s) \right)
\]

\[
= \sum_{s \in S} P(s) X_1(s) + \cdots + \sum_{s \in S} P(s) X_n(s)
\]

\[
= \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n].
\]
4.9 The Geometric Distribution

Definition 4.9. A random variable \(X\) is said to have the geometric distribution with parameter \(p \in (0, 1]\) if \(X\) takes values in the positive integers with probability mass function

\[
p(n) = \mathbb{P}(X = n) = (1 - p)^{n-1}p, \quad n \geq 1.
\]

Remark 4.3. Some authors define the geometric distribution differently, by allowing \(X\) to take values in the set of non-negative integers with probability mass function

\[
p(n) = \mathbb{P}(X = n) = (1 - p)^n p, \quad n \geq 0.
\]

Sometimes this is clear from context, but you should always take care to indicate which convention is being followed.

Geometric random variables arise in problems where an experiment is performed repeatedly until a particular outcome is observed. For example, suppose that \(Y_n, n \geq 1\) is a sequence of independent Bernoulli trials, each with probability of success \(p\), and let \(X\) be the number of the first trial at which a success is recorded. Then

\[
\mathbb{P}(X = n) = \mathbb{P}(Y_1 = 0, \ldots, Y_{n-1} = 0, Y_n = 1) = (1 - p)^{n-1}p,
\]

and so it follows that \(X\) is a geometric random variable with parameter \(p\). (If, instead, we take \(X\) to be the number of failures until the first success is observed, then \(X\) is geometrically-distributed according to the alternative convention described in the preceding remark.)

Example 4.16. (Ross, Example 8a): An urn contains \(N\) white and \(M\) black balls. If balls are sampled with replacement until the first black one is obtained, then the number of balls sampled is geometrically distributed with parameter \(p = M/(N + M)\).

Properties: Let \(X\) have the geometric distribution with parameter \(p\) and set \(q = 1 - p\). Then the PGF of \(X\) is

\[
\phi_X(s) = \mathbb{E}[s^X] = \sum_{n=1}^{\infty} pq^{n-1}s^n = ps \sum_{n=0}^{\infty} q^n s^n = \frac{ps}{1 - qs},
\]

while the mean and variance are given by

\[
\mathbb{E}[X] = \phi_X'(1) = \frac{p}{(1 - q)^2} = \frac{1}{p},
\]

\[
\text{Var}(X) = \phi_X''(1) + \mathbb{E}[X] - \mathbb{E}[X]^2 = \frac{2q}{p^2} - \frac{1}{p} + \frac{1}{p^2} = \frac{q}{p^2}.
\]

As expected, the mean number of trials needed to observe the first success increases as the probability of success, \(p\), decreases. Notice that the variance also increases with decreasing \(p\). Per the remark, the mean and the variance depend on whether the geometric distribution is defined on just the positive integers or on all of the non-negative integers.
4.10 The Negative Binomial Distribution

Definition 4.10. An integer-valued random variable $X$ is said to have the negative binomial distribution with parameters $r$ and $p$ if the probability mass function is

$$p(n) = \Pr(X = n) = \begin{cases} 0 & \text{if } n < r \\ \binom{n-1}{r-1} p^r (1-p)^{n-r} & \text{if } n \geq r. \end{cases}$$

The negative binomial distribution is a generalization of the geometric distribution in which an experiment is performed repeatedly until $r$ successes have been observed. As above, let $Y_n, n \geq 1$ be a sequence of independent Bernoulli trials, each with probability of success $p$, but now let $X$ be the number of the trial at which the $r$'th success is obtained. By definition, $X \geq r$ and $X = n$ if and only if there are $r - 1$ successes and $n - r$ failures among the first $n - 1$ trials and the $n$'th trial is a success. Notice that there are $\binom{n-1}{r-1}$ different outcomes that have $r - 1$ successes among the first $n - 1$ trials as well as a success on the $n$'th trials, and that the probability of any one such outcome is simply $p^r (1-p)^{n-r}$. Since the different outcomes are mutually exclusive, it follows that

$$\Pr(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r},$$

for $n \geq r$, and so $X$ is a negative binomial random variable with parameters $(r, p)$.

Example 4.17. (Ross, Example 8d): If independent trials, each having success probability $p$ are performed, then the probability that $r$ successes occur before $m$ failures is

$$\sum_{n=r}^{r+m-1} \binom{n-1}{r-1} p^r (1-p)^{n-r}.$$ 

A decomposition: In the above formulation, let $X_1$ denote the number of trials until the first success is observed, let $X_2$ denote the number of additional trials occurring that occur after the first success until the second success is observed, and (in general), let $X_i$ denote the number of trials taking place after the $-1$'st success has been observed and until the $i$'th success occurs. Clearly, $X_1$ is a geometric random variable with parameter $p$. In addition, because all of the trials are independent, $X_2, \cdots, X_r$ are also geometric random variables with parameter $p$, and all of the random variables $X_1, \cdots, X_r$ are independent. Since $X = X_1 + \cdots + X_r$, we see that the distribution of the sum of $r$ independent geometric random variables with parameter $p$ is the negative binomial distribution with parameter $(r, p)$.

Remark 4.4. This result only holds when all of the geometric random variables have the same parameter $p$.

Properties: Let $X$ have the negative binomial distribution with parameters $(r, p)$. Later in the course we will use the decomposition of $X$ into a sum of independent geometric random variables to show that the PGF of $X$ is:

$$\phi_X(x) = \mathbb{E}[s^X] = \left( \frac{ps}{1-q s} \right)^r.$$
A simple calculation then shows that

\[ E[X] = \phi_X'(1) = \frac{r}{p} \]

\[ \text{Var}(X) = \phi_X''(1) + E[X] - E[X]^2 = \frac{rq}{p^2}, \]

i.e., the mean and variance of \( X \) are equal to \( r \) times the mean and variance of a geometric random variable with parameter \( p \). Both of these results can also be deduced from the relationship between the negative binomial distribution and the geometric distribution.

### 4.11 The Hypergeometric Distribution

**Definition 4.11.** A random variable \( X \) is said to have the **hypergeometric distribution with parameters** \( (N, m, n) \) if \( X \) has values in the set \( \{0, 1, \ldots, n\} \) with probability mass function

\[ p(i) = \mathbb{P}(X = i) = \binom{m}{i} \binom{N-m}{n-i} / \binom{N}{n}. \]

That this is a legitimate distribution follows from the following combinatorial identity (cf. Exercise (1.8TE)),

\[ \sum_{i=0}^{n} \binom{m}{i} \binom{N-m}{n-i} = \binom{N}{n}, \]

which in turn implies that

\[ \sum_{i=0}^{n} p(i) = 1. \]

The hypergeometric distribution arises in problems involving sampling without replacement. Suppose that an urn contains \( m \) red balls and \( N - m \) blue balls (so, \( N \) balls in total), and that we choose \( n \) balls at random without replacement. Then, if \( X \) denotes the number of red balls in our sample, a simple combinatorial argument shows that

\[ \mathbb{P}(X = i) = \binom{m}{i} \binom{N-m}{n-i} / \binom{N}{n}, \]

and so \( X \) has the hypergeometric distribution with parameters \( (N, m, n) \).

**Remark 4.5.** Since

\[ \lim_{N \to \infty} \frac{\binom{[Np]}{i} \binom{[N(1-p)]}{n-i}}{\binom{N}{n}} = \binom{n}{i} p^i (1-p)^{n-i}, \]

we see that when \( N \) and \( m \) are both large in comparison to \( n \), the distribution of a hypergeometric random variable with parameters \( (N, m, n) \) is approximately the same as the distribution of a binomial random variable with parameters \( n \) and \( p \), where \( p = m/N \). This is unsurprising, since sampling \( n \) individuals without replacement from a very large population is ‘nearly’ the same as sampling \( n \) individuals from that population with replacement.
4.11. THE HYPERGEOMETRIC DISTRIBUTION

Properties: In general, the PGF of a hypergeometric random variable does not have a ‘nice’ closed-form expression. (It can be written as a hypergeometric function, hence the name of the distribution.) However, the mean and the variance of the hypergeometric distribution can be calculated using simple combinatorial manipulation. Let $X$ be a $HG(N,m,n)$-distributed random variable. Using the identities

$$i \binom{m}{i} = m \binom{m-1}{i-1} \quad \text{and} \quad n \binom{N}{n} = N \binom{N-1}{n-1},$$

we have

$$\mathbb{E}[X^k] = \sum_{i=1}^{n} i^k \binom{m}{i} \frac{(N-m)}{(n-i)}$$

$$= \frac{nm}{N} \sum_{i=1}^{n} i^{k-1} \binom{m-1}{i-1} \frac{(N-m)}{(n-i)}$$

$$= \frac{nm}{N} \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{m-1}{j} \frac{(N-m)}{(n-1-j)}$$

$$= \frac{nm}{N} \mathbb{E}[(Y+1)^{k-1}],$$

where $Y$ is a hypergeometric RV with parameters $(N-1, m-1, n-1)$. Taking $k = 1$ and defining $p = m/N$ gives

$$\mathbb{E}[X] = np \cdot \mathbb{E}[(Y+1)^0] = np.$$

Also, taking $k = 2$, we have

$$\mathbb{E}[X^2] = np \mathbb{E}[Y + 1] = np \left[ \frac{(n-1)(m-1)}{N-1} + 1 \right],$$

and so

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np(1 - p) \left( 1 - \frac{n-1}{N-1} \right).$$

These quantities can also be derived more directly using techniques to be introduced in the next lecture.
Chapter 5

Continuous Random Variables

5.1 Cumulative Distribution Functions

Recall that if $X$ is a real-valued random variable, then the cumulative distribution function (CDF) of $X$ is the function $F : \mathbb{R} \rightarrow [0, 1]$ defined by

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

**Proposition 5.1.** Let $X$ be a real-valued random variable (not necessarily discrete) with cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$. Then

1. $F$ is non-decreasing, i.e., if $x < y$, then $F(x) \leq F(y)$.
2. $\lim_{x \to \infty} F(x) = 1$.
3. $\lim_{x \to -\infty} F(x) = 0$.
4. $F$ is right-continuous, i.e., for any $x$ and any decreasing sequence $(x_n, n \geq 1)$, that converges to $x$, $\lim_{n \to \infty} F(x_n) = F(x)$.

**Proof.** 1) If $x < y$, then $\{X \leq x\} \subset \{X \leq y\}$, which implies that $F(x) \leq F(y)$.

2) If $(x_n, n \geq 1)$ is an increasing sequence such that $x_n \to \infty$, then the events $E_n = \{X \leq x_n\}$ form an increasing sequence with

$$\{X < \infty\} = \bigcup_{n=1}^{\infty} E_n.$$

It follows from the continuity properties of probability measures (Lecture 2) that

$$\lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} \mathbb{P}(E_n) = \mathbb{P}(X < \infty) = 1.$$

3) Likewise, if $(x_n, n \geq 1)$ is a decreasing sequence such that $x_n \to -\infty$, then the events $E_n = \{X \leq x_n\}$ form a decreasing sequence with

$$\emptyset = \bigcap_{n=1}^{\infty} E_n.$$

In this case, the continuity properties of measures imply that

$$\lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} \mathbb{P}(E_n) = \mathbb{P}(\emptyset) = 0.$$
4) If \( x_n, n \geq 1 \) is a decreasing sequence converging to \( x \), then the sets \( E_n = \{ X \leq x_n \} \) also form a decreasing sequence with
\[
\{ X \leq x \} = \bigcap_{n=1}^{\infty} E_n.
\]
Consequently,
\[
\lim_{n \to \infty} F(x_n) = \lim_{n \to \infty} P(E_n) = P\{ X \leq x \} = F(x).
\]

Some other notable properties of the CDF are:

(i) \( P(a < X \leq b) = F(b) - F(a) \) for all \( a < b \).

(ii) Another application of the continuity property shows that the probability that \( X \) is strictly less than \( x \) is equal to
\[
P(X < x) = \lim_{n \to \infty} P(X \leq x - 1/n) = \lim_{n \to \infty} F(x - 1/n).
\]
In other words, \( P(X < x) \) is equal to the left limit of \( F \) at \( x \), which is often denoted \( F(x-) \). However, notice that in general this limit need not be equal to \( F(x) \):
\[
F(x) = P(X < x) + P(X = x),
\]
and so \( P(X < x) = F(x) \) if and only if \( P(X = x) = 0 \).

(iii) The converse of this proposition is also true: namely, if \( F : \mathbb{R} \to [0,1] \) is a function satisfying properties (1) - (4), then \( F \) is the CDF of some random variable \( X \). We will prove this later on in the course.

5.2 Continuous Random Variables

Definition 5.1. A real-valued random variable \( X \) is said to be a continuous random variable if there is a non-negative function \( p : \mathbb{R} \to [0,\infty) \) such that
\[
P(a \leq X \leq b) = \int_{a}^{b} p(x)dx
\]
for all \( a < b \). The function \( p \) is called the probability density function (pdf) of \( X \).

Because the density of a continuous random variable is not a probability, it can exceed 1 as we shall see in several examples. However, since
\[
1 = P(-\infty < X < \infty) = \int_{-\infty}^{\infty} p(x)dx,
\]
it follows that the integral of a density function over the entire real line must equal 1. As in the discrete setting, this identity can be used to calculate normalizing constants for density functions.
Furthermore, since
\[ P(X = x) = \int_{x}^{\infty} p(y)dy = 0, \]
the probability that a continuous RV will assume any particular value is 0.

There is a close connection between the density of a random variable \( X \) and its CDF. First, observe that
\[ F(x) = P(X \leq x) = \int_{-\infty}^{x} p(y)dy. \]
Since the right-hand side is a differentiable function of \( x \), it follows that \( F(x) \) is also differentiable and that
\[ F'(x) = p(x), \]
i.e., the density is the derivative of the cumulative distribution function. Conversely, if \( F(x) \) is known to be differentiable, then by the fundamental theorem of calculus, \( F(x) \) is the anti-derivative of its derivative and thus \( F'(x) \) is the density of \( X \). Notice that if \( F \) is differentiable, then \( F \) is necessarily continuous, i.e., the cumulative distribution function of a continuous random variable is necessarily continuous. However, the converse is not true: there are random variables with continuous cumulative distribution functions that are not continuous variables. An example is given later in this chapter.

**Example 5.1.** If \( X \) is a continuous RV with CDF \( F_X \) and density \( f_X \), find the CDF and the density function of \( Y = 2X \).

**Solution:** The CDF of \( Y \) is equal to
\[ F_Y(x) = P(Y \leq x) = P(2X \leq x) = P(X \leq x/2) = F_X(x/2), \]
and differentiation with respect to \( x \) gives
\[ f_Y(x) = \frac{d}{dx} F_Y(x) = \frac{1}{2} f_X(x/2). \]

### 5.3 Expectation

**Definition 5.2.** Let \( X \) be a continuous RV with density function \( p(x) \). Then the expected value of \( X \) (also called the mean or expectation of \( X \)) is defined as
\[ E[X] = \int_{-\infty}^{\infty} xp(x)dx, \]
provided that the integral on the right-hand side exists.

Thus, as with discrete random variables, the expected value of a continuous random variable can be thought of as a weighted average of the values that the variable can assume, where now the weights are provided by the density of the variable.

