CLASS NOTES

MATH 334
Introduction to Probability

FALL 2017
# Contents

1 **Basic Ideas of Probability** 1

1.1 Probability Axioms ......................................................... 2
1.2 Sample Spaces with Equally Likely Outcomes ..................... 6

2 **Counting** 8

2.1 Fundamental Principle of Counting ........................................ 8
2.2 Permutations .............................................................. 8
2.3 Combinations .............................................................. 9
2.4 Computing Probabilities by Counting .................................... 13

3 **Conditional Probability and Independence** 15

3.1 Definition and Examples .................................................... 15
3.2 The Law of Total Probability ............................................... 18
3.3 Bayes’ Rule ................................................................. 19
3.4 Probabilities of Intersections of Events ............................... 21
3.5 Independence ............................................................... 22

4 **Discrete Random Variables** 27

4.1 Probability mass functions ............................................... 27
4.2 Expectation of Discrete Random Variables ........................... 29
4.3 Variance and Standard Deviation of Discrete Random Variables 34

5 **Common Parametric Families of Discrete Distributions** 35
5.1 The Discrete Uniform Distribution ........................................ 36
5.2 The Bernoulli Distribution ................................................. 36
5.3 The Binomial Distribution .................................................. 37
5.4 The Hypergeometric Distribution ......................................... 37
5.5 The Poisson Distribution .................................................... 39
5.6 The Geometric Distribution ................................................. 45
5.7 The Negative Binomial Distribution ....................................... 47

6 Continuous Random Variables ............................................. 48

7 Common Parametric Families of Continuous Distributions .......... 50

7.1 The Uniform Distribution .................................................. 50
7.2 The Normal Distribution ................................................... 51
7.3 The Exponential Distribution .............................................. 54
7.4 The Poisson Process ......................................................... 55
7.5 The Gamma Distribution ................................................... 55
7.6 The Beta Distribution ...................................................... 57
7.7 The Lognormal Distribution ............................................... 58

8 The Central Limit Theorem .................................................. 58

8.1 Sample Mean and Sample Variance ....................................... 58
8.2 Statement of the Central Limit Theorem ................................ 60
8.3 Normal Approximation to the Binomial ................................ 63
8.4 Normal Approximation to the Poisson .................................. 64
9 Jointly Distributed Random Variables 65

9.1 Jointly Discrete Random Variables .......................... 65
9.2 The Multinomial Distribution ................................ 67
9.3 Jointly Continuous Random Variables ....................... 67
9.4 Conditional Distributions ...................................... 69
9.5 Independent Random Variables ............................... 70
9.6 Computing Expectations from Joint Distributions ......... 73
9.7 Conditional Expectation ........................................ 74

10 Covariance and Correlation 77

11 Distributions of Functions of Random Variables 83

11.1 Distributions of Functions of Discrete Random Variables .... 83
11.2 Distributions of Functions of Continuous Random Variables .... 83
11.3 Distributions of Sums ........................................... 86

12 Moment Generating Functions 87

12.1 Basic Results .................................................. 87
12.2 Proof of the Central Limit Theorem ........................... 90

13 Sampling from the Normal Distribution 91

13.1 The Chi-Square Distribution ................................ 91
13.2 The $F$-distribution .......................................... 93
13.3 The Student’s $t$ distribution ................................ 94
14 Asymptotics  

14.1 Convergence in Probability ........................................ 95
14.2 The Weak Law of Large Numbers ................................ 95
14.3 Convergence in Law .................................................. 98

Tables ................................................................. 99
1 Basic Ideas of Probability

Intuitively, the concept of probability is applied to the outcomes of a random experiment. A random experiment is some sort of process than will result in one of a number of outcomes. The possible outcomes are known, but it cannot be predicted in advance which one will occur. Following are some examples.

Example 1.1. A fair coin is tossed. The set of possible outcomes is \{Heads, Tails\}. Intuitively, a “fair” coin is one that is equally likely to come up Heads as Tails. This description is not sufficiently precise, since it is not clear what “equally likely” means. When we say that a fair coin is equally likely to come up Heads as Tails, we mean that in a series of coin tosses, the proportion of times that the coin comes up Heads will converge to $1/2$ as the number of tosses increases. It follows that the proportion of times the coin comes up Tails must also converge to $1/2$.

We use the word “probability” to describe this long-run proportion. We say that the probability that a fair coin comes up Heads is $1/2$, and the probability that it comes up Tails is $1/2$. The notation is

$$P(\text{Heads}) = 1/2 \quad P(\text{Tails}) = 1/2$$

Example 1.2. A fair, six-sided die is rolled. The possible outcomes are \{1,2,3,4,5,6\}. Saying that the die is fair means that in a series of rolls, the proportion of times each number comes up will converge to $1/6$, or equivalently, that the probability of each outcome is $1/6$. In notation,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$$

In addition to the six individual outcomes, we can compute probabilities of other events as well. For example, consider the event that the number on the die is even. This event occurs if the outcome that occurs is in the set \{2,4,6\}. Intuitively, we can see that the proportion of rolls whose outcome is in the set \{2,4,6\} will approach $1/2$ as the number of rolls increases. It follows that the probability of this event is $3/6$, or $1/2$. The notation is

$$P(\{2,4,6\}) = 1/2$$

Notice that the event \{2,4,6\} occurs if the outcome is any member of the set \{2,4,6\}.  

There are a total of 64 \(2^6\) events in the die rolling experiment, corresponding to the 64 subsets of the set \(\{1,2,3,4,5,6\}\). These events include the empty set, which has probability zero \((P(\emptyset) = 0)\), and the whole set \(\{1,2,3,4,5,6\}\), which has probability 1.

We will now summarize this intuitive introduction: A random experiment has a number of possible outcomes. When the experiment is performed, one of the outcomes will occur. Given a subset of outcomes, the probability of that subset is the long-run proportion of times that the outcome is in that subset.

### 1.1 Probability Axioms

In order to develop the basic ideas of probability further, we must make them mathematically rigorous. This we now do.

**Definition 1.1.** A sample space (outcome space) is a nonempty set containing the outcomes of a random experiment.

**Definition 1.2.** Given a sample space \(\Omega\), an event \(A\) is a subset of \(\Omega\) to which a probability is assigned.

**Definition 1.3.** Given a sample space \(\Omega\), a probability function is a function \(P\) that assigns to each event \(A\) a number \(P(A)\) in the closed interval \([0,1]\), and which satisfies the following properties:
1. For any event \(A\), \(P(A) \geq 0\).
2. \(P(\Omega) = 1\).
3. If \(A_1, \ldots\) are disjoint events, that is, if \(A_i \cap A_j = 0\) whenever \(i \neq j\), then \(P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)\).

Properties 1–3 above are known as the **Probability Axioms**. Note that Axiom 3 is stated for an infinite sequence of events, but it holds for finite sequences as well. That is, if \(A_1, \ldots, A_n\) are disjoint events, then \(P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)\). This can be shown rigorously by extending the finite sequence \(A_1, \ldots, A_n\) to an infinite sequence by defining \(A_j = \emptyset\) for \(j > n\).

Following are some examples of probability functions.
Example 1.3. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, For each set $A \subset \Omega$, define $P(A) = \frac{\text{number of elements in } A}{6}$. This is an appropriate model for the rolling of a fair die.

Example 1.4. Same $\Omega$, but now define $P(1) = 1/2$, $P(2) = 1/4$, $P(3) = 1/8$, $P(4) = 1/16$, $P(5) = 1/32$, $P(6) = 1/32$. This is a model for a loaded die.

Example 1.5. In the previous example, compute $P(\text{even number})$.

Solution: The event “even number” is $\{2, 4, 6\}$. We know that $P(2) = 1/4$, $P(4) = 1/16$, and $P(6) = 1/32$. Now $\{2, 4, 6\} = \{2\} \cup \{4\} \cup \{6\}$, and the sets $\{2\}$, $\{4\}$, and $\{6\}$ are disjoint. It follows from Axiom 3 that $P(\{2, 4, 6\}) = P(2) + P(4) + P(6) = 1/4 + 1/16 + 1/32 = 11/32$.

In the previous example, probabilities were defined for individual outcomes in the sample space, and the probability of an event was the sum of the probabilities of the outcomes contained in the event. This holds for any event with a countable number of elements, as the following theorem shows:

Theorem 1.1. Let $A = \{e_1, e_2, \ldots\}$ be an event with a countable number of elements. Then $P(A) = \sum P(e_i)$.

Proof: $A = \{e_1\} \cup \{e_2\} \cup \cdots$, and $\{e_1\}$, $\{e_2\}$, $\ldots$ are disjoint. By Axiom 3, $P(A) = \sum P(e_i)$.

Example 1.6. Let $\Omega = \mathbb{N}$. For each $x \in \mathbb{N}$, define $P(x) = (1/2)^x$.

Example 1.7. Let $\Omega$ be the set of nonnegative integers, and let $\lambda$ be any positive number. For each $x \in \Omega$, define $P(x) = e^{-\lambda} \lambda^x / x!$. This is called the Poisson distribution.

The sample spaces we have seen so far have all been countable. In many cases, sample spaces are uncountable. As we have mentioned, when the sample space is countable, we will take the event space $B$ to be the collection of all subsets. Unfortunately, this doesn’t work for uncountable sample spaces, as the following theorem implies:
Theorem 1.2. Let $\Omega$ be an uncountable sample space. Then it is impossible to define probabilities on all the subsets of $\Omega$ in a way that satisfies the Probability Axioms.