**Remark 5.1.** The expected value of a random variable \( X \) may be equal to \( \infty \) or \( -\infty \). For example, suppose that \( X \) has the following density
\[ p(x) = \begin{cases} \frac{1}{x^2} & \text{if } x > 1 \\ 0 & \text{otherwise.} \end{cases} \]
Notice that \( \int_{-\infty}^{\infty} p(x)dx = 1 \), so \( p(x) \) is a legitimate probability density. Nonetheless, the expected value of \( X \) is \( \infty \) since

\[
E[X] = \int_{1}^{\infty} xp(x)dx = \int_{1}^{\infty} \frac{dx}{x} = \infty.
\]

**Expectations of functions:** As in the discrete case, the expectation of a real-valued function of a random variable can be evaluated directly in terms of the distribution of the random variable (cf. the Law of the Unconscious Statistician). Although this result is true in general, we will only prove it under the assumption that \( f(X) \) is non-negative. Our proof will rely on the following lemma.

**Lemma 5.1.** If \( Y \) is a nonnegative random variable with density \( p(y) \), then

\[
E[Y] = \int_{0}^{\infty} P(Y > y)dy.
\]

**Proof.** Using the expression \( P\{Y > y\} = \int_{y}^{\infty} p(x)dx \), we can rewrite the right-hand side as

\[
\int_{0}^{\infty} P(Y > y)dy = \int_{0}^{\infty} \left( \int_{y}^{\infty} p(x)dx \right) dy
= \int_{0}^{\infty} \left( \int_{0}^{\infty} dy \right) p(x)dx
= \int_{0}^{\infty} xp(x)dx = E[Y].
\]

Notice that in passing from the first to the second line, we have interchanged the order of integration. (Draw a picture if this calculation is not clear.) This interchange is justified by the fact that the integrand is non-negative.

**Proposition 5.2.** If \( X \) is a continuous random variable with density \( p(x) \) and if \( f : \mathbb{R} \to \mathbb{R} \), then

\[
E[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx.
\]

**Proof.** Assume that \( f \) is non-negative. By applying the preceding lemma to the random variable \( Y = f(X) \), we have

\[
E[f(X)] = \int_{0}^{\infty} P\{f(X) > y\}dy
= \int_{0}^{\infty} \left( \int_{x:f(x)>y} p(x)dx \right) dy
= \int_{-\infty}^{\infty} \left( \int_{0}^{f(x)} dy \right) p(x)dx
= \int_{-\infty}^{\infty} f(x)p(x)dx,
\]

where the last equality follows from the fact that \( f(x) \geq 0 \) for all \( x \in \mathbb{R} \).
Example 5.2. As in the discrete setting, the expected value of an affine function of a continuous random variable is equal to that affine function evaluated at the expected value of the random variable, i.e., if $X$ has pdf $p(x)$ and $a, b \in \mathbb{R}$ are real numbers, then
\[
E[aX + b] = \int_{-\infty}^{\infty} (ax + b)p(x)dx = aE[X] + b.
\]

Example 5.3. If $n$ is a non-negative integer, then the $n$'th moment of the continuous random variable $X$ is the expected value of $X^n$, which can be calculated as:
\[
E[X^n] = \int_{-\infty}^{\infty} x^n p(x)dx.
\]

Definition 5.3. Let $X$ be a continuous RV with density function $p(x)$. Then the variance of $X$ is
\[
\]

Remark 5.2. The equality of the first and second lines of Definition 5.3 can be established by the same calculation that we used for discrete random variables.

5.4 The Uniform Distribution

Definition 5.4. A continuous random variable $X$ is said to have the uniform distribution on the interval $(a,b)$ (denoted $U(a,b)$) if the density function of $X$ is
\[
p(x) = \begin{cases} 
\frac{1}{b-a} & \text{if } x \in (a,b) \\
0 & \text{otherwise}
\end{cases}
\]

Interpretation: To say that $X$ is uniformly distributed on an interval $(a,b)$ means that each point in $(a,b)$ is equally likely to be a value of $X$. In particular, the probability that $X$ lies in any subinterval of $(a,b)$ is proportional to the length of the subinterval: if $(l,r) \subseteq (a,b)$, then
\[
P(X \in (l,r)) = \frac{r-l}{b-a}.
\]

Moments: Suppose that $X$ has the $U(a,b)$ distribution. Then the $n$'th moment of $X$ is given by
\[
E[X^n] = \frac{1}{b-a} \int_{a}^{b} x^n dx = \frac{1}{b-a} \frac{1}{n+1} x^{n+1}\Big|_{a}^{b} = \frac{1}{n+1} \left( \frac{b^{n+1} - a^{n+1}}{b-a} \right).
\]

It follows that the mean and the variance of $X$ are
\[
E[X] = \frac{1}{2}(a+b)
\]
\[
Var(X) = E[X^2] - E[X]^2 = \frac{1}{12}(b-a)^2.
\]

The case $U(0,1)$ is of particular importance. If $X$ is uniform on $(0,1)$, then
\[
E[X] = \frac{1}{2} \quad \text{and} \quad Var(X) = \frac{1}{12}.
\]
Remark: Notice that there is no random variable that is uniformly distributed over an infinite interval such as \((0, \infty)\) or \((-\infty, \infty)\). This follows from the fact that the integral of a constant density over such a region is necessarily infinite.

Construction: Suppose that \(X_1, X_2, \cdots\) is a sequence of independent Bernoulli random variables, each with parameter \(1/2\). If we define
\[
X = \sum_{n=1}^{\infty} X_n \left( \frac{1}{2} \right)^n,
\]
then \(X\) is a standard uniform random variable. Notice that the random sequence \((X_1, X_2, \cdots)\) is the binary representation of the random number \(X\). That \(X\) is uniformly distributed can be deduced as follows:

**Step 1:** We say that a number \(x \in \mathbb{R}\) is a dyadic rational number if \(x\) has a finite binary expansion, i.e., there exists an integer \(n \geq 1\) and numbers \(c_1, \cdots, c_n \in \{0, 1\}\) such that
\[
x = \sum_{k=1}^{\infty} c_k \left( \frac{1}{2} \right)^k.
\]
Notice that every dyadic rational number is rational, but that not every rational number is dyadic rational, e.g., \(x = 1/3\) is not dyadic rational. On the other hand, the set of dyadic rational numbers is dense in \(\mathbb{R}\), i.e., for every \(z \in \mathbb{R}\) and every \(\epsilon > 0\), there exists a dyadic rational number \(x\) such that \(|z - x| < \epsilon\).

**Step 2:** Every number \(x \in \mathbb{R}\) either has a unique, non-terminating binary expansion, or it has two binary expansions, one ending in an infinite sequence of 0’s and the other ending in an infinite sequence of 1’s. In either case,
\[
\mathbb{P}\{X = x\} = 0.
\]

**Step 3:** Given numbers \(c_1, \cdots, c_n \in \{0, 1\}\), let \(x\) be the corresponding dyadic rational defined by the equation shown in Step 1 and let \(E(c_1, \cdots, c_n)\) be the interval \((x, x + 2^{-n}]\). Then
\[
\mathbb{P}\{X \in E(c_1, \cdots, c_n)\} = \mathbb{P}\{X_1 = c_1, \cdots, X_n = c_n\} - \mathbb{P}\{X = x\} = \left( \frac{1}{2} \right)^n.
\]
Furthermore, if \(0 \leq a < b \leq 1\), where \(a\) and \(b\) are dyadic rational numbers, then because the interval \((a, b]\) can be written as the disjoint union of finitely many sets of the type \(E(c_1, \cdots, c_n)\) for suitable choices of the \(c_i\)'s, we have that
\[
\mathbb{P}\{X \in (a, b]\} = (b - a).
\]

**Step 4:** Given any two numbers \(0 \leq a < b \leq 1\), not necessarily dyadic rational, we can find a decreasing sequence \((a_n; n \geq 1)\) and an increasing sequence \((b_n; n \geq 1)\) such that \(a_n < b_n\), \(a_n \to a\) and \(b_n \to b\). Then the sets \((a_n, b_n)\) form an increasing sequence, with limit set \((a, b]\), and by the continuity properties of probability measures, we know that
\[
\mathbb{P}\{X \in (a, b]\} = \lim_{n \to \infty} \mathbb{P}\{X \in (a_n, b_n]\} = \lim_{n \to \infty} (b_n - a_n) = b - a.
\]
However, this result implies that \(X\) is uniform on \([0, 1]\).
5.5 The Normal Distribution

Definition 5.5. A continuous random variable $X$ is said to have the normal distribution with mean $\mu$ and variance $\sigma^2$ (denoted $\mathcal{N}(\mu, \sigma^2)$) if the density of $X$ is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad -\infty < x < \infty.$$  

Remark 5.3. The normal distribution is also called the Gaussian distribution.

We can show that the integral of $p(x)$ over $\mathbb{R}$ is equal to 1 by using a geometrically-inspired trick. First, by making the substitution $y = (x - \mu)/\sigma$, we see that

$$\int_{-\infty}^{\infty} p(x) dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy.$$  

Then, letting $I = \int_{-\infty}^{\infty} \exp(-\frac{y^2}{2}) dy$, we obtain

$$I^2 = \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(x^2+y^2)}{2}} dxdy = \int_0^{2\pi} \int_0^{\infty} re^{-\frac{r^2}{2}} d\theta dr = 2\pi.$$  

Here we have changed to polar coordinates in passing from the second to the third line, i.e., we set $x = r \cos(\theta)$, $y = r \sin(\theta)$, and $dxdy = r d\theta dr$. This calculation shows that $I = \sqrt{2\pi}$, which confirms that $p(x)$ is a probability density.

A simple calculation using cumulative distribution functions and densities shows that an affine transformation of a normal random variable is also normally distributed. Specifically, if $X$ is normally distributed with parameters $\mu$ and $\sigma^2$, then the random variable $Y = aX + b$ is normally distributed with parameters $a\mu + b$ and $a^2 \sigma^2$. In particular, if $X$ is a normal random variable with parameters $\mu$ and $\sigma^2$, then $Z = (X - \mu)/\sigma$ is a normal random variable with parameters 0 and 1. $Z$ is known as a standard normal random variable. This identity has several applications.

For statistical applications of the normal distribution, we are often interested in probabilities of the form $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x)$. Although a simple analytical expression is unavailable, these quantities can be calculated numerically and have been extensively tabulated for the standard normal distribution (see Table 5.1 in Ross). Fortunately, because every normal random variable can be expressed in terms of a standard normal random variable, we can use these lookup tables for more general problems. For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then by defining $Z = (X - \mu)/\sigma$, we see that

$$\mathbb{P}(X > x) = \mathbb{P}(\sigma Z + \mu > x) = 1 - \mathbb{P}(Z \leq (x - \mu)/\sigma) = 1 - \Phi((x - \mu)/\sigma),$$  

where $\Phi(x)$ is the CDF of the standard normal distribution:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy.$$
Another application of the relationship between the standard normal distribution and the other normal distributions is illustrated in the proof of the following proposition.

**Proposition 5.3.** Let \( X \) be a normal random variable with parameters \( \mu \) and \( \sigma^2 \). Then the expected value of \( X \) is \( \mu \), while the variance of \( X \) is \( \sigma^2 \).

**Proof.** Let \( Z = (X - \mu)/\sigma \), so that \( Z \sim N(0, 1) \). Then

\[
\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2} dx = 0,
\]

while

\[
\text{Var}(Z) = \mathbb{E}[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx = 1.
\]

Since \( X = \mu + \sigma Z \), the proposition follows from the identities

\[
\mathbb{E}[X] = \mu + \sigma \mathbb{E}[Z] = \mu,
\]

\[
\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2.
\]

\( \square \)

**Remark 5.4.** Notice that each normal distribution is uniquely determined by its mean and variance.

One of the reasons that the normal distribution is so important in statistics is that it provides a counterpart to the Poisson approximation for a binomial distribution with parameters \( n \) and \( p \) when the success probability \( p \) is not small. Recall that if \( X_1, \ldots, X_n \) are independent Bernoulli random variables, each having parameter \( p \), then the distribution of the sum \( S_n = X_1 + \cdots + X_n \) is binomial with parameters \((n, p)\).

**Theorem 5.1.** **DeMoivre-Laplace Limit Theorem:** If \( S_n \) is a binomial random variable with parameters \( n \) and \( p \), then

\[
\lim_{n \to \infty} \mathbb{P} \left\{ a \leq \frac{S_n - np}{\sqrt{npq}} \leq b \right\} = \Phi(b) - \Phi(a).
\]

**Remark 5.5.** One way to interpret this theorem is that it says that when \( n \) is large, the distribution of the random variable \( Z_n = \frac{S_n - np}{\sqrt{npq}} \) can be approximated by the normal distribution with parameters \((0, 1)\). In particular, notice that for all \( n \), \( Z_n \) has mean 0 and variance 1. This result is a special case of a much more general result known as the **Central Limit Theorem**, which states that the distribution of the sum of a large number of independent, identically-distributed random variables, when suitably normalized, is approximately normal. In this particular case, the individual random variables are all Bernoulli. We will prove the Central Limit Theorem for sums of random variables with finite moment generating functions at the end of the course and will be able to deduce the DeMoivre-Laplace theorem as a corollary.
Example 5.4. Suppose that a fair coin is tossed 100 times. What is the probability that the number of heads obtained is between 45 and 55 (inclusive)?

Solution: If $X$ denotes the number of heads obtained in 100 tosses, then $X$ is a binomial random variable with parameters $(100, 1/2)$. By the Demoivre-Laplace theorem, we know that

$$
\Pr\{45 \leq X \leq 55\} = \Pr\left\{-1 \leq \frac{X - 50}{5} \leq 1\right\} \\
\approx \Phi(1) - \Phi(-1) \\
= \Phi(1) - (1 - \Phi(1)) \\
= 2\Phi(1) - 1 = 0.683,
$$

where Table 5.1 in Ross (p. 201) has been consulted for the numerical value of $\Phi(1) = 0.8413$. Notice that we have also made use of the identity

$$
\Phi(-x) = 1 - \Phi(x),
$$

which follows from the fact that if $X \sim \mathcal{N}(0, 1)$, then

$$
\Pr\{X \leq -x\} = \Pr\{X > x\} = 1 - \Pr\{X \leq x\}.
$$

5.6 The Exponential Distribution and Relatives

Definition 5.6. A continuous random variable $X$ is said to have the exponential distribution with parameter $\lambda$ if the density of $X$ is

$$
p(x) = \begin{cases} 
\lambda e^{-\lambda x} & \text{if } x > 0 \\
0 & \text{otherwise.}
\end{cases}
$$

In this case, a simple calculation confirms that $\int_{-\infty}^{\infty} p(x)dx = 1$. Also, the CDF of $X$ can be calculated explicitly:

$$
F_X(x) = \Pr\{X \leq x\} = \int_{0}^{x} \lambda e^{-\lambda y}dy \\
= 1 - e^{-\lambda x}.
$$

The moments of the exponential distribution can be calculated recursively using integration-by-parts. Let $X$ be an exponential random variable with parameter $\lambda$. Then

$$
E[X^n] = \int_{0}^{\infty} x^n \lambda e^{-\lambda x}dx \\
= -x^n e^{-\lambda x}\bigg|_{0}^{\infty} + \int_{0}^{\infty} nx^{n-1} e^{-\lambda x}dx \\
= 0 + \frac{n}{\lambda} \int_{0}^{\infty} x^{n-1} e^{-\lambda x}dx \\
= \frac{n}{\lambda} E[X^{n-1}] \\
= \frac{n!}{\lambda^n},
$$
using the fact that $\mathbb{E}[X^0] = 1$. Taking $n = 1$ and then $n = 2$ gives:

\[
\begin{align*}
\mathbb{E}[X] &= \frac{1}{\lambda} \\
\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.
\end{align*}
\]

Thus, the mean of an exponential random variable is equal to the reciprocal of its parameter, while the variance is equal to the mean squared.

An important property of the exponential distribution is that it is memoryless: if $X$ is exponentially distributed with parameter $\lambda$, then for all $t, s \geq 0$,

\[
\mathbb{P}\{X > t + s | X > t\} = \mathbb{P}\{X > s\}.
\]

To verify this, use the CDF calculated above to obtain

\[
\mathbb{P}\{X > t\} = 1 - \mathbb{P}\{X \leq t\} = e^{-\lambda t},
\]

and then calculate

\[
\mathbb{P}\{X > t + s | X > t\} = \frac{\mathbb{P}\{X > t + s\}}{\mathbb{P}\{X > t\}} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}\{X > s\}.
\]

In other words, if we think of $X$ as being the (random) time until some event occurs, then conditional on the event not having occurred by time $t$, i.e., on $X > t$, the distribution of the time remaining until the event does occur, $X - t$, is also exponential with parameter $\lambda$. It is as if the conditional probability of the event occurring between times $t$ and $t + s$ has no ‘memory’ of the amount of time that has elapsed up to that point. Remarkably, the converse of this statement is also true.

**Proposition 5.4.** Suppose that $X$ is a continuous random variable with values in $[0, \infty)$. Then $X$ is memoryless if and only $X$ is exponentially distributed for some parameter $\lambda > 0$.

**Proof.** See footnote on page 211 of Ross.

**Interpretation:** Exponential random variables are often used to model the distribution of the amount of time elapsed until some particular event occurs. In this case, the memorylessness property implies that the likelihood that the event occurs in some short interval $[t, t+\delta t]$ conditional on it not yet having occurred does not depend on $t$. Notice that for events that occur at discrete times, the geometric distribution has a similar property: the probability of observing the event in question on any particular trial given that it has not yet occurred is just $p$. In fact, the exponential distribution can be thought of as a limiting case for the geometric distribution when the success probability $p$ is small and when time is measured in units that are of size $1/p$. 
Proposition 5.5. For each \( n \geq 1 \), let \( X_n \) be a geometric random variable with parameter \( \lambda/n \). Then, for every \( t \geq 0 \),
\[
\lim_{n \to \infty} P\{X_n \leq nt\} = 1 - e^{-\lambda t}.
\]

Proof. First, recall that
\[
\lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{nx} = e^{-\lambda x}.
\]
The result can then be deduced by writing \( x = k/n \) below and approximating the sum by an integral with respect to \( x \):
\[
P\{X_n \leq nt\} = \sum_{k=1}^{\left\lfloor nt\right\rfloor} \left(1 - \frac{\lambda}{n}\right)^{k-1} \frac{\lambda}{n} \\
= \frac{1}{n} \sum_{k=0}^{\left\lfloor nt\right\rfloor-1} \left(1 - \frac{\lambda}{n}\right)^{n-k/n} \lambda \\
\approx \frac{1}{n} \sum_{k=0}^{\left\lfloor nt\right\rfloor-1} \lambda e^{-\lambda k/n} \\
\approx \int_0^t \lambda e^{-\lambda x} dx \\
= 1 - e^{-\lambda t}. 
\]

This proposition implies that if \( X_n \) is geometric with parameter \( \lambda/n \), then the distribution of the random variable \( T_n = \frac{1}{n}X_n \) is approximately exponential with parameter \( \lambda \) when \( n \) is large.

5.6.1 The Gamma Distribution

Definition 5.7. A random variable \( X \) is said to have the gamma distribution with shape parameter \( \alpha > 0 \) and scale parameter \( \lambda > 0 \) if its density is
\[
p(x) = \begin{cases} 
\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{if } x > 0 \\
0 & \text{otherwise,}
\end{cases}
\]
where the gamma function \( \Gamma(\alpha) \) is defined (for \( \alpha > 0 \)) by the formula
\[
\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.
\]

Remark 5.6. Notice that the exponential distribution with parameter \( \lambda \) is identical to the gamma distribution with parameters \((1, \lambda)\). Later we will show that if \( X_1, \ldots, X_n \) are independent exponential RVs, each with parameter \( \lambda \), then the sum \( X = X_1 + \cdots + X_n \) is a gamma RV with parameters \((n, \lambda)\). Because of this relationship, the gamma distribution is often used to model life spans.
Integration by parts shows that for \( \alpha > 1 \), the gamma function satisfies the following important recursion:

\[
\Gamma(\alpha) = -y^{-\alpha - 1}e^{-y}\bigg|_0^\infty + (\alpha - 1)\int_0^\infty y^{\alpha - 2}e^{-y}dy \\
= (\alpha - 1)\Gamma(\alpha - 1).
\]

In particular, if \( \alpha = n \) is a positive integer, then because

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

we obtain

\[
\Gamma(n) = (n - 1)\Gamma(n - 2) = (n - 1)(n - 2)\Gamma(n - 3) = (n - 1)(n - 2) \cdots 3 \cdot 2\Gamma(1) = (n - 1)!
\]

This recursion can be used to calculate the moments of the gamma distribution. If \( X \) has the gamma distribution with parameters \((\alpha, \lambda)\), then

\[
E[X^n] = \int_0^\infty x^n \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1}e^{-\lambda x}dx \\
= \lambda^{-n}\frac{\Gamma(n + \alpha)}{\Gamma(\alpha)} \int_0^\infty \frac{\lambda^{n+\alpha}}{\Gamma(n + \alpha)} x^{n+\alpha - 1}e^{-\lambda x}dx \\
= \lambda^{-n} \prod_{k=0}^{n-1} (\alpha + k).
\]

In particular, taking \( n = 1 \) and \( n = 2 \) gives

\[
E[X] = \frac{\alpha}{\lambda} \\
Var(X) = E[X^2] - E[X]^2 = \frac{\alpha(\alpha + 1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.
\]

### 5.6.2 The Beta Distribution

**Definition 5.8.** A random variable \( X \) is said to have the beta distribution with parameters \( a, b > 0 \) if its density is

\[
p(x) = \begin{cases} 
\frac{1}{\beta(a,b)} x^{a-1}(1 - x)^{b-1} & \text{if } x \in (0, 1) \\
0 & \text{otherwise},
\end{cases}
\]

where the beta function \( \beta(a,b) \) is defined (for \( a, b > 0 \)) by the formula

\[
\beta(a,b) = \int_0^1 x^{a-1}(1 - x)^{b-1}dx.
\]
Remark 5.7. Notice that if $a = b = 1$, then $X$ is simply a standard uniform random variable. Also, if $X_1$ and $X_2$ are independent gamma-distributed RVs with parameters $(a, \theta)$ and $(b, \theta)$, respectively, then the random variable $X = X_1/(X_1 + X_2)$ is beta-distributed with parameters $(a, b)$. In particular, if $X_1$ and $X_2$ are independent exponential RVs with parameter $\lambda$, then $X = X_1/(X_1 + X_2)$ is uniformly distributed on $[0, 1]$. The beta distribution is often used to model distributions of frequencies.

It can be shown that the beta function and the gamma function are related by the following identity:

$$\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}.$$ 

In turn, we can use this identity to evaluate the moments of the beta distribution. Let $X$ be a beta random variable with parameters $(a, b)$. Then

$$E[X^n] = \frac{1}{\beta(a, b)} \int_0^1 x^n x^{a-1}(1 - x)^{b-1} dx$$

$$= \frac{\beta(n + a, b)}{\beta(a, b)}$$

$$= \frac{\Gamma(n + a)\Gamma(b)}{n + a + b} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}$$

$$= \frac{n!}{\Gamma(n + a + b)}$$

$$= \prod_{k=0}^{n-1} \frac{a + k}{a + b + k}.$$

Taking $n = 1$ and $n = 2$ gives

$$E[X] = \frac{a}{a + b}$$

$$Var(X) = E[X^2] - E[X]^2$$

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

$$= \frac{ab}{(a + b)^2(a + b + 1)}.$$

5.7 Compositions and Distributions

Suppose that $X$ is a real-valued random variable and let $Y = g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is a measurable function. Then

$$P(Y \in E) = P(g(X) \in E) = P(X \in g^{-1}(E)),$$

where the set $g^{-1}(E)$ is the preimage of $E$ under $g$:

$$g^{-1}(E) = \{x \in \mathbb{R} : g(x) \in E\}.$$ 

This result shows how the distribution of a random variable is altered by the transformation $g$. Provided that $g$ is not too complicated, we can sometimes express the cumulative distribution function of $Y$ in terms of that of $X$. We first see how this is done in two examples and then give some general results.
Example 5.5. *(Ross, Example 7b)*: Let $X$ be a non-negative random variable with CDF $F_X$ and let $Y = X^n$, where $n \geq 1$. Then, for any $x \geq 0$,

\[
F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(X^n \leq x) = \mathbb{P}(X \leq x^{1/n}) = F_X(x^{1/n}),
\]

while for any $x < 0$, we have $F_Y(x) = 0$. Similarly, if $X$ has a density $p_X = F_X'(x)$, then $Y$ also has a density given by

\[
p_Y(x) = \frac{d}{dx} F_Y(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\frac{1}{n} x^{\frac{1}{n}-1} p_X(x^{1/n}) & \text{if } x \geq 0.
\end{cases}
\]

Example 5.6. *(Ross, Example 7c)* Let $X$ be a continuous real-valued random variable with density $p_X(x) = F_X'(x)$ and let $Y = X^2$. Then, for $x \geq 0$,

\[
F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(X^2 \leq x) = \mathbb{P}(-\sqrt{x} \leq X \leq \sqrt{x}) = F_X(\sqrt{x}) - F_X(-\sqrt{x}),
\]

while for any $x < 0$, we have $F_Y(x) = 0$. In this case, the density of $Y$ is given by

\[
p_Y(x) = \frac{d}{dx} F_Y(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\frac{1}{2\sqrt{x}} (p_X(\sqrt{x}) + p_X(-\sqrt{x})) & \text{if } x \geq 0.
\end{cases}
\]

Notice that we obtain different expressions for $F_Y(x)$ in Examples 5.5 and 5.6 in the case $n = 2$ depending on whether we allow $X$ to take only positive values or both positive and negative values. The reason for this is that the function $x \rightarrow x^2$ is one-to-one when restricted to $[0, \infty)$ but is two-to-one when considered on the entire real line.

**Monotonic transformations:** Suppose that $X$ is a real-valued random variable with CDF $F_X$ and let $Y = g(X)$, where the function $g : \mathbb{R} \rightarrow (l, r)$ is strictly increasing, continuous, one-to-one, and onto, i.e., if $x < y$, then $g(x) < g(y)$. Then $g$ is invertible on the interval $(l, r)$ and so the CDF of $Y$ can be expressed as

\[
F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(g(X) \leq x) = \mathbb{P}(X \leq g^{-1}(x)) = F_X(g^{-1}(x)),
\]

for any $x \in (l, r)$. Of course, if $x \geq r$, then $\mathbb{P}(Y \leq x) = 1$, while if $x \leq l$, then $\mathbb{P}(Y \leq x) = 0$. Combining these facts, we obtain the general expression

\[
F_Y(x) = \begin{cases} 
0 & \text{if } x \leq l \\
F_X(g^{-1}(x)) & \text{if } x \in (l, r) \\
1 & \text{if } x \geq r.
\end{cases}
\]
If instead $g$ is a decreasing function, then
\[ F_Y(x) = \mathbb{P}\{g(X) \leq x\} = \mathbb{P}\{X \geq g^{-1}(x)\} = 1 - F_X(g^{-1}(x)-) = 1 - F_X(g^{-1}(x)-) \]
since
\[ F_X(x-) = F_X(x) \]
whenever $F$ is continuous. In this case, the CDF of $Y$ has the form
\[
F_Y(x) = \begin{cases} 
0 & \text{if } x \leq l \\
1 - F_X(g^{-1}(x)) & \text{if } x \in (l, r) \\
1 & \text{if } x \geq r. 
\end{cases}
\]

The next proposition describes an important special case of these identities.

**Theorem 5.2.** Suppose that $U$ is a uniform random variable on $(0, 1)$ and let $X$ be a real-valued random variable with a strictly increasing continuous CDF $F_X(\cdot)$. Then the random variable $Y = F_X^{-1}(U)$ has the same distribution as $X$, while the random variable $Z = F_X(X)$ is uniformly distributed on $(0, 1)$.

**Proof.** First consider the distribution of $Y$:
\[
F_Y(x) = \mathbb{P}\{Y \leq x\} = \mathbb{P}\{F_X^{-1}(U) \leq x\} = \mathbb{P}\{U \leq F_X(x)\} = F_X(x),
\]
since $F_X(x) \in [0, 1]$ for all $x$ and $\mathbb{P}\{U \leq y\} = y$ whenever $y \in [0, 1]$. This shows that $Y$ and $X$ have the same distribution.

Similarly, the CDF of $Z = F_X(X)$ is
\[
F_Z(x) = \mathbb{P}\{Z \leq x\} = \mathbb{P}\{F_X(X) \leq x\} = \mathbb{P}\{X \leq F_X^{-1}(x)\} = F_X(F_X^{-1}(x)) = x,
\]
for any $x \in [0, 1]$, which shows that $Z$ is a uniform random variable on $[0, 1]$.

**Remark 5.8.** This result has important applications in computational statistics and stochastic simulations. For example, many computational statistics packages include routines that will simulate a sequence of independent uniform random variables $U_1, U_2, \ldots$ (or some suitable approximation thereof). Then, if $X$ is a random variable with a strictly increasing CDF, we can generate a sequence of independent random variables $X_1, X_2, \ldots$, each having the same distribution as $X$, by setting $X_i = F_X^{-1}(U_i)$. This is computationally efficient if the inverse $F_X^{-1}$ can be easily evaluated.
Example 5.7. Recall that the CDF of the exponential distribution with parameter $\lambda$ is
\[ F_X(x) = 1 - e^{-\lambda x}. \]
A simple calculation shows that
\[ F_X^{-1}(x) = -\frac{1}{\lambda} \ln(1 - x), \]
and so it follows that if $U$ is uniform on $[0, 1]$, then
\[ Y = -\frac{1}{\lambda} \ln(1 - U) \]
is exponentially distributed with parameter $\lambda$. In fact, because the distribution of $1 - U$ is also uniform on $[0, 1]$, the random variable
\[ Y' = -\frac{1}{\lambda} \ln(U) \]
is also exponentially distributed with parameter $\lambda$.

These results of the last section can also be used to express the density of $Y = g(X)$ in terms of the density of $X$ whenever $X$ is a continuous random variable and $g$ is strictly monotonic.