Proof: Omitted.

When $\Omega$ is an uncountable sample space, there will be some subsets that do not have probabilities. Fortunately, we will not encounter any of these subsets. Therefore, we will assume that every subset of every sample space we encounter will have a probability.

Here are some basic properties of probability functions that can be deduced from the Axioms.

Theorem 1.3. Let $P$ be a probability function on a sample space $\Omega$. Let $A$ and $B$ be events. Then
1. $P(\emptyset) = 0$
2. $P(A^c) = 1 - P(A)$
3. $P(A) = P(A \cap B) + P(A \cap B^c)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. If $A \subset B$, then $P(A) \leq P(B)$
6. $P(A) \leq 1$

Proof: 1. Since $\emptyset = \emptyset \cup \emptyset$, $P(\emptyset) = P(\emptyset \cup \emptyset)$. Since $\emptyset \cap \emptyset = \emptyset$, $P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset)$. Therefore $P(\emptyset) = P(\emptyset) + P(\emptyset)$, so $P(\emptyset) = 0$.

2. $1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$

3. $A = (A \cap B) \cup (A \cap B^c)$ and $(A \cap B) \cap (A \cap B^c) = \emptyset$. Therefore $P(A) = P(A \cap B) + P(A \cap B^c)$.

4. $A \cup B = B \cup (A \cap B)$, and $B \cap (A \cap B^c) = \emptyset$. So $P(A \cup B) = P(B) + P(A \cup B^c)$. It follows from (3) that $P(A \cup B^c) = P(A) - P(A \cap B)$.

5. $B = A \cup (B \cap A^c)$, and $A \cap (B \cap A^c) = \emptyset$. It follows that $P(B) = P(A) + P(B \cap A^c) \geq P(A)$.

6. $A \subset \Omega$, so $P(A) \leq P(\Omega) = 1$.

Theorem 1.4. DeMorgan’s Laws: Let $A$ and $B$ be sets. Then
1. $A^c \cap B^c = (A \cup B)^c$ and $A^c \cup B^c = (A \cap B)^c$
2. Let $A_1, \ldots$ be a sequence of sets. Then $\cap_i A_i^c = (\cup_i A_i)^c$ and $\cup_i A_i^c = (\cap_i A_i)^c$. 

4
Example 1.8. In a bolt manufacturing process, 10% of the bolts are too short, 5% are too thick, and 3% are both too short and too thick. What is the probability that a randomly chosen bolt is acceptable?

Solution: Let \( S \) be the event that a bolt is too short, and let \( T \) be the event that it is too thick. We want to find \( P(S^c \cap T^c) \). We know that \( P(S) = 0.10 \), \( P(T) = 0.05 \), and \( P(S \cup T) = 0.03 \).

By DeMorgan’s laws, \( S^c \cap T^c = (S \cup T)^c \).

Now \( P(S \cup T) = P(S) + P(T) - P(S \cap T) = 0.10 + 0.05 - 0.03 = 0.12 \).

It follows that \( P(S^c \cap T^c) = 1 - P(S \cup T) = 0.88 \).

Example 1.9. Of those with a particular type of injury, 22% visit both a physical therapist and a chiropractor and 12% visit neither a physical therapist nor a chiropractor. The proportion who visit a chiropractor exceeds by 0.14 the proportion who visit a physical therapist. Find the proportion who visit a physical therapist.

Solution: Let \( T \) be the event that someone visits a physical therapist and let \( C \) be the event that someone visits a chiropractor. We know that \( P(T \cap C) = 0.22 \), \( P(T^c \cap C^c) = 0.12 \), and \( P(C) = P(T) + 0.14 \). We must find \( P(T) \).

First, \( P(T \cup C) = P(T) + P(C) - P(T \cap C) \).

We are given that \( P(T \cap C) = 0.22 \) and \( P(C) = P(T) + 0.14 \).

Also, \( P(T \cup C) = 1 - P(T^c \cap C^c) = 1 - 0.12 = 0.88 \).

Therefore, \( 0.88 = 2P(T) + 0.14 - 0.22 \). Solving, we obtain \( P(T) = 0.48 \).

Example 1.10. If \( P(A \cup B) = 0.7 \) and \( P(A \cup B^c) = 0.9 \), find \( P(A) \).

Solution:

\[
0.7 = P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \text{and} \quad 0.9 = P(A \cup B^c) = P(A) + P(B^c) - P(A \cap B^c)
\]

Therefore \( 0.7 + 0.9 = 2P(A) + P(B) + P(B^c) - [P(A \cap B) + P(A \cap B^c)] \).

Now \( P(A \cap B) + P(A \cap B^c) = P(A) \) and \( P(B) + P(B^c) = 1 \).

Therefore \( 1.6 = P(A) + 1 \), so \( P(A) = 0.6 \).
Theorem 1.5. Bonferroni’s Inequality: Let $A$ and $B$ be events. Then $P(A \cap B) \geq P(A) + P(B) - 1$.

Proof: $P(A) + P(B) - 1 \leq P(A) + P(B) - P(A \cup B) = P(A \cap B)$.

Theorem 1.6. Boole’s Inequality, Finite Sequence: Let $A_1, \ldots, A_n$ be a sequence of events. Then $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$.

Proof: $n = 2$: $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$.

Now let $k$ be an integer for which $P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \cdots + P(A_k)$. Then $P(A_1 \cup \cdots \cup A_k \cup A_{k+1}) = P[(A_1 \cup \cdots \cup A_k) \cup A_{k+1}] \leq P[(A_1 \cup \cdots \cup A_k)] + P(A_{k+1}) \leq P(A_1) + \cdots + P(A_k) + P(A_{k+1})$.

Theorem 1.7. Boole’s Inequality, Infinite Sequence: Let $A_1, \ldots$ be a sequence of events. Then $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.

Proof: Define $A_1^* = A_1$, $A_2^* = A_2 \cap A_1^c$, $A_3^* = A_3 \cap (A_1 \cup A_2)^c$, and generally for each $n$, $A_n^* = A_n \cap (\bigcup_{i=1}^{n-1} A_i)^c$.

Now $\bigcup_{i=1}^{\infty} A_i^* = \bigcup_{i=1}^{\infty} A_i$. Therefore

$$P(\bigcup_{i=1}^{\infty} A_i) = P(\bigcup_{i=1}^{\infty} A_i^*)$$

Furthermore, $A_1^*, \ldots$ are disjoint. Therefore

$$P(\bigcup_{i=1}^{\infty} A_i^*) = \sum_{i=1}^{\infty} P(A_i^*)$$

Finally, $P(A_i^*) \leq P(A_i)$ since $A_i^* \subset A_i$. Therefore

$$\sum_{i=1}^{\infty} P(A_i^*) \leq \sum_{i=1}^{\infty} P(A_i)$$

1.2 Sample Spaces with Equally Likely Outcomes
Theorem 1.8. Let $\Omega = \{e_1, ..., e_n\}$ be a sample space with $n$ equally likely outcomes. Then

1. $P(e_i) = \frac{1}{n}$ for all $i$.

2. Let $A$ be an event containing $k$ outcomes. Then $P(A) = \frac{k}{n}$.

Proof: 1. $1 = P(\Omega) = P(e_1) + \cdots + P(e_n)$. Since $P(e_1) = P(e_2) = \cdots = P(e_n)$, $P(e_i) = \frac{1}{n}$ for all $i$.

2. Renumbering outcomes if necessary, let $A = \{e_1, ..., e_k\}$. Then $A = \{e_1\} \cup \{e_2\} \cup \cdots \cup \{e_k\}$, which is a union of disjoint sets. Therefore $P(A) = P(e_1) + \cdots + P(e_k) = \frac{k}{n}$.

Example 1.11. Two coins are tossed. Find the probability that both come up heads.

Model 1: Equally likely outcomes are $\{0, 1, 2\}$, so probability is $1/3$.
Model 2: Pretend that one coin is red and one coin is blue. List the outcomes, with the red coin first: $\{HH, HT, TH, TT\}$. Now probability is $1/4$.

Experiments show that the long run frequency approaches $1/4$. So model 2 is correct. (They are distinguishable particles.)

Example 1.12. Two dice are rolled. Find a sample space with equally likely outcomes for the sum of the dice.

Solution: One sample space is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, but these outcomes are not equally likely. To construct a sample space with equally likely outcomes, consider rolling the dice one at a time, and make a $2 \times 2$ table listing all 36 outcomes.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

From this table we can compute any probabilities involving the sum of two dice. For example, $P(6) = 5/36$, $P(7) = 1/6$, $P(2) = P(12) = 1/36$, $P(\text{doubles}) = 1/6$, $P(\text{even number}) = 1/2$. 

7
2 Counting

2.1 Fundamental Principle of Counting

Here is a simple problem. You are going to buy a car. You have three choices of color: red, blue, and green, and two choices of engine size: large or small. How many different choices of car do you have? We can answer this question by using the Fundamental Principle of Counting.

**Fundamental Principle of Counting:** If an operation can be performed in any of \( m \) ways, and another operation can be performed in any of \( n \) ways, then the total number of ways to perform the sequence of two operations is \( mn \).

Now let’s say you have three choices of sound system for your car. Now how many choices of car do you have? We can answer this question by using the following generalization of the Fundamental Principle of Counting.

**Fundamental Principle of Counting (General Case):** If a sequence of \( k \) operations is to be performed, and the first operation can be performed in any of \( n_1 \) ways, the second in any of \( n_2 \) ways, and so on, then the total number of ways to perform the sequence of operations is \( n_1 n_2 \cdots n_k \).