Proposition 5.6. Let $X$ be a continuous random variable with density $p_X(\cdot)$ and suppose that $g$ is a strictly monotonic continuous function with differentiable inverse $g^{-1}$. Then $Y = g(X)$ is a continuous random variable with density function
\[
p_Y(y) = \begin{cases} 
\left| \frac{d}{dy} g^{-1}(y) \right| \cdot p_X(g^{-1}(y)) & \text{if } y = g(x) \text{ for some } x \\
0 & \text{otherwise.}
\end{cases}
\]
Proof. Without loss of generality, we may assume that $g$ is increasing. Suppose that $y = g(x)$ for some $x$. Then, as shown above, $F_Y(y) = F_X(g^{-1}(y))$, and differentiating with respect to $y$ gives
\[
p_Y(y) = \frac{d}{dy} F_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| \cdot p_X(g^{-1}(y)).
\]
Chapter 6

Random Vectors

6.1 Multivariate Distributions

Except in rare cases, most efforts to use probabilistic models to describe real-world phenomena must take into account more than one random quantity at a time. For example, if we are interested in understanding temporal fluctuations in quantities such as ocean surface temperatures or currency exchange rates, then we will need to study time series of data that can be represented as a sequence of random variables, $X_1, X_2, \cdots$, where the variable $X_t$ denotes the quantity at interest at the time of the $t$’th measurement. Similarly, if we are conducting a genome-wide association study to identify genetic variants that predispose individuals to complex diseases such as cancer or diabetes, then we will need to consider an array of random variables, $X_{i1}, \cdots, X_{in}, Y_i$, where $Y_i$ indicates whether the $i$’th individual in the study is affected (a case) or not (a control) and $X_{ij}$ indicates which nucleotide \{A,T,C,G\} is present at the $j$’th site in this individual.

To build probabilistic models for these kinds of data, we must introduce groups of random variables $X_1, \cdots, X_n$ that are all defined on the same probability space. Provided that these are ordered in some manner, we can treat each group of variables as a random vector $X = (X_1, \cdots, X_n)$ which is a map from the sample space $\Omega$ into $\mathbb{R}^n$:

$$X(\omega) \equiv (X_1(\omega), \cdots, X_n(\omega)) \in \mathbb{R}^n.$$  

Our main goal in this chapter is to develop some tools that can be used to describe and study such random vectors.

**Definition 6.1.** Suppose that $X_1, X_2, \cdots, X_n$ are real-valued random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the **joint distribution** of these variables is the probability distribution $Q$ of the random vector $X = (X_1, \cdots, X_n)$ defined by

$$Q(E) = \mathbb{P}\{(X_1, \cdots, X_n) \in E\}$$  

for any (measurable) subset $E \subset \mathbb{R}^n$. Likewise, the **joint cumulative distribution function** of $(X_1, \cdots, X_n)$ is defined by

$$F(x_1, \cdots, x_n) = \mathbb{P}(X_1 \leq x_1, \cdots, X_n \leq x_n).$$

As with real-valued random variables, both the joint distribution and the joint CDF are fully determined by the other. In particular, this means that if two random vectors $X = (X_1, \cdots, X_n)$ and $Y = (Y_1, \cdots, Y_n)$ have the same joint cumulative distribution function, then their components have the same joint distribution.
If \( X = (X_1, \cdots, X_n) \) is a random vector, then each variable \( X_i \) is itself a real-valued random variable and the distribution of \( X_i \) is called the **marginal distribution** of \( X_i \). Note that each component variable has a marginal distribution. By exploiting the continuity properties of probability distributions, the marginal cumulative distribution function of \( X_i \) can be determined from the joint cumulative distribution of the random vector. For example, the marginal CDF of \( X_1 \) is equal to

\[
F_{X_1}(x) = \mathbb{P}\{X_1 \leq x\} = \mathbb{P}\{X_1 \leq x, X_2 < \infty, \cdots, X_n < \infty\} = \mathbb{P}\left( \bigcup_{M \geq 1} \{X_1 \leq x, X_2 < M, \cdots, X_n < M\} \right) = \lim_{M \to \infty} F_X(x_1, M, \cdots, M).
\]

This process of recovering the marginal distribution of a single random variable from the joint distribution of the random vector is called **marginalisation**.

**Caveat:** The joint distribution of the variables \( (X_1, \cdots, X_n) \) fully determines all of the marginal distributions. However, knowledge of the marginal distributions alone is not sufficient to determine the joint distribution: there are usually infinitely many joint distributions that have the same marginal distributions.

**Definition 6.2.** If \( X_1, \cdots, X_n \) are all discrete random variables and \( X_i \) takes values in the countable set \( E_i \), then \( X = (X_1, \cdots, X_n) \) is a discrete random vector with values in the product space \( E = E_1 \times \cdots \times E_n \) and the **joint probability mass function** of \( X \) is defined by

\[
p(x_1, \cdots, x_n) = \mathbb{P}\{X_1 = x_1, \cdots, X_n = x_n\}.
\]

If we are given the joint probability mass function \( p_X \) of a discrete random vector \( X = (X_1, \cdots, X_n) \), then we can find the marginal probability mass functions of the variables \( X_i \) by summing \( p_X \) over appropriate subsets of \( E \). For example, the marginal probability mass function of \( X_1 \) is given by the following sum:

\[
p_{X_1}(x) = \mathbb{P}(X_1 = x) = \sum_{x_2, \cdots, x_n} p_X(x, x_2, \cdots, x_n).
\]

In other words, we can calculate the marginal probability mass function for \( X_i \) at a point \( x \in E_i \) by summing \( p_X \) over all possible vectors in \( E \) with the \( i \)'th coordinate held fixed at \( x \).

**Example 6.1.** Suppose that \( n \) independent experiments are performed, each of which can have one of \( r \) possible outcomes (labeled \( 1, \cdots, r \)) with respective probabilities \( p_1, \cdots, p_r \), where \( p_1 + \cdots + p_r = 1 \). Let \( X_i \) denote the number of experiments with outcome \( i \). Then \( X = (X_1, \cdots, X_n) \) is a discrete random vector with probability mass function

\[
p(n_1, \cdots, n_r) = \mathbb{P}\{X_1 = n_1, \cdots, X_r = x_r\} = \binom{n}{n_1, \cdots, n_r} p_1^{n_1} \cdots p_r^{n_r},
\]

and \( X \) is said to have the **multinomial distribution** with parameters \( n \) and \( (p_1, \cdots, p_r) \). Notice that \( X_1 + \cdots + X_n = n \).
In other words, the random variables for all \( x \in \mathbb{R}^n \) are independent if every finite subcollection is independent.

**Definition 6.4.** When a collection of \( n \) random variables \( X_1, \ldots, X_n \) are independent, the joint probability density function exists a function \( p(x_1, \ldots, x_n) \), called the **joint probability density function**, such that

\[
\mathbb{P}\{(X_1, \ldots, X_n) \in E\} = \int \cdots \int_{(x_1, \ldots, x_n) \in E} p(x_1, \ldots, x_n) dx_1 \cdots dx_n
\]

for any (measurable) subset \( E \subset \mathbb{R}^n \).

As in the univariate case, the joint density function can be found by differentiating the joint cumulative distribution function. However, when there are multiple variables, we must take partial derivatives for each component of the vector:

\[
p_X(x_1, \ldots, x_n) = \partial_{x_1} \cdots \partial_{x_n} F_X(x_1, \ldots, x_n).
\]

Furthermore, if the variables \( X_1, \ldots, X_n \) are jointly continuous, then each variable \( X_i \) is marginally continuous and its marginal density function can be found by integrating the joint density function over all coordinates except \( x_i \):

\[
p_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_X(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.
\]

**Caveat:** \( X_1, \ldots, X_n \) can all be individually continuous even when they are not jointly continuous. For example, the random vector \((X, X)\) is never continuous even if \( X \) is itself a continuous random variable.

**Example 6.2.** The random vector \( X = (X_1, \ldots, X_{n-1}) \) is said to have the **Dirichlet distribution** with parameters \( \alpha_1, \ldots, \alpha_n > 0 \) if its components are jointly continuous with joint density function

\[
p_X(x_1, \ldots, x_{n-1}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \prod_{i=1}^{n} x_i^{\alpha_i}
\]

for all \( x_1, \ldots, x_{n-1} \geq 0 \) satisfying \( x_1 + \cdots + x_{n-1} \leq 1 \), where \( x_n = 1 - x_1 - \cdots - x_{n-1} \). Since \( x_1 + \cdots + x_n = 1 \) by construction, the Dirichlet distribution is usually regarded as a continuous distribution on the standard \( n-1 \)-dimensional simplex

\[
K_{n-1} = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i\right\}
\]

When \( n = 2 \), the Dirichlet distribution reduces to the Beta distribution.

### 6.2 Independent Random Variables

**Definition 6.4.** A collection of \( n \) random variables \( X_1, \ldots, X_n \), all defined on the same probability space, are said to be **independent** if for every collection of (measurable) sets \( A_1, \ldots, A_n \),

\[
\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in A_i\}.
\]

In other words, the random variables \( X_1, \ldots, X_n \) are independent if and only if for all such subsets \( A_1, \ldots, A_n \subset \mathbb{R} \), the sets \( \{X_1 \in A_1\}, \ldots, \{X_n \in A_n\} \) are independent. Also, we say that an infinite collection of random variables is independent if every finite subcollection is independent.
6.2. INDEPENDENT RANDOM VARIABLES

Heuristically, $X_1$ and $X_2$ are independent if knowing the value of $X_1$ provides us with no information about the value of $X_2$, and vice versa.

**Proposition 6.1.** Let $X_1, \ldots, X_n$ be jointly distributed random variables. Then each of the following conditions implies independence:

(a) The joint cumulative distribution function factors into a product of the marginal cumulative distribution functions, i.e., for all real numbers $x_1, \ldots, x_n$,

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i).$$

(b) $X_1, \ldots, X_n$ are discrete and the joint probability mass function factors into a product of the marginal probability mass functions:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i)$$

for all $x_1, \ldots, x_n$.

(c) $X_1, \ldots, X_n$ are jointly continuous and the joint probability density function factors into a product of the marginal probability density functions:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_i(x_i)$$

for all $x_1, \ldots, x_n$.

**Example 6.3.** (Ross, Example 2b) Let $Z$ be a Poisson random variable with parameter $\lambda$, let $W_1, W_2, \ldots$ be a collection of independent Bernoulli random variables, each with parameter $p$, and define

$$X = \sum_{i=1}^{Z} W_i,$$

$$Y = Z - X.$$

Then $X$ and $Y$ are independent Poisson random variables with parameters $\lambda p$ and $\lambda (1-p)$, respectively. To prove this, we calculate

$$\mathbb{P}\{X = i, Y = j\} = \mathbb{P}\{X = i, Z = i + j\}$$

$$= \mathbb{P}\{X = i | Z = i + j\} \cdot \mathbb{P}\{Z = i + j\}$$

$$= \binom{i + j}{i} p^i (1-p)^j \cdot e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}$$

$$= \frac{e^{-\lambda p} (\lambda p)^i}{i!} \cdot \frac{e^{-\lambda (1-p)} (\lambda (1-p))^j}{j!}$$

$$\equiv p_X(i) \cdot p_Y(j),$$

where $p_X$ and $p_Y$ are the probability mass functions of Poisson random variables with parameters $\lambda p$ and $\lambda (1-p)$. Independence of $X$ and $Y$ then follows from the factorization of the joint probability mass function.
6.3 Sums of Independent Random Variables

**Definition 6.5.** If $f, g : \mathbb{R} \to \mathbb{R}$ are two integrable real-valued functions, then the **convolution** of $f$ and $g$ is the real-valued function $f * g : \mathbb{R} \to \mathbb{R}$ defined by the formula

$$(f * g)(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx = \int_{-\infty}^{\infty} f(z-x)g(x)dx = (g * f)(z).$$

The identity between the first and second line follows from a simple change of variables and shows that convolution is a **commutative** operation: $f * g = g * f$.

**Application:** In probability theory, convolutions arise when we consider the distribution of sums of independent random variables. To see this, suppose that $X$ and $Y$ are independent, continuous random variables with densities $p_x$ and $p_y$. Then $X + Y$ is a continuous random variable with cumulative distribution function

$$F_{X+Y}(z) = \mathbb{P}\{X + Y \leq z\} = \int_{x+y\leq z} p_X(x)p_Y(y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} p_X(x)p_Y(y)dxdy = \int_{-\infty}^{\infty} F_X(z-y)p_Y(y)dy = \int_{-\infty}^{\infty} p_Y(y)p_X(x)dydx = \int_{-\infty}^{\infty} F_Y(z-x)p_X(x)dx,$$

where the expression in the fourth line is the convolution $F_X * p_Y$ and the expression in the sixth line is the convolution $F_Y * p_X$. The density of $X + Y$ can then be found by differentiating the CDF, giving

$$p_{X+Y}(z) = \frac{d}{dz} F_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(z-y)p_Y(y)dy = \int_{-\infty}^{\infty} p_Y(z-x)p_X(x)dx,$$

which is equal to the convolution of the density functions of $X$ and $Y$:

$$p_{X+Y}(z) = (p_X * p_Y)(z) = (p_Y * p_X)(z).$$

**Proposition 6.2.** If $X_1, \ldots, X_n$ are independent gamma-distributed random variables with parameters $(t_i, \lambda)$, then the sum $X = X_1 + \cdots + X_n$ is a gamma-distributed RV with parameters $(\sum_{i=1}^{n} t_i, \lambda)$. 
Proof. It suffices to consider the case $n = 2$, since the full result then follows by induction on $n$. In this case, the result follows from the calculation

$$p_{X_1+X_2}(z) = \frac{1}{\Gamma(s+t)} \int_0^z \lambda e^{-\lambda(z-x)} (\lambda(z-x))^{s-1} \lambda e^{-\lambda x} (\lambda x)^{t-1} dx$$

where $C$ is a constant. However, since $p_{X_1+X_2}$ is a density, we know that it integrates to 1 and so we can calculate

$$C = \left( \int_0^\infty e^{-\lambda z} z^{s+t-1} dz \right)^{-1} = \frac{\lambda^{s+t}}{\Gamma(s+t)}.$$

This shows that

$$p_{X_1+X_2}(z) = \frac{\lambda^{s+t}}{\Gamma(s+t)} z^{s+t-1} e^{-\lambda z},$$

and so $X_1 + X_2$ is Gamma-distributed with parameters $(s+t, \lambda)$ as claimed.

Corollary 6.1. Because the exponential distribution with parameter $\lambda$ is the same as the gamma distribution with parameters $(1, \lambda)$, it follows that if $X_1, \cdots, X_n$ are independent exponential RVs all with parameter $\lambda$, then the sum $X = X_1 + \cdots + X_n$ is a gamma RV with parameters $(n, \lambda)$.

Proposition 6.3. If $X_1, \cdots, X_n$ are independent normal RVs with parameters $(\mu_i, \sigma_i^2), i = 1, \cdots, n$, then their sum $X = X_1 + \cdots + X_n$ is a normal RV with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.

Proof. It suffices to consider the case $n = 2$, since the full result then follows by induction on $n$. Also, because $X_i - \mu_i$ is normally distributed with parameters $(0, \sigma_i^2)$, we may assume that $\mu_1 = \mu_2 = 0$.