**Example 2.1.** A license plate consists of three digits followed by three letters. How many different license plates can be made?

**Solution:** There are 10 choices for each digit, and 26 choices for each letter. There are thus \( 10 \cdot 10 \cdot 10 \cdot 26 \cdot 26 \cdot 26 = 17,576,000 \) different license plates.

**Example 2.2.** In how many different ways can the letters ABCDE be ordered?

**Solution:** There are 5 choices for the first letter, then 4 choices for the second letter, and so on, so the total number of orders is \( (5)(4)(3)(2)(1) = 120 \).

**Definition 2.1.** Let \( n \) be a positive integer. Then \( n! = n(n - 1) \cdots (3)(2)(1) \). Also, \( 0! = 1 \).

2.2 Permutations
Definition 2.2. A permutation of objects is an ordering of them.

Theorem 2.1. The number of distinct permutations of $n$ objects is $n!$.

Proof: There are $n$ ways to choose the first object, etc.

Example 2.3. Eight people enter a race. The person finishing first will receive a gold medal, second place silver, third place bronze. In how many different ways can the medals be awarded?

Solution: There are 8 choices for the gold medal, then 7 choices for the silver medal and 6 choices for the bronze. There are thus $8 \cdot 7 \cdot 6 = 336$ ways to award the medals.

In Example 2.3, we counted the number of permutations of 3 items chosen from 8. The following theorem presents the general principle.

Theorem 2.2. The number of ways to choose $k$ objects from $n$ and order them is

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

Proof: There are $n$ ways to choose the first object, etc.

2.3 Combinations

In some cases, when choosing a set of objects from a larger set, we don’t care about the ordering of the chosen objects, we care only which objects are chosen. Each distinct group of objects that can be selected, without regard to order, is called a combination. We will now show how to determine the number of combinations of $k$ objects chosen from a set of $n$ objects.

Example 2.4. Five people, A, B, C, D, and E enter a race. The top three finishers win medals, but all the medals are the same. In how many different ways can the medals be awarded?

Solution: There are $\frac{5!}{(5-3)!} = 60$ permutations of 3 items chosen from 5. We list them.
The 60 permutations are arranged in 10 columns of 6 permutations each. Within each column, the three objects are the same, and the column contains the 6 different permutations of those three objects. Therefore each column represents a distinct combination of 3 objects chosen from 5, and there are 10 such combinations. We can see that the number of combinations of 3 objects chosen from 5 can be found by dividing the number of permutations of 3 objects chosen from 5, which is $5!/(5-3)!$, by the number of permutations of 3 objects, which is $3!$ Therefore the number of combinations of 3 objects chosen from 5 is \[ \frac{5!}{3!(5-3)!} \] 

The number of combinations of $k$ objects chosen from $n$ is often denoted by the symbol $\binom{n}{k}$. The reasoning above can be generalized to derive an expression for $\binom{n}{k}$.

**Theorem 2.3.** The number of combinations of $k$ objects chosen from a group of $n$ objects is \[ \binom{n}{k} = \frac{n!}{k!(n-k)!} \]

**Example 2.5.** Sally has 12 books. She wants to put four of them on her bookshelf. How many ways can she do this if
a) The order matters
b) The order doesn’t matter

**Solution:** a) The number of permutations of 4 objects chosen from 12 is \[ \frac{12!}{(12-4)!} = 11,880. \]
b) The number of combinations of 4 objects chosen from 12 is \[ \binom{12}{4} = \frac{12!}{4!(12-4)!} = 495. \]
Example 2.6. Five of Sally’s books are novels, two are biographies, three are probability textbooks, and two are comic books. She wants to put one of each type of book on her shelf. How many ways can she do this if
a) The order doesn’t matter
b) The order matters

Solution: a) Sally must choose a novel, a biography, a probability textbook, and a comic book. There are 5 ways to choose the novel, 2 ways to choose the biography, 3 ways to choose the probability textbook, and 2 ways to choose the comic book. Using the Fundamental Principle of Counting, we see that there are $5 \cdot 2 \cdot 3 \cdot 2 = 60$ ways to choose the books.

b) Sally must first choose the books to put on the shelf, then she must choose an order. From part a) we know that there are 60 ways to choose the books. Now there are $4! = 24$ permutation of 4 items. Therefore there are $60 \cdot 24 = 1440$ ways to choose the books and order them.

Example 2.7. Sally has just bought a bigger bookshelf, and now she wants to put all 12 books on it. For the purposes of ordering, she considers all the books of a given type to be equivalent. So therefore one ordering would be NPBCNNPPBCNN. How many different ways are there to arrange the books?

Solution: There are 12 spaces on the bookshelf. Sally must choose 5 of them for the novels. Then she must choose 2 of the remaining 7 spaces for the biographies, 3 of the remaining 5 for the probability textbooks, and 2 of the remaining 2 for the comic books. The number of ways to do this is

$$\binom{12}{5} \cdot \binom{7}{2} \cdot \binom{5}{3} \cdot \binom{2}{2} = \frac{12!}{5!(12-5)!} \cdot \frac{7!}{2!(7-2)!} \cdot \frac{5!}{3!(5-3)!} \cdot \frac{2!}{2!(2-2)!} = \frac{12!}{5!2!3!2!} = 166,320$$

Theorem 2.4. In a group of $n$ objects, the objects are of $k$ different types. Assume $n_1$ are of type 1, $n_2$ are of type 2, and so on. The number of permutations of the $n$ objects is

$$\frac{n!}{n_1!n_2!\cdots n_k!}.$$  

Proof:
First we choose $n_1$ positions for the objects of type 1. There are $\binom{n}{n_1}$ ways to do this. Then we choose $n_2$ positions for the objects of type 2. There are $\binom{n-n_1}{n_2}$ ways to do this.
Then we choose \( n_3 \) positions for the objects of type 3. There are \( \binom{n - n_1 - n_2}{n_3} \) ways to do this. Continuing, the total number of ways to arrange the objects is

\[
\binom{n}{n_1} \binom{n - n_1}{n_2} \cdots \binom{n}{n_k} = \frac{n!}{n_1!n_2!\cdots n_k!}
\]

**Example 2.8.** How many orderings are there of the letters in the word “Connecticut”?

**Solution:** There are 11 letters, of 7 types: There are 3 C’s, 1 O, 2 N’s, 1 E, 2 T’s, 1 I, and 1 U. The number of orderings is \( \frac{11!}{3!2!2!1!1!} = 1,663,200 \).

The formula for combinations can be used to derive an important theorem in algebra, known as the Binomial Theorem.

**Theorem 2.5. The Binomial Theorem:** For any two numbers \( x \) and \( y \), and any positive integer \( n \),

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

**Proof:**

\( (x + y)^n = (x + y)(x + y) \cdots (x + y) \). The product consist of \( 2^n \) terms, with each term being formed by choosing either an \( x \) or a \( y \) from each factor. The number of terms that are \( x^k y^{n-k} \) is equal to the number of ways to choose \( k \) factors from \( n \) factors. This number is \( \binom{n}{k} \).

**Theorem 2.6.** For any positive integer \( n \), \( \sum_{k=0}^{n} \binom{n}{k} = 2^n \).

**Proof:**

\[
2^n = (1 + 1)^n = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k}
\]
2.4 Computing Probabilities by Counting

Example 2.9. In the Powerball lottery, five balls are drawn from an urn containing balls numbered from 1 to 59.

a) What is the probability that #1 is the first ball drawn?
b) What is the probability that #1 is the second ball drawn?
c) What is the probability that one of the five balls is #1?

Solution:
a) There are 59 equally likely outcomes, and one of them is #1. So \( P(\text{#1 is first ball drawn}) = \frac{1}{59} \).

b) We will do this one two ways. First the hard way. Consider the sample space to consist of all sequences of two balls. There are 59 possibilities for the first ball, then 58 possibilities for the second. So there are 59 \( \cdot \) 58 outcomes in the sample space. Now we count the number of sequences in which the second ball is #1. There are 58 possibilities for the first ball and 1 possibility for the second. Therefore \( P(\text{second ball is #1}) = \frac{(58)(1)}{(59)(58)} = \frac{1}{59} \).

Now for the easy way: Any of the 59 balls could be the second one chosen, and they are all equally likely. So \( P(\text{second ball is #1}) = \frac{1}{59} \).

c) Imagine that all the balls are drawn. There are 59 possible locations for ball #1; from 1st ball drawn to 59th ball drawn. Each location is equally likely. The event that ball #1 is one of the first five drawn consists of 5 of the 59 locations. Therefore \( P(\text{ball #1 is in the first five drawn}) = \frac{5}{59} \).

Example 2.10. Fifteen men and ten women buy raffle tickets. Three tickets are chosen at random to be winners. What is the probability that two winners are men and one is a woman?

Solution: The total number of outcomes is the number of combinations of 3 chosen from 25, which is \( \frac{25!}{3!22!} = 2300 \). Now we have to compute the number of combinations in which two are men and one is a woman. There are \( \binom{15}{2} \) ways to choose two men from 15. There are \( \binom{10}{1} \) ways to choose one woman from 10. So the number of ways to choose 2 men and
1 woman is \( \binom{15}{2} \binom{10}{1} = 105 \cdot 10 = 1050 \). So

\[
P(\text{two winners are men and one is a woman}) = \frac{1050}{2300} = 0.4565
\]

**Example 2.11.** Texas Hold’em is a variant of poker in which each player is initially dealt two cards, called hole cards, face down. During play, a total of five cards (called the board) are dealt face up in the middle of the table. Each player makes the best possible five-card hand from his own hole cards combined with the cards on the board.