Then, using the convolution formula, we see that the density of $X + Y$ is

$$p_{X+Y}(z) = \int_{-\infty}^{\infty} p_X(z-x)p_Y(x)dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left(-\frac{(z-x)^2}{2\sigma_1^2}\right) \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left(-\frac{x^2}{2\sigma_2^2}\right) dx$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{z^2}{2\sigma_1^2}\right) \int_{-\infty}^{\infty} \exp\left(-\left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right) x^2 + \frac{2}{\sigma_1^2} x + \frac{z}{\sigma_1^2}\right) dx$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{z^2}{2\sigma_1^2}\right) \exp\left(2\left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} \frac{z^2}{4\sigma_1^2}\right)$$

$$\times \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right) \left(x - \left(\frac{1}{2\sigma_1^2} + \frac{1}{2\sigma_2^2}\right)\right)^{-1} z \frac{z}{\sigma_1^2}\right] dx$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} (2\pi)^{1/2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1/2} \exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right)$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} \exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right),$$

which shows that $X_1 + X_2 \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$.

\qed
Convolutions can also be defined for discrete distributions. For example, if $X$ and $Y$ are independent integer-valued random variables with probability mass functions $p_X$ and $p_Y$, then $X + Y$ is also an integer-valued random variable with probability mass function

$$p_{X+Y}(n) = \mathbb{P}\{X + Y = n\} = \sum_k \mathbb{P}\{X = k, Y = n-k\} = \sum_k \mathbb{P}\{X = k\} \mathbb{P}\{Y = n-k\} = \sum_k p_X(k)p_Y(n-k).$$

The expression in the last line of this series of equations can be interpreted as a discrete convolution.

**Proposition 6.4.** $X_1, \ldots, X_n$ are independent Poisson RVs with parameters $\lambda_1, \ldots, \lambda_n$, respectively, then the sum $X = X_1 + \cdots + X_n$ is a Poisson RV with parameter $\lambda_1 + \cdots + \lambda_n$.

**Proof.** It suffices to prove the result for the case when $n = 2$, since the general result follows by induction. Using the discrete convolution formula (and noting that $X_1$ and $X_2$ are both non-negative), the probability mass function of $X_1 + X_2$ is

$$p_{X_1+X_2}(n) = \sum_{k=0}^n p_{X_1}(k)p_{X_2}(n-k) = \sum_{k=0}^n e^{-\lambda_1} \frac{\lambda_1^k}{k!} e^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{k=0}^n \binom{n}{k} \lambda_1^k (1 - \lambda_2)^{n-k} = e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!},$$

which is also the probability mass function for the Poisson distribution with parameter $\lambda_1 + \lambda_2$. \qed

### 6.4 Conditional Distributions

#### 6.4.1 Discrete Conditional Distributions

**Definition 6.6.** If $X$ and $Y$ are discrete random variables with joint probability mass function $p_{X,Y}(x,y)$, then we define the **conditional probability mass function** of $X$ given $Y = y$ by

$$p_{X\mid Y}(x\mid y) = \mathbb{P}\{X = x\mid Y = y\} = \frac{\mathbb{P}\{X = x, Y = y\}}{\mathbb{P}\{Y = y\}} = \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$
In other words, since $X$ and $Y$ are defined on the same probability space, we can consider the distribution of $X$ conditional on the event $Y = y$. This is of great importance in statistics, since we are often interested in the distributions of variables that we can only indirectly estimate using observations of other presumably correlated random variables. Of course, if $X$ and $Y$ are independent, then the conditional probability mass function of $X$ given $Y = y$ is equal to the unconditioned probability mass function:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x).$$

In other words, if $X$ and $Y$ are independent, then $Y$ provides no information concerning $X$.

**Example 6.4.** (Ross, Example 4b) Suppose that the $X$ and $Y$ are independent Poisson RVs with parameters $\lambda_1$ and $\lambda_2$. Then, using the fact that $X + Y$ is a Poisson RV with parameter $\lambda_1 + \lambda_2$, the conditional probability mass function of $X$ given that $X + Y = n$ is

$$\Pr\{X = k|X + Y = n\} = \frac{\Pr\{X = k, Y = n - k\}}{\Pr\{X + Y = n\}}$$

$$= \frac{\Pr\{X = k\} \Pr\{Y = n - k\}}{\Pr\{X + Y = n\}}$$

$$= e^{-\lambda_1} \left(\frac{\lambda_1^k}{k!}\right) e^{-\lambda_2} \left(\frac{\lambda_2^{n-k}}{(n-k)!}\right) \left(e^{-(\lambda_1 + \lambda_2)}\right)^n \frac{1}{n!}$$

$$= \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k},$$

for $k = 0, \ldots, n$, and so it follows that the conditional distribution of $X$ given that $X + Y = n$ is Binomial with parameters $n$ and $p = \lambda_1/(\lambda_1 + \lambda_2)$. Notice that once we condition on the event $X + Y = n$, $X$ and $Y$ are no longer independent.

The next example illustrates how the conditional joint distribution of more than two RVs can be calculated.

**Example 6.5.** (Ross, Example 4c) Recall that $(X_1, \ldots, X_k)$ has the multinomial distribution with parameters $n$ and $(p_1, \ldots, p_k)$, $p_1 + \cdots + p_k = 1$, if the joint probability mass function has the form

$$\Pr\{X_1 = n_1, \ldots, X_k = n_k\} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}, \quad n_i \geq 0, \sum_{i=1}^k n_i = n.$$

Notice that by combining groups of the component variables, we again obtain a multinomial RV. For example, if we combine the first $r$ categories into a single type and we define $F_r = 1 - p_{r+1} - \cdots - p_k = p_1 + \cdots + p_r$, then the random vector $(X_1 + \cdots + X_r, X_{r+1}, \ldots, X_k)$ has the multinomial distribution with parameters $n$ and $(F_r, p_{r+1}, \ldots, p_k)$. This can be shown by calculating

$$\Pr\{X_1 + \cdots + X_r = l, X_{r+1} = n_{r+1}, \ldots, X_k = n_k\}$$

$$= \sum_{n_1 + \cdots + n_r = l} \Pr\{X_1 = n_1, \ldots, X_k = n_k\}$$

$$= \sum_{n_1 + \cdots + n_r = l} \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_{r-1}^{n_{r-1}} p_r^{n_r} p_{r+1}^{n_{r+1}} \cdots p_k^{n_k}$$

$$= \frac{n!}{(n_{r+1}! \cdots n_k!)} F_r^{n_{r+1}} \cdots p_k^{n_k},$$
provided \( l + n_{r+1} + \cdots + n_k = n \). The last identity is a consequence of the multinomial formula.

Now suppose that we observe that \( X_{r+1} = n_{r+1}, \cdots, X_k = n_k \). Then, if we set \( m = n_{r+1} + \cdots + n_k \) and let \( F_r \) be defined as in the preceding paragraph, the conditional joint distribution of \((X_1, \cdots, X_r)\) given this observation is equal to

\[
\mathbb{P}\{X_1 = n_1, \cdots, X_r = n_r | X_{r+1} = n_{r+1}, \cdots, X_k = n_k\} = \frac{\mathbb{P}\{X_1 = n_1, \cdots, X_k = n_k\}}{\mathbb{P}\{X_{r+1} = n_{r+1}, \cdots, X_k = n_k\}} = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_r^{n_{r+1}} \cdots p_k^{n_k} \left( \frac{n!}{(n-m)! n_{r+1}! \cdots n_k!} F_r^{m-m} p_{r+1} \cdots p_k \right)^{-1} = \frac{(n-m)!}{n_1! \cdots n_r!} \left( \frac{p_1}{F_r} \right)^{n_1} \cdots \left( \frac{p_r}{F_r} \right)^{n_r}.
\]

In other words, the conditional distribution of \((X_1, \cdots, X_r)\) given the values of the random variables \(X_{r+1}, \cdots, X_k\) is again multinomial with parameters \( n - m \) and \((p_1/F_r, \cdots, p_r/F_r)\).

### 6.4.2 Continuous Conditional Distributions

**Definition 6.7.** If \( X \) and \( Y \) are jointly continuous random variables with joint probability density \( p_{X,Y}(x,y) \) and marginal densities \( p_X(x) \) and \( p_Y(y) \), then for any \( y \) such that \( p_Y(y) > 0 \), we can define the **conditional probability density** of \( X \) given that \( Y = y \) by

\[
p_{X|Y}(x|y) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\{X \in (x, x+h]|Y \in (y, y+h)\} = \lim_{h \to 0} \frac{1}{h} \mathbb{P}\{X \in (x, x+h]|Y \in (y, y+h)\} = \frac{p_{X,Y}(x,y)}{p_Y(y)}.
\]

Similarly, we can define the **conditional cumulative distribution function** of \( X \) given that \( Y = y \) by

\[
F_{X|Y}(x|y) = \lim_{h \to 0} \mathbb{P}\{X \leq x|Y \in (y-h, y+h)\} = \int_{-\infty}^{x} f_{X|Y}(z,y)dz.
\]

Because \( Y \) is marginally continuous, we are effectively conditioning on an event with zero probability: \( \mathbb{P}\{Y = y\} = 0 \). That we are able to make sense of this is a consequence of the fact that \( X \) and \( Y \) are jointly continuous: this controls the singularity in the ratio that defines the conditional probability.

**Remark:** If \( X \) and \( Y \) are independent, then the conditional probability density of \( X \) given \( Y = y \) is equal to the unconditioned probability density:

\[
p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(y)} = p_X(x).
\]
Example 6.6. (Ross, Example 5c) We say that a random vector \((X,Y)\) has a \textbf{bivariate normal distribution} if there exist parameters \((\mu_x, \mu_y)\) and \(\sigma_x > 0, \sigma_y > 0, -1 < \rho < 1\) such that the joint density of \(X\) and \(Y\) is

\[
p_{X,Y}(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x - \mu_x}{\sigma_x} \right)^2 + \left( \frac{y - \mu_y}{\sigma_y} \right)^2 - 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y} \right] \right\}.
\]

The conditional density of \(X\) given that \(Y = y\) is equal to

\[
p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} = C_1 p_{X,Y}(x, y)
\]

\[
= C_2 \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x - \mu_x}{\sigma_x} \right)^2 - 2\rho \frac{x(y - \mu_y)}{\sigma_x \sigma_y} \right] \right\}
\]

\[
= C_3 \exp \left\{ -\frac{1}{2\sigma_x^2(1-\rho^2)} \left[ x^2 - 2x \left( \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \right) \right] \right\}
\]

\[
= C_4 \exp \left\{ -\frac{1}{2\sigma_x^2(1-\rho^2)} \left[ x - \left( \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y) \right)^2 \right] \right\}.
\]

Apart from the constant \(C_4\), which can be evaluated explicitly by using the fact that \(\int_{-\infty}^{\infty} p_{X|Y}(x|y)dx = 1\), we recognize the conditional density as the density of a normal RV with

- \textbf{mean} = \(\mu_x + \rho(\sigma_x/\sigma_y)(y - \mu_y)\)
- \textbf{variance} = \(\sigma_x^2(1-\rho^2)\).

Remark 6.1. Notice that \(X\) and \(Y\) are independent if and only if \(\rho = 0\).
Chapter 7

Expectations

7.1 Expectations of Sums

We begin with a two-dimensional version of the law of the unconscious statistician.

**Proposition 7.1.** If $X$ and $Y$ have a joint probability mass function $p_{X,Y}(x,y)$ and $g : \mathbb{R}^2 \to \mathbb{R}$, then

$$\mathbb{E}[g(X,Y)] = \sum_y \sum_x p_{X,Y}(x,y)g(x,y).$$

Similarly, if $X$ and $Y$ are jointly continuous with joint density $p_{X,Y}(x,y)$, then

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{X,Y}(x,y)g(x,y)dx dy.$$ 

**Proof.** By expressing $g$ as the difference between its positive and negative parts, $g = g^+ - g^-$, it suffices to prove the result under the assumption that $g$ is non-negative. Here we will assume that $X$ and $Y$ are jointly continuous; a similar proof works in the case where both are discrete. Then, by Lemma 5.1, we have

$$\mathbb{E}[g(X,Y)] = \int_0^\infty \mathbb{P}\{g(X,Y) > t\}dt = \int_0^\infty \int_{(x,y) : g(x,y) > t} p_{X,Y}(x,y)dx dy dt = \int_{-\infty}^\infty \int_{-\infty}^\infty g(x,y)dx dy dt = \int_{-\infty}^\infty \int_{-\infty}^\infty p_{X,Y}(x,y)g(x,y)dx dy,$$

where the interchange of integrals in passing from the second to the third line is justified by the fact that the integrand is everywhere non-negative.

□

**Corollary 7.1.** Suppose that $X_1, X_2, \cdots, X_n$ are random variables such that $\mathbb{E}[X_i]$ is finite for all $i = 1, \cdots, n$. Then

$$\mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$
7.1. EXPECTATIONS OF SUMS

Proof. If the \( X_i \) are all discrete or all jointly continuous, the corollary follows from Proposition 7.1 by taking \( g(x, y) = x + y \) and then proceeding by induction on \( n \).

Corollary 7.2. Suppose that \( X \) and \( Y \) are random variables such that \( \mathbb{P}\{X \geq Y\} = 1 \). Then \( \mathbb{E}[X] \geq \mathbb{E}[Y] \).

Proof. We will prove this under the assumption that \( X \) and \( Y \) are both discrete. The assumption that \( \mathbb{P}\{X \geq Y\} = 1 \) implies that \( p_{X,Y}(x, y) = 0 \) whenever \( x < y \). Consequently,

\[
\mathbb{E}[X] - \mathbb{E}[Y] = \mathbb{E}[X - Y] = \sum_{(x,y)} p_{X,Y}(x, y)(x - y) \geq 0.
\]

Example 7.1. (Ross, Example 2a) Suppose that \( X \) and \( Y \) are independent RVs that are uniformly distributed on \([0,1]\). Then

\[
\mathbb{E}[|X - Y|] = \int_0^1 \int_0^1 |x - y| dy \, dx
\]

\[
= \int_0^1 \left( \int_0^x (x - y) \, dy + \int_x^1 (y - x) \, dy \right) \, dx
\]

\[
= \int_0^1 \left( x^2 - x^2/2 + (1 - x^2)/2 - x(1 - x) \right) \, dx
\]

\[
= \int_0^1 \left( x^2 - x + \frac{1}{2} \right) \, dx
\]

\[
= \frac{1}{3}.
\]

Example 7.2. (Ross, Example 2c) Suppose that \( X_1, \ldots, X_n \) are independent, identically-distributed random variables, each with expected value \( \mu \). Then the sample mean of the \( X_i \)'s is just the unweighted average of the sample:

\[
\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.
\]

Notice that

\[
\mathbb{E}[\bar{X}] = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i]
\]

\[
= \frac{1}{n} \cdot n \mu = \mu,
\]

i.e., the expected value of the sample mean is equal to the mean of the distribution. For this reason, the sample mean is said to be an unbiased estimator of the expected value.
Example 7.3. **Expectation of a binomial RV:** Recall that if \( X_1, \ldots, X_n \) are independent Bernoulli RVs, each with parameter \( p \), then their sum \( X = X_1 \cdots X_n \) is a Binomial RV with parameters \( n \) and \( p \). A simple calculation shows that \( \mathbb{E}[X_i] = p \cdot 1 + (1 - p) \cdot 0 = p \) for each \( i \). Consequently,

\[
\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np,
\]

as we showed earlier using probability generating functions.

It is important to realize that linearity of expectations holds even when the summands are not independent.