The best hole cards you can get are two aces, known as pocket aces, pocket rockets, or occasionally, American Airlines (AA). What is the probability that you are dealt pocket aces?

**Solution:**

The number of possible sets of hole cards is \( \binom{52}{2} \). Since there are four aces in the deck, the number of sets of hole cards that are both aces is \( \binom{4}{2} \).

The probability is

\[
\frac{\binom{4}{2}}{\binom{52}{2}} = \frac{1}{221} = 0.004525.
\]

**Example 2.12.** The second-best hand you can be dealt in Texas Hold’em is two kings (pocket kings). You are playing against a single opponent. You are dealt pocket kings. What is the probability that your opponent has pocket aces?

**Solution:**

The experiment consists of choosing two cards to deal to your opponent from the 50 cards that you don’t hold. There are \( \binom{50}{2} \) ways to do this. The event consists of being dealt pocket aces. There are \( \binom{4}{2} \) ways to do this. The probability is

\[
\frac{\binom{4}{2}}{\binom{50}{2}} = \frac{6}{1225} = 0.004898.
\]
**Example 2.13.** You are dealt pocket aces. What is the probability that one or both of the remaining aces come up on the board?

**Solution:** This is an example of a problem in which it is easiest to compute the probability of the complement of the event and subtract from 1. We will find the probability that no aces come up on the board.

There are 50 cards that are not in your hand. The number of combinations of 5 cards chosen from 50 is \( \binom{50}{5} = 2,118,760 \). Of the 50 cards, 48 are not aces. The number of combinations of 5 non-aces chosen from 48 is \( \binom{48}{5} = 1,712,304 \). Therefore

\[
P(\text{no aces come up on the board}) = \frac{1,712,304}{2,118,760} = 0.8082
\]

and

\[
P(\text{one or more aces come up on the board}) = 1 - 0.8082 = 0.1918
\]

### 3 Conditional Probability and Independence

#### 3.1 Definition and Examples

Sometimes, when we want to compute the probability of an event, we are given information about another event that has already occurred or that will occur. This information can change the probability we wish to compute. We illustrate the idea with the following dice table.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>

**Example 3.1.** Find the probability of rolling an 8.

**Solution:**

There are 36 equally likely outcomes in the sample space. Of these, 5 correspond to an 8. Therefore \( P(8) = 5/36 \).
Example 3.2. Find the probability of rolling an 8, given that the first die has been rolled and came up 3.

Solution:
We have been given the condition that the first die is a 3. This reduces the sample space to six possible outcomes, (3,1), (3,2), (3,3), (3,4), (3,5), and (3,6). These six outcomes are equally likely, and one of them, (3,5) corresponds to rolling an 8. The probability of rolling an 8, given that the first die has been rolled and came up 3 is 1/6.

Terminology: We refer to a probability of an event, given a condition involving another event, as a conditional probability. In Example 3.2, we would say that the conditional probability of an 8, given that the first die is 3, is 1/6.

Notation: Let A and B be events. The conditional probability of A given B is denoted P(A|B). So we would write the result of Example 3.2 as P(8|first die is 3) = 1/6.

Note that P(8|first die is 3) \neq P(8). The condition that the first die is 3 changes the probability of rolling an 8.

Example 3.3. Find P(7), P(7|first die is 3).

Solution:
Using the dice table we see that P(7) = 1/6. Of the six outcomes corresponding to the first die coming up 3, one, (3,4) corresponds to a 7. Therefore P(7|first die is 3) = 1/6.

Note that P(7) = P(7|first die is 3). The condition that the first die is 3 does not change the probability of rolling a 7.

Example 3.4. Find P(6|doubles), and P(doubles|6).

Solution:
There are six outcomes corresponding to doubles, and one of them, (3,3) also corresponds to a 6. Therefore P(6|doubles) = 1/6. There are five outcomes corresponding to rolling a 6, and one of them, (3,3), also corresponds to rolling doubles. Therefore P(doubles|6) = 1/5.

Note that P(doubles) = P(6|doubles). In most cases, P(A|B) \neq P(B|A).
In the previous examples, we computed conditional probabilities by counting outcomes in a sample space. In many cases this is not feasible, for example when the outcomes are not equally likely. We therefore derive a method that can be used in general. To see how it works, consider the calculation of $P(8 \mid \text{first die is 3})$. We found that $P(8 \mid \text{first die is 3}) = 1/6$. The denominator represents the number of outcomes that correspond to the event that first die being 3. The numerator represents the number of outcomes that correspond to event that the sum is 8 and the first die is 3. Now divide both numerator and denominator by 36, the total number of outcomes in the sample space. We obtain

$$P(8 \mid \text{first die is 3}) = \frac{1/36}{6/36}$$

The numerator now represents the probability that the sum is 8 and the first die is 3, that is, $P(8 \cap \text{first die is 3})$. The denominator now represents the probability that the first die is 3, that is, $P(\text{first die is 3})$. We have now expressed the conditional probability in terms of other probabilities. This idea provides the definition of conditional probability.