Example 7.4. *(Ross, Example 2h)* Let \( X \) be the number of cards that are left in their original position by a random permutation of a deck of \( n \) cards. Then \( X \) can be written as the sum \( X = X_1 + \cdots + X_n \), where \( X_i \) is the indicator function of the event that the \( i \)’th card is left in its original position. Notice that each \( X_i \) is a Bernoulli RV with parameter \( p = \frac{1}{n} \), but the \( X_i \)’s are not independent. Nonetheless,

\[
\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot \frac{1}{n} = 1.
\]

Example 7.5. **Example: Random Walks on \( \mathbb{Z} \).** Suppose that \( X_1, X_2, \ldots \) is a collection of independent RVs, each with distribution \( \mathbb{P}\{X_i = 1\} = \mathbb{P}\{X_i = -1\} = \frac{1}{2} \), and let \( S_N = X_1 + \cdots + X_N \). We can think of \( S_N \) as the position at time \( N \) of a particle which in each unit interval of time is equally likely to move to the right or to the left. Notice that

\[
\mathbb{E}[S_N] = \sum_{i=1}^{N} \mathbb{E}[X_i] = 0,
\]

since \( \mathbb{E}[X_i] = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 0 \) for each \( i \). Thus, the expected position of the particle never changes. A better measure of the displacement of the particle away from its initial position at 0 is given by the root mean square displacement (RMSQ) of \( S_N \)

\[
D_N \equiv \sqrt{\mathbb{E}[S_N^2]}.
\]

To find \( D_N \), we begin by calculating

\[
D_N^2 = \mathbb{E}[S_N^2] = \mathbb{E}[(X_1 + \cdots + X_N)^2] = \mathbb{E} \left[ \sum_{i=1}^{N} X_i^2 \right] + \mathbb{E} \left[ \sum_{1 \leq i \neq j \leq N} X_i X_j \right] = N \mathbb{E}[X_1^2] + N(N-1) \mathbb{E}[X_1 X_2],
\]

where to obtain the last line, we have used the identities \( \mathbb{E}[X_i^2] = \mathbb{E}[X_i^2] \) and \( \mathbb{E}[X_i X_j] = \mathbb{E}[X_1 X_2] \) for all \( i \) and \( j \) such that \( i \neq j \). Furthermore,

\[
\mathbb{E}[X_1^2] = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot (-1)^2 = 1,
\]

\[
\mathbb{E}[X_1 X_2] = \frac{1}{4} \cdot 1 \cdot 1 + \frac{1}{4} \cdot 1 \cdot (-1) + \frac{1}{4} \cdot (-1) \cdot 1 + \frac{1}{4} \cdot (-1) \cdot (-1) = 0.
\]
Substituting these values into the previous equation shows that \( D^2_N = N \) and so the RMSD of the particle is

\[
D_N = N^{1/2}.
\]

Notice that \( D_N \) diverges as \( N \to \infty \), but at a rate that is sublinear in \( N \).

Example 7.6. (Ross, Example 2n: The Inclusion-Exclusion Formula Revisited) Let \( A_1, \ldots, A_N \) be a collection of events and let \( X_i \) be the indicator variable of \( A_i \), i.e., \( X_i = 1 \) if \( A_i \) occurs and \( X_i = 0 \) otherwise. Notice that

\[
1 - \prod_{i=1}^{n} (1 - X_i) = \begin{cases} 
1 & \text{if } \bigcup_{i=1}^{n} A_i \text{ occurs} \\
0 & \text{otherwise,}
\end{cases}
\]

and so

\[
\mathbb{E} \left[ 1 - \prod_{i=1}^{n} (1 - X_i) \right] = \mathbb{P} \left( \bigcup_{i=1}^{n} A_i \right).
\]

Expanding the polynomial on the left-hand side of the preceding formula yields

\[
\mathbb{P} \left( \bigcup_{i=1}^{n} A_i \right) = \mathbb{E} \left[ \sum_{i=1}^{n} X_i - \sum_{i<j} X_iX_j + \cdots + (-1)^{n+1}X_1X_2\cdots X_n \right] = \sum_{i=1}^{n} \mathbb{E}[X_i] - \sum_{i<j} \mathbb{E}[X_iX_j] + \cdots + (-1)^{n+1} \mathbb{E}[X_1X_2\cdots X_n].
\]

Because

\[
X_{i_1}X_{i_2}\cdots X_{i_k} = \begin{cases} 
1 & \text{if } \bigcap_{j=1}^{k} A_{i_j} \text{ occurs} \\
0 & \text{otherwise,}
\end{cases}
\]

each term in this expression can written as the probability of an intersection of events:

\[
\mathbb{E}[X_{i_1}X_{i_2}\cdots X_{i_k}] = \mathbb{P} \left( \bigcap_{j=1}^{k} A_{i_j} \right).
\]

Upon substituting these formulas into the above expression, we obtain the inclusion-exclusion formula for the probability of a union of events:

\[
\mathbb{P} \left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1} \mathbb{P} \left( \bigcap_{i=1}^{n} A_i \right).
\]

7.2 Variance and Covariance

Proposition 7.2. If \( X \) and \( Y \) are independent, then for any pair of real-valued functions \( h \) and \( g \),

\[
\mathbb{E}[h(X)g(Y)] = \mathbb{E}[h(X)] \cdot \mathbb{E}[g(Y)],
\]

provided that the expectations appearing on both sides of the identity exist.
Proof. We will prove this under the assumption that $X$ and $Y$ are jointly continuous with joint density $p_{X,Y} = p_X \cdot p_Y$. Then

$$\mathbb{E}[h(X)g(Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)g(y)p_{X,Y}(x,y)dxdy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)g(y)p_X(x)p_Y(y)dxdy$$

$$= \int_{-\infty}^{\infty} p_X(x)h(x)dx \int_{-\infty}^{\infty} p_Y(y)g(y)dy$$

$$= \mathbb{E}[h(X)] \cdot \mathbb{E}[g(Y)].$$

Corollary 7.3. Taking $h(x) = x$ and $g(y) = y$ in Proposition 7.2 yields the important identity:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y],$$

provided that $X$, $Y$ and $XY$ all have finite expectations.

Corollary 7.4. If $X_1, \ldots, X_n$ are independent RVs, then a simple induction argument on $n$ shows that

$$\mathbb{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}[f_i(X_i)],$$

again provided that all of the expectations appearing in the formula exist.

Definition 7.1. If $X$ and $Y$ are RVs with means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, then the covariance between $X$ and $Y$ is the quantity

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

$$= \mathbb{E}[XY] - \mu_X \cdot \mu_Y.$$

Remark: The covariance of two random variables is a measure of their association. The covariance is positive if when $X$ is larger than $\mu_X$, then $Y$ also tends to be larger than $\mu_Y$, while the covariance is negative if the opposite pattern holds. In particular, if $X$ and $Y$ are independent, then

$$\text{Cov}(X,Y) = \mathbb{E}[X - \mu_X]\mathbb{E}[Y - \mu_Y] = 0,$$

because knowledge of how $X$ compares with its mean provides no information about $Y$ or vice versa. Observe that the converse is not true. There are random variables that have zero covariance but which are not independent. For example, suppose that $X$ has the distribution

$$\mathbb{P}\{X = 0\} = \mathbb{P}\{X = 1\} = \mathbb{P}\{X = -1\} = \frac{1}{3},$$

and that the random variable $Y$ is defined by

$$Y = \begin{cases} 
0 & \text{if } X \neq 0 \\
1 & \text{if } X = 0.
\end{cases}$$

Then $\mathbb{E}[X] = 0$ and also $\mathbb{E}[XY] = 0$ because $XY = 0$ with probability 1, so that

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0,$$

because knowledge of how $X$ compares with its mean provides no information about $Y$ or vice versa. Observe that the converse is not true. There are random variables that have zero covariance but which are not independent. For example, suppose that $X$ has the distribution

$$\mathbb{P}\{X = 0\} = \mathbb{P}\{X = 1\} = \mathbb{P}\{X = -1\} = \frac{1}{3},$$

and that the random variable $Y$ is defined by

$$Y = \begin{cases} 
0 & \text{if } X \neq 0 \\
1 & \text{if } X = 0.
\end{cases}$$

Then $\mathbb{E}[X] = 0$ and also $\mathbb{E}[XY] = 0$ because $XY = 0$ with probability 1, so that

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0,$$
but $X$ and $Y$ clearly are not independent. For example,

$$P\{X = 0|Y = 1\} = 1 \neq \frac{1}{3} = P\{X = 0\}.$$ 

Any two random variables $X$ and $Y$ are said to be uncorrelated if $\text{Cov}(X, Y) = 0$, regardless of whether or not they are independent.

The next proposition lists some useful properties of covariances.

**Proposition 7.3.**

(a) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

(b) $\text{Cov}(X, X) = \text{Var}(X)$

(c) $\text{Cov}(aX, Y) = a\text{Cov}(X, Y)$ for any scalar $a$

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

Each identity can be proved directly from the definition of the covariance and the linearity properties of expectations. (See Ross, pp. 323-324.)

A very useful formula follows by combining parts (b) and (d) of the preceding proposition:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{1 \leq i \neq j \leq n} \text{Cov}(X_i, X_j).$$

In particular, if the $X_i$’s are pairwise independent, then $\text{Cov}(X_i, X_j) = 0$ whenever $i \neq j$, so that

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).$$

If, in addition to being pairwise independent, the $X_i$’s are identically distributed, say with variance $\text{Var}(X_i) = \sigma^2$, then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2.$$ 

**Example 7.7. (Ross, Example 4b)** If $X_1, \cdots, X_n$ are independent Bernoulli RVs, each with parameter $p$, then $X = X_1 + \cdots + X_n$ is a Binomial RV with parameters $(n, p)$. Because the variance of each Bernoulli RV is $p(1-p)$, the preceding formula for the variance of a sum of IID RVs shows that

$$\text{Var}(X) = np(1-p),$$

confirming our previous calculation using probability generating functions.

**Example 7.8. (Ross, Example 4a)** If $X_1, \cdots, X_n$ is a collection of IID (independent, identically-distributed) random variables with mean $\mu$ and variance $\sigma^2$, then the sample mean $\bar{X}$ and the
**Sample Variance** $S^2$ are the quantities defined by

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

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$  

If we think of the $X_i$’s as the outcomes of a sequence of independent replicates of some experiment, then the sample mean and sample variance are often used to estimate the true mean and true variance of the underlying distribution. Recall that in the previous section we showed that the sample mean is an unbiased estimator for the true mean:

$$E[\bar{X}] = \mu.$$  

Although unbiasedness is a desirable property of an estimator, we also care about the variance of the estimator about its expectation. For the sample mean, we can calculate

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n},$$

which shows that the variance of the sample mean is inversely proportional to the number of independent replicates that have been performed. We can also show that the sample variance is an unbiased estimator of the true variance. We first observe that

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + \sum_{i=1}^{n} (\bar{X} - \mu)^2\right)$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu)^2 - \frac{n}{n-1}(\bar{X} - \mu)^2.$$  

Taking expectations gives

$$E[S^2] = \frac{1}{n-1} \sum_{i=1}^{n} E[(X_i - \mu)^2] - \frac{n}{n-1} Var(\bar{X})$$

$$= \frac{n}{n-1} \sigma^2 - \frac{n}{n-1} \frac{\sigma^2}{n}$$

$$= \sigma^2.$$  

Informally, the reason that we must divide by $n-1$ rather than $n$ to obtain an unbiased estimate of the variance is that the deviations between the $X_i$’s and the sample mean (which is computed using the observed data) tend to be smaller than the deviations between the $X_i$’s and the true mean.

The next definition provides us with a **dimensionless** measure of the association between two random variables.
Definition 7.2. The correlation of two random variables $X$ and $Y$ with variances $\sigma_X^2 > 0$ and $\sigma_Y^2 > 0$ is the quantity

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

Because

$$\rho(aX,bY) = \frac{ab \cdot \text{Cov}(X,Y)}{a \sigma_X \cdot b \sigma_Y} = \rho(X,Y),$$

for any pair of scalar quantities $a, b \neq 0$, the correlation between two random variables does not depend on the units in which they are measured, i.e., the correlation is dimensionless. Another important property of the correlation is that

$$-1 \leq \rho(X,Y) \leq 1. \quad (7.1)$$

This can be proved as follows. First note that

$$0 \leq \text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = \frac{\text{Var}(X)}{\sigma_X^2} + \frac{\text{Var}(Y)}{\sigma_Y^2} + 2\frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = 2(1 + \rho(X,Y)),$$

which implies that $\rho(X,Y) \geq -1$. A similar calculation involving the variance of the difference of $X/\sigma_X$ and $Y/\sigma_Y$ shows that $\rho(X,Y) \leq 1$. If we interpret the standard deviation of a random variable as a measure of its length or norm, then (7.1) can be interpreted as a version of the Cauchy-Schwartz inequality.

It can be shown that if $\rho(X,Y) = 1$, then with probability one such that

$$Y = \mu_Y + \frac{\sigma_Y}{\sigma_X} (X - \mu_X),$$

while if $\rho(X,Y) = -1$, then

$$Y = \mu_Y - \frac{\sigma_Y}{\sigma_X} (X - \mu_X).$$

Thus the correlation constant can also be treated as a measure of the collinearity of two random variables.

7.3 Conditional Expectation

Definition 7.3. Recall that, if $X$ and $Y$ are jointly discrete random variables, then the conditional probability mass function of $X$ given $Y = y$ is defined to be

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

provided $p_Y(y) > 0$. In this case, we can define the conditional expectation of $X$ given $Y = y$ by

$$\mathbb{E}[X|Y = y] = \sum_x x \cdot \mathbb{P}\{X = x|Y = y\} = \sum_x p_{X|Y}(x|y) \cdot x.$$ 

Similarly, if $X$ and $Y$ are jointly continuous with conditional density function $p_{X|Y}(x|y)$, then the conditional expectation of $X$ given $Y = y$ is defined by

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x \cdot p_{X|Y}(x|y) dx.$$
Example 7.9. Suppose that $X$ and $Z$ are independent Poisson RVs, both with parameter $\lambda$, and let $Y = X + Z$. We wish to find the conditional expectation of $X$ given that $Y = n$. To do so, we first observe that in light of Example 7.4, the conditional distribution of $X$ given that $Y = n$ is binomial with parameters $(n, 1/2)$:

$$p_{X|Y}(k|n) = \binom{n}{k} \left(\frac{1}{2}\right)^n,$$

for $k = 0, \cdots, n$. Consequently,

$$E[X|Y = n] = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^n k = \frac{n}{2}.$$

We can also solve this problem by observing that

$$n = E[Y|Y = n] = E[X|Y = n] + E[Z|Y = n] = 2E[X|Y = n],$$

using the fact that $E[X|Y = n] = E[Z|Y = n]$ since $X$ and $Y$ have the same conditional distribution given $Z$.

Remark: In general, all of the results that hold for expectations also hold for conditional expectations, including

$$E[g(X)|Y = y] = \begin{cases} \sum_x g(x)p_{X|Y}(x|y) & \text{discrete case} \\ \int_{-\infty}^{\infty} g(x)p_{X|Y}(x|y)dx & \text{continuous case}, \end{cases}$$

and

$$E \left[ \sum_{i=1}^{n} X_i | Y = y \right] = \sum_{i=1}^{n} E[X_i|Y = y].$$

The next definition introduces a subtle concept, but we will see that it leads to a powerful tool for calculating expected values. This is particularly important in the theory of stochastic processes and Markov chains.