**Definition 3.1.** Let $A, B$ be events with $P(B) > 0$. The conditional probability of $A$ given $B$, denoted $P(A \mid B)$, is $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$.

**Example 3.5.** You roll two dice. Find $P(8 \mid 1\text{st is 3})$ using the formula.

**Solution:**

$$P(8 \mid 1\text{st die is 3}) \cdot \frac{P(8 \cap 1\text{st die is 3})}{P(1\text{st die is 3})} = \frac{1/36}{1/6} = \frac{1}{6}.$$ 

**Example 3.6.** You roll two dice until either a 6 or a 7 comes up. What is the probability that a 6 comes up before a 7?

**Solution:** We know that the last roll will be either a 6 or a 7. Given this condition, we want to find the probability that the last roll is a 6. Thus

$$P(6 \text{ before } 7) = P(\text{last roll is 6} \mid \text{last roll is 6 or 7}) = \frac{P(\text{last roll is 6} \cap \text{last roll is 6 or 7})}{P(\text{last roll is a 6 or 7})}$$

$$= \frac{P(\text{last roll is 6})}{P(\text{last roll is a 6 or 7})} = \frac{5/36}{5/36 + 6/36} = \frac{5}{11}$$

The following problem, known as Bertrand’s box paradox, is one of the most famous problems in probability theory.
Example 3.7. A chest contains three drawers. One drawer contains two gold coins, another contains two silver coins, and the third contains one gold and one silver coin. A drawer is chosen at random, and a coin is chosen at random from that drawer. If the chosen coin is gold, what is the probability that the other coin in the chosen drawer is gold?

Solution: Let $G_1$ denote the event that the chosen coin is gold, and let $G_2$ denote the event that the other coin in the same drawer is gold. We must find $P(G_2 \mid G_1)$.

Using the definition of conditional probability, $P(G_2 \mid G_1) = \frac{P(G_2 \cap G_1)}{P(G_1)}$. Now $G_2 \cap G_1$ is the event that the drawer with two gold coins was chosen, so $P(G_2 \cap G_1) = 1/3$. To compute $P(G_1)$, note that there are six coins in all, each of which is equally likely to be chosen, and three of them are gold. Therefore $P(G_1) = 1/2$. We conclude that $P(G_2 \mid G_1) = \frac{1/3}{1/2} = 2/3$.

The following proposition follows immediately from the definition of conditional probability. It provides a method of computing the probability of the intersection of two events.

**Proposition 3.1.** $P(A \cap B) = \begin{cases} P(A \mid B)P(B) \\ P(A)P(B \mid A) \end{cases}$

Example 3.8. You toss a fair coin. If it comes up heads, you roll a die. If it comes up tails, you roll two dice. What is the probability that the coin comes up heads and you roll a 6?

Solution: If the coin comes up heads, we roll one die, so the probability of a 6 is 1/6. In other words, $P(6 \mid H) = 1/6$. We also know that $P(H) = 1/2$. Therefore $P(6 \cap H) = P(H)P(6 \mid H) = (1/2)(1/6) = 1/12$.

### 3.2 The Law of Total Probability

Sometimes we can use conditional probabilities to compute an unconditional probability. The rule most often used to do this is the **Law of Total Probability**. Example 3.9 presents an illustration.
Example 3.9. You toss a fair coin. If it comes up heads, you roll a die. If it comes up tails, you roll two dice. What is the probability that you roll a 6?

Solution: If we knew the outcome of the coin, we could find the probability of rolling a 6. Specifically, if we toss heads, we roll one die, so the conditional probability of a 6 given heads is \( P(6 | \text{Heads}) = \frac{1}{6} \). If we toss tails, we roll two dice, so the conditional probability of a 6 given tails is \( P(6 | \text{Tails}) = \frac{5}{36} \).

Now \( P(6 \cap \text{Heads}) = P(6 | \text{Heads})P(\text{Heads}) = (1/6)(1/2) = 1/12 \).
\( P(6 \cap \text{Tails}) = P(6 | \text{Tails})P(\text{Tails}) = (5/36)(1/2) = 5/72 \).

Finally, \( P(6) = P(6 \cap \text{Heads}) + P(6 \cap \text{Tails}) = 1/12 + 5/72 = 11/72 \).

Theorem 3.1 presents the general principle used to solve Example 3.9.

Theorem 3.1. Law of Total Probability (Rule of Average Conditional