Definition 7.4. The conditional expectation of $X$ given $Y$ is the random variable

$$E[X|Y] = H(Y),$$

where $H(y)$ is the (deterministic) function defined by the formula

$$H(y) = E[X|Y = y].$$

Perhaps the most important application of conditional expectations is described in the next proposition.

Proposition 7.4. For any two random variables $X$ and $Y$, we have

$$E[X] = E[E[X|Y]],$$
provided that both $E[X]$ and $E[X|Y]$ exist. In particular, if $Y$ is discrete, then

$$E[X] = \sum_y E[X|Y = y]P\{Y = y\},$$

while if $Y$ is continuous with density $p_Y(y)$, then

$$E[X] = \int_{-\infty}^{\infty} E[X|Y = y]p_Y(y)dy.$$

Proof. We will assume that $X$ and $Y$ are both discrete. Then

$$\sum_y E[X|Y = y]P\{Y = y\} = \sum_y \sum_x xP\{X = x, Y = y\}P\{Y = y\} = \sum_y \sum_x xP\{X = x, Y = y\}P\{Y = y\} = \sum_x xP\{X = x\} = E[X].$$

Example 7.10. (Wald’s Identity) Let $N$ be a RV with values in the natural numbers and finite mean $E[N] < \infty$, and let $X_1, X_2, \ldots$ be a sequence of independent, identically-distributed random variables that are independent of $N$ and have finite mean $E[X]$. Our goal is to calculate the expected value of the random sum $X_1 + \cdots + X_N$, which we can do by conditioning on $N$:

$$E \left[ \sum_{i=1}^{N} X_i \right] = E \left[ E \left[ \sum_{i=1}^{N} X_i | N \right] \right].$$

Notice that

$$E \left[ \sum_{i=1}^{N} X_i | N = n \right] = E \left[ \sum_{i=1}^{n} X_i | N = n \right] = E \left[ \sum_{i=1}^{n} X_i \right] \quad \text{since } X_i \text{ and } N \text{ are independent} = nE[X],$$

which implies that

$$E \left[ \sum_{i=1}^{N} X_i | N \right] = NE[X]$$

and consequently

$$E \left[ \sum_{i=1}^{N} X_i \right] = E[N \cdot E[X]] = E[N] \cdot E[X].$$
Example 7.11. Let $\xi_{i,n}, i \geq 1, n \geq 0$ be a collection of I.I.D. non-negative integer-valued random variables with finite mean $\mu$ and define the variables $X_0, X_1, \cdots$ recursively by setting $X_0 = 1$ and

$$X_{n+1} = \sum_{i=1}^{X_n} \xi_{i,n}$$

for every $n \geq 0$. The sequence of random variables $(X_n : n \geq 0)$ is said to be a Galton-Watson process with offspring distribution $\xi$. We can interpret $X_n$ as the number of adults alive in the $n$th generation of a population in which individuals reproduce independently of one another and $\xi_{i,n}$ is the number of (surviving) offspring born to the $i$th individual alive in that generation. Wald’s identity can be used to calculate the expected number of individuals alive in generation $n$:

$$E[X_n] = \mu \cdot E[X_{n-1}] = \mu^2 \cdot E[X_{n-2}] \cdots = \mu^n,$$

where the last identity uses the fact that $E[X_0] = 1$. If $\mu < 1$, then $\mu^n \to 0$ as $n \to \infty$ and it can be shown that the population is certain to go extinct in finite time. In this case, the process is said to be subcritical. In contrast, if $\mu > 1$, then the process is said to be supercritical and it can be shown that there is a positive probability (not necessarily equal to 1) that the population will persist indefinitely.

Example 7.12. A discrete random variable $X$ is said to have a compound Poisson distribution if there exists a Poisson RV $N$ with parameter $\lambda > 0$ and a set of IID random variables $Y_1, Y_2, \cdots$, all independent of $N$, such that

$$X = \sum_{i=1}^{N} Y_i.$$ 

If $E[Y_i] = \mu_y < \infty$, then by Wald’s identity we have

$$E[X] = \lambda \cdot \mu_y.$$ 

Conditioning is also sometimes helpful when calculating probabilities. Suppose that $E$ is some event and let $X$ be the indicator function of $E$:

$$X = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

Then, for any random variable $Y$ defined on the same probability space as $X$,

$$E[X] = P(E) \quad \quad \quad E[X|Y=y] = P(E|Y=y).$$

Furthermore, by using Proposition 5.1, we have

$$P(E) = \begin{cases} \sum_y P(E|Y=y)P(Y=y) & \text{if } Y \text{ is discrete} \\ \int_{-\infty}^{\infty} P(E|Y=y)p_Y(y)dy & \text{if } Y \text{ is continuous.} \end{cases}$$

Example 7.13. (Ross, Example 5I) Let $U$ be a uniform RV on $(0,1)$, and suppose that the conditional distribution of $X$, given $U = p$, is binomial with parameters $(n,p)$. Then the
probability mass function of $X$ can be calculated by conditioning on $p$:
\[
P\{X = k\} = \mathbb{E}[P\{X = k|U = p\}]
= \int_0^1 P\{X = k|U = p\}dp
= \int_0^1 \binom{n}{k} p^k(1-p)^{n-k}dp
= \frac{n!}{k!(n-k)!} \int_0^1 p^k(1-p)^{n-k}dp.
\]
Using the Beta function, it can be shown that the value of the integral in this last line is
\[
\int_0^1 p^k(1-p)^{n-k}dp = \beta(k+1, n-k+1) = \frac{k!(n-k)!}{(n+1)!},
\]
so upon substituting this expression back into the preceding series of equations, we obtain:
\[
P\{X = k\} = \frac{1}{n+1},
\]
i.e., $X$ is uniformly distributed on the set $\{0, 1, \cdots, n\}$. $X$ is an example of a mixture model: the distribution of $X$ is equal to the mixture of a family of probability distributions with respect to another distribution called the mixing distribution. These play an important role in statistics.

### 7.3.1 Conditional Variance

**Definition 7.5.** If $X$ and $Y$ are random variables defined on the same probability space, then the **conditional variance** of $X$ given that $Y = y$ is the quantity
\[
\text{Var}(X|Y = y) \equiv \mathbb{E}[(X - \mathbb{E}[X|Y = y])^2|Y = y].
\]
Similarly, the conditional variance of $X$ given $Y$ is the random variable
\[
\text{Var}(X|Y) \equiv \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y].
\]
As the following proposition shows, in general, the expected conditional variance of a random variable is less than its variance.

**Proposition 7.5.** *Law of Total Variance*

\[
\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y]).
\]

**Proof.** First observe that
\[
\text{Var}(X|Y) = \mathbb{E}[X^2|Y] - \left(\mathbb{E}[X|Y]\right)^2,
\]
and so
\[
\mathbb{E}[\text{Var}(X|Y)] = \mathbb{E}[\mathbb{E}[X^2|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2]
= \mathbb{E}[X^2] - \mathbb{E}[(\mathbb{E}[X|Y])^2].
\]
Also, because $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$, the variance of the conditional expectation of $X$ given $Y$ can be written as
\[
\text{Var}(\mathbb{E}[X|Y]) = \mathbb{E}[(\mathbb{E}[X|Y])^2] - (\mathbb{E}[X])^2.
\]
The law of total variance follows upon adding these two equations.
Example 7.14. (Ross, Example 5p) Let $N$ and $X_1, X_2, \ldots$ be as in Example (5d), with the added assumption that $\text{Var}(N) < \infty$ and $\text{Var}(X) < \infty$. Then

\[
\mathbb{E} \left[ \sum_{i=1}^{N} X_i \mid N \right] = N \cdot \mathbb{E}[X] \\
\text{Var} \left( \sum_{i=1}^{N} X_i \mid N \right) = N \cdot \text{Var}(X),
\]

and so the law of total variance implies that

\[
\text{Var} \left( \sum_{i=1}^{N} X_i \right) = \mathbb{E}[N] \cdot \text{Var}(X) + (\mathbb{E}[X])^2 \cdot \text{Var}(N).
\]

7.4 Moment Generating Functions

Earlier, we defined the probability generating function of a non-negative integer-valued random variable $X$ to be the function

\[
\phi_X(s) = \mathbb{E} \left[ s^X \right] = \sum_{n=0}^{\infty} \mathbb{P}\{X = n\} s^n.
\]

In this section, we will introduce a different kind of generating function which is defined for a larger class of random variables.

Definition 7.6. If $X$ is a real-valued random variable, then the moment generating function of $X$ is the function $\psi_X : \mathbb{R} \rightarrow [0, \infty]$ defined by the formula

\[
\psi_X(t) = \mathbb{E} \left[ e^{tX} \right] = \sum_{n=0}^{\infty} \frac{M_n}{n!} t^n,
\]

where $M_n = \mathbb{E}[X^n]$ is the $n$'th moment of $X$.

Remarks:

1. Although the moment generating function is defined for all $t \in \mathbb{R}$, it may be infinite for some of these values (cf. Example 7.15).
2. The moments of $X$ can be computed from $\phi_X(t)$ by differentiation:

\[
M_n = \phi_X^{(n)}(0).
\]

3. A distribution is uniquely determined by its probability generating function whenever this function is defined on some open interval containing 0.
4. If $X \in \{0, 1, 2, \ldots\}$, then the probability generating function and the moment generating function of $X$ are related by the simple identity

\[
\phi_X(t) = \psi_X(\ln(t)),
\]

for $t \in (0, 1]$. 

Example 7.15. (Ross, Example 7b) If $X$ is a Poisson RV with parameter $\lambda$, then the moment generating function of $X$ is

$$
\psi_X(t) = E[e^{tX}] = \sum_{n=0}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} e^{tn} = \exp(-\lambda(1-e^t)).
$$

Example 7.16. (Ross, Example 7c) If $X$ is an exponential RV with parameter $\lambda$, then

$$
\phi_X(t) = E[e^{tX}] = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda-t} \quad \text{for } t < \lambda
$$

$$
= \sum_{n=0}^{\infty} \lambda^{-n} t^n.
$$

Thus, the $n$'th moment of this distribution is simply $M_n = \lambda^{-n} n!$.

Example 7.17. (Ross, Example 7d) If $Z$ is a standard normal RV (with mean 0 and variance 1), then

$$
M_Z(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx
$$

$$
= e^{t^2/2} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx
$$

$$
= e^{t^2/2} \sum_{n=0}^{\infty} \frac{1}{2^n n!} t^{2n}.
$$

Thus, the moments of $Z$ are

$$
M_{2n} = \frac{(2n)!}{2^n n!}
$$

$$
M_{2n+1} = 0
$$

for all $n \geq 0$. To calculate the moment generating function of a normal RV with mean $\mu$ and $\sigma$, recall that $Z \equiv (X - \mu)/\sigma)$ is a standard normal RV. Then, writing $X = \sigma Z + \mu$, we have

$$
\psi_X(t) = E[e^{tX}] = E[e^{t(\sigma Z + \mu)}]
$$

$$
= e^{t \mu} \psi_Z(\sigma t)
$$

$$
= \exp\left(\frac{\sigma^2 t^2}{2} + \mu t\right).
$$

Our next result shows that moment generating functions behave particularly nicely when applied to sums of independent random variables. This is arguably their most important property.
CHAPTER 7. EXPECTATIONS

Proposition 7.6. Suppose that $X_1, \ldots, X_n$ are independent RVs with moment generating functions $\psi_{X_i}(t), \ldots, \psi_{X_n}(t)$. Then the moment generating function of the sum $X = X_1 + \cdots + X_n$ is

$$
\psi_X(t) = \mathbb{E} \left[ e^{t(X_1 + \cdots + X_n)} \right] = \prod_{i=1}^n \mathbb{E} \left[ e^{tX_i} \right] = \prod_{i=1}^n \psi_{X_i}(t),
$$

i.e., the moment generating function of a sum of independent RVs is just the product of the moment generating functions of the individual RVs.

Example 7.18. Suppose that $X_1, \ldots, X_n$ are independent Poisson RVs with parameters $\lambda_1, \ldots, \lambda_n$, respectively. Then the moment generating function of the sum $X = X_1 + \cdots + X_n$ is

$$
\psi_X(t) = \prod_{i=1}^n e^{-\lambda_i (1 - e^t)} = \exp \left( -(1 - e^t) \sum_{i=1}^n \lambda_i \right),
$$

which confirms our earlier finding that $X$ is a Poisson RV with parameter $\sum_{i=1}^n \lambda_i$.

Example 7.19. Suppose that $X_1, \ldots, X_n$ are independent normal RVs with parameters $(\mu_1, \sigma^2_1), \ldots, (\mu_n, \sigma^2_n)$, respectively. Then the moment generating function of the sum $X = X_1 + \cdots + X_n$ is

$$
\psi_X(t) = \prod_{i=1}^n \exp \left( \frac{\sigma^2_i t^2}{2} + \mu_i t \right) = \exp \left( \frac{t^2}{2} \sum_{i=1}^n \sigma^2_i + t \sum_{i=1}^n \mu_i \right),
$$

which confirms our earlier finding that $X$ is a normal RV with mean $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma^2_i$.

Example 7.20. (Ross, Example 7j) Suppose that $N$ is a non-negative integer-valued RV with moment generating function $\psi_N(t)$ and that $X_1, X_2, \cdots$ are IID random variables that are independent of $N$, each with moment generating function $\psi_X(t)$. Our goal is to calculate the moment generating function of the random sum $Z = X_1 + \cdots + X_N$. To do so, we first condition on $N$:

$$
\mathbb{E} \left[ \exp \left\{ t \sum_{i=1}^N X_i \right\} \middle| N = n \right] = \mathbb{E} \left[ \exp \left\{ t \sum_{i=1}^n X_i \right\} \middle| N = n \right] = \mathbb{E} \left[ \exp \left\{ t \sum_{i=1}^n X_i \right\} \right] = (\psi_X(t))^n,
$$
where the second line exploits the independence of \( N \) from the \( X_i \). Consequently,

\[
\psi_Z(t) = \mathbb{E}[e^{tX}]
= \mathbb{E} \left[ \mathbb{E}[e^{tX} | N] \right]
= \mathbb{E} \left[ (\psi_X(t))^N \right]
= \mathbb{E} \left[ \exp \left( N \cdot \ln(\psi_X(t)) \right) \right]
= \psi_N \left( \ln \left( \psi_X(t) \right) \right).
\]

**Definition 7.7.** If \( X_1, \cdots, X_n \) are random variables defined on the same sample space, then their joint moment generating function is the function \( \psi : \mathbb{R}^n \to \mathbb{R} \) defined by

\[
\psi(t_1, \cdots, t_n) = \mathbb{E} \left[ e^{t_1 X_1 + \cdots + t_n X_n} \right].
\]

As in the univariate case, it can be shown that the moment generating function uniquely determines the joint distribution of the RVs \( X_1, \cdots, X_n \) provided that this function is finite on some open disk around the origin \((0, \cdots, 0)\).

The following proposition describes a powerful method for verifying independence.

**Proposition 7.7.** Suppose that the joint moment generating function of the random variables \( X_1, \cdots, X_n \) is finite in some open disk. Then these RVs are independent if and only if

\[
\psi(t_1, \cdots, t_n) = \psi_{X_1}(t_1) \cdots \psi_{X_n}(t_n)
\]

for all \((t_1, \cdots, t_n)\) in the domain of \( \psi \).

**Example 7.21.** *(Ross, Example 7I)* Suppose that \( X \) and \( Y \) are independent standard normal RVs and define \( W = X + Y \) and \( Z = X - Y \). Then the joint moment generating function of \( W \) and \( Z \) is

\[
\mathbb{E} \left[ e^{tW+sZ} \right] = \mathbb{E} \left[ e^{t(X+Y)+s(X-Y)} \right]
= \mathbb{E} \left[ e^{(t+s)X+(t-s)Y} \right]
= \mathbb{E} \left[ e^{(t+s)X} \right] \mathbb{E} \left[ e^{(t-s)Y} \right]
= e^{(t+s)^2/2} e^{(t-s)^2/2}
= e^{2t^2/2} e^{2s^2/2},
\]

which we recognize as the joint moment generating function of two independent normal RVs with mean 0 and variance 2. This shows that \( W \) and \( Z \) are independent.
Chapter 8

Limit Theorems for Sequences of Random Variables

8.1 The Weak Law of Large Numbers

Proposition 8.1. (Markov’s Inequality) If $X$ is a nonnegative real-valued RV, then for any positive real number $a > 0$,

$$
\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.
$$

Proof. Let $I_a$ be the indicator function of the event \{ $X \geq a$ \}

$$
I_a(\omega) = \begin{cases} 
1 & \text{if } X(\omega) \geq a \\
0 & \text{otherwise}.
\end{cases}
$$

Then, since $X \geq 0$, we have

$$
I_a \leq \frac{X}{a}.
$$

Taking expectations of both sides of this inequality gives

$$
\mathbb{P}(X \geq a) = \mathbb{E}[I_a] \leq \frac{\mathbb{E}[X]}{a}.
$$

Proposition 8.2. (Chebyshev’s Inequality) If $X$ is a RV with finite mean $\mu$ and finite variance $\sigma^2$, then for any positive real number $k > 0$,

$$
\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.
$$

Proof. The result follows by taking $a = k^2$ and applying Markov’s inequality to the non-negative random variable $(X - \mu)^2$

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

While Markov’s inequality provides a simple bound on the tail of the distribution of $X$, Chebyshev’s inequality allows us to estimate the probability that $X$ differs greatly from its mean. In general, both estimates tend to be crude in the sense that the probabilities in question are typically much smaller than the upper bounds given by these inequalities. However, as the following corollary and theorem will illustrate, even these crude inequalities can be used to prove some important results.
Corollary 8.1. If $\text{Var}(X) = 0$, then
\[ \mathbb{P}(X = \mathbb{E}[X]) = 1, \]
i.e., a random variable with zero variance is almost surely constant.

Proof. For each $n \geq 1$, Chebyshev's inequality implies that
\[ \mathbb{P}\left(|X - \mu| \geq \frac{1}{n}\right) = 0. \]
Since the sets $\{|X - \mu| \geq \frac{1}{n}\}$ are increasing as $n \to \infty$ and since
\[ \{X - \mu > 0\} = \bigcup_{n=1}^{\infty} \left\{|X - \mu| \geq \frac{1}{n}\right\}, \]
it follows from the continuity of probability measures that
\[ \mathbb{P}(X - \mu > 0) = \lim_{n \to \infty} \mathbb{P}\left(|X - \mu| \geq \frac{1}{n}\right) = 0. \]

Our next result, called the weak law of large numbers, is one of the celebrated limit theorems of probability theory. Before stating this theorem, we introduce one of the modes of convergence for a sequence of random variables. Other notions of convergence can also be defined and we will see one of these in the next section when we state the strong law of large numbers.

Definition 8.1. We say that a sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable $X$ if for every $\epsilon > 0$,
\[ \lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0. \]
Notice that the variables $X_1, X_2, \ldots$ and $X$ must all be defined on the same probability space for this definition to make sense.

Theorem 8.1. (The Weak Law of Large Numbers) Suppose that $X_1, X_2, \ldots$ is a sequence of I.I.D. variables with finite mean $\mu = \mathbb{E}[X_1]$ and let $S_n = X_1 + \cdots + X_n$. Then, for any $\epsilon > 0$,
\[ \lim_{n \to \infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0, \]
i.e., the sample averages $\frac{1}{n}S_n$ converge in probability to the mean.

Proof. We will prove the theorem under the additional assumption that $\text{Var}(X_i) = \sigma^2 < \infty$. In this case we have
\[ \mathbb{E}\left[\frac{1}{n}S_n\right] = \mu \quad \text{and} \quad \text{Var}\left(\frac{1}{n}S_n\right) = \frac{\sigma^2}{n}, \]
and so Chebyshev’s inequality implies that for any $\epsilon > 0$,
\[ \mathbb{P}\left(|S_n - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}, \]
which tends to 0 as $n \to \infty$. \qed
8.2 The Strong Law of Large Numbers

Theorem 8.2. (Strong Law of Large Numbers) Let \( X_1, X_2, \ldots \) be a sequence of I.I.D. random variables, each with finite mean \( \mu = \mathbb{E}[X] \), and let \( S_n = X_1 + \cdots + X_n \). Then,

\[
P \left( \lim_{n \to \infty} \frac{1}{n} S_n = \mu \right) = 1,
\]

i.e., the sequence \( \frac{1}{n} S_n \) converges almost surely to \( \mu \).

Proof. We will prove this result under the additional assumption that the \( X_i \) have a finite fourth moment \( K = \mathbb{E}[X_i^4] < \infty \). Without loss of generality, we may also assume that \( \mu = 0 \) (since otherwise we can simply de-mean the variables \( X_i \) by subtracting \( \mu \)). Then

\[
\mathbb{E}[S_n^4] = n \mathbb{E}[X_i^4] + 6 \binom{n}{2} \mathbb{E}[X_i^2 X_j^2] = nK + 3n(n-1)\sigma^4.
\]

Dividing both sides by \( n^4 \) gives

\[
\mathbb{E} \left[ \frac{S_n^4}{n^4} \right] = \frac{K}{n^3} + \left( 1 - \frac{1}{n} \right) \frac{3\sigma^4}{n^2},
\]

and so

\[
\mathbb{E} \left[ \sum_{n=1}^{\infty} \frac{S_n^4}{n^4} \right] < \infty.
\]

In particular, this last inequality implies that

\[
P \left( \sum_{n=1}^{\infty} \frac{S_n^4}{n^4} < \infty \right) = 1,
\]

since otherwise the expectation would be infinite. Furthermore, because a series converges only if its terms tend to 0, we can conclude that

\[
P \left( \lim_{n \to \infty} \frac{1}{n} S_n = 0 \right) = P \left( \lim_{n \to \infty} \frac{1}{n^4} S_n^4 = 0 \right) = 1,
\]

which completes the proof. \( \square \)

Remark: The significance of the strong law of large numbers is that it shows that the frequentist concept of probabilities can be derived using the measure-theoretic machinery introduced by Kolmogorov in the 1930’s. In particular, suppose that \((\Omega, \mathcal{F}, \mathbb{P})\) is a probability space and that \( E \in \mathcal{F} \) is an event, and let \( X_1, X_2, \ldots \) be independent indicator variables for \( E \):

\[
X_i = \begin{cases} 
1 & \text{if } E \text{ occurs on the } i\text{'th trial} \\
0 & \text{otherwise}
\end{cases}
\]

Then \( \mathbb{E}[X_i] = \mathbb{P}(E) \) and so the strong law tells us that

\[
P \left( \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{P}(E) \right) = 1,
\]

i.e., the limiting frequency of \( E \) is equal to its probability.
8.3 The Central Limit Theorem

Although the Strong Law of Large Numbers tells us that the sample means of a sequence of I.I.D. random variables converge almost surely to the expected value of the sampling distribution, it does not tell us anything about the rate of convergence or about the size of the fluctuations of the sample means about this limit. This information is important because we might like to know, for example, how likely it is that the sample mean of say 10,000 independent variates will differ from the true mean by more than 10% of the value of the latter. The Central Limit Theorem is one of several results that gives us some insight into the fluctuations of the sequence of sample means under the additional assumption that the sampling distribution has finite variance. Before stating this theorem, we begin by defining a new mode of convergence for a sequence of random variables.

**Definition 8.2.** Suppose that $F$ and $F_n, n \geq 1$ are cumulative distribution functions on $\mathbb{R}$. Then $F_n$ is said to converge weakly to $F$ if $\lim_{n \to \infty} F_n(x) = F(x)$ at every continuity point $x$ of $F$. Likewise, a sequence of random variables $X_n$ is said to converge in distribution to a random variable $X$ if the cumulative distribution functions $F_n(x) = \mathbb{P}(X_n \leq x)$ converge weakly to $F(x) = \mathbb{P}(X \leq x)$.

**Example 8.1.** To see why we only require pointwise convergence at continuity points, let $X_n$ and $X$ be defined by setting $\mathbb{P}(X_n = 1/n) = 1$ and $\mathbb{P}(X = 0) = 1$. Then

$$F_n(x) = \begin{cases} 0 & \text{if } x < 1/n \\ 1 & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{otherwise,} \end{cases}$$

which shows that $F_n(x)$ converges to $F(x)$ as $n \to \infty$ at every $x \neq 0$. Since $F(x)$ is continuous everywhere except at $x = 0$, it follows that the sequence $X_n$ converges in distribution to $X$ even though the sequence $F_n(0) = 0$ does not converge to $F(0) = 1$.

The next proposition shows how we can use moment generating functions to deduce that a sequence of random variables converges in distribution. The assumption that the m.g.f. of the limit $X$ is finite is essential.

**Proposition 8.3.** Let $X_1, X_2, \ldots$ be a sequence of random variables with moment generating functions $\psi_{X_n}(t)$, and let $X$ be a random variable with moment generating function $\psi_X$. Then the sequence $X_n$ converges in distribution to $X$ if

$$\lim_{n \to \infty} \psi_{X_n}(t) = \psi_X(t) < \infty$$

for all $t \in \mathbb{R}$.

We now come to the main result.

**Theorem 8.3. (The Central Limit Theorem)** Suppose that $X_1, X_2, \ldots$ is a sequence of I.I.D. variables with finite mean $\mu$ and variance $\sigma^2$, and let $S_n = X_1 + \cdots + X_n$. Then the sequence of random variables

$$Z_n \equiv \frac{S_n - n\mu}{\sigma \sqrt{n}} = \frac{1}{\sigma n^{1/2}} \left( \frac{1}{n} S_n - \mu \right)$$

converges in distribution to the standard normal $\mathcal{N}(0, 1)$, i.e., for every real number $z$,

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx.$$
Proof. We will prove the C.L.T. under the assumption that the moment generating function of the random variables \( X_i \), denoted \( \psi(t) \), is finite on the entire real line. Let us first assume that \( \mu = 0 \) and \( \sigma^2 = 1 \). Notice that the moment generating function of the scaled random variable \( X_i / \sqrt{n} \) is

\[
\mathbb{E} \left[ \exp \left\{ t \frac{X_i}{\sqrt{n}} \right\} \right] = \psi \left( \frac{t}{\sqrt{n}} \right)
\]

and that the moment generating function of the sum \( Z_n = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}} \) is equal to

\[
\left[ \psi \left( \frac{t}{\sqrt{n}} \right) \right]^n.
\]

Let

\[
L(t) = \log \psi(t)
\]

and observe that \( L(0) = 0 \) and \( L'(0) = 0 \) and \( L''(0) = 1 \). By Proposition 8.3, it suffices to show that

\[
\lim_{n \to \infty} \left[ \psi \left( \frac{t}{\sqrt{n}} \right) \right]^n = e^{t^2/2},
\]

which is equivalent to

\[
\lim_{n \to \infty} nL \left( \frac{t}{\sqrt{n}} \right) = \frac{t^2}{2}.
\]

However, this last identity can be verified using L'Hôpital’s rule

\[
\lim_{n \to \infty} \frac{L(t/\sqrt{n})}{n^{-1}} = \lim_{x \to 0} \frac{L(tx)}{x^2} = \lim_{x \to 0} \frac{tL'(tx)}{2x} = \lim_{x \to 0} \frac{t^2L''(tx)}{2} = \frac{t^2}{2}.
\]

The general case can then be handled by applying this result to the standardized variables \( X_i^* = (X_i - \mu) / \sigma \), which have mean 0 and variance 1.

\[
\square
\]

What is remarkable about the central limit theorem is that it shows that the distribution of a sum of sufficiently many I.I.D. random variables with finite variance is approximately normal no matter what distribution the individual variables have. This may explain why the normal distribution is encountered in so many seeming unrelated phenomena.

Remark 8.1. The restriction imposed by the use of the moment generating function in the proof of the C.L.T. can be surmounted by instead working with characteristic functions:

\[
\chi_X(t) = \mathbb{E} \left[ e^{itX} \right].
\]

Provided that \( X \) is real-valued, we have \( |e^{itX}| = 1 \) and so the modulus of the characteristic function is less than or equal to 1: \( |\chi_X(t)| \leq 1 \) for all \( t \in (-\infty, \infty) \). This means that the characteristic function of any real-valued random variable is finite on the entire real line. Furthermore, Proposition (8.3) still holds if we replace pointwise convergence of moment generating functions by pointwise convergence of characteristic functions.

Example 8.2. Let \( X_1, \ldots, X_{10} \) be independent random variables, each uniformly distributed on \([0,1]\), and let \( X = X_1 + \cdots + X_{10} \). The central limit theorem can be used to approximate the distribution of \( X \). For example,

\[
\mathbb{P} \{ X > 6 \} = \mathbb{P} \left\{ \frac{X - 5}{\sqrt{10/12}} > \frac{6 - 5}{\sqrt{10/12}} \right\} \approx 1 - \Phi(\sqrt{1.2}) \approx 0.137.
\]
A version of the central limit theorem holds even when the random variables \(X_1, X_2, \ldots\) are independent but not necessarily identically-distributed.

**Theorem 8.4.** Let \(X_1, X_2, \ldots\) be a sequence of independent random variables with \(\mu_i = \mathbb{E}[X_i]\) and \(\sigma_i^2 = \text{Var}(X_i)\). If the \(X_i\) are uniformly bounded, i.e., there exists \(M < \infty\) such that \(\mathbb{P}\{X_i < M\} = 1\) for all \(i \geq 1\), and if \(\sum_{i \geq 1} \sigma_i^2 = \infty\), then

\[
\lim_{n \to \infty} \mathbb{P}\left\{ \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \leq x \right\} = \Phi(x)
\]

for each \(x \in \mathbb{R}\).

This more general result can be used to prove the following surprising statement about the number of prime divisors of a randomly chosen positive integer.

**Theorem 8.5. (Erdős-Kac)** For each \(n \geq 1\), let \(X_n\) be uniformly distributed on the set \(\{1, \cdots, n\}\), and define \(\phi(n)\) to be the number of prime divisors of the integer \(n\), e.g., \(\phi(2) = 1\), \(\phi(3) = 1\), \(\phi(4) = 1\), \(\phi(5) = 1\), \(\phi(6) = 2\), etc. Then, for every real number \(x\),

\[
\lim_{n \to \infty} \mathbb{P}\left( \frac{\phi(X_n) - \log \log(n)}{\sqrt{\log \log(n)}} \leq x \right) = \Phi(x).
\]

In other words, if an integer is chosen at random between 1 and \(n\), then for large \(n\) the number of prime divisors of that integer is approximately normally distributed with mean and variance both equal to \(\log \log(n)\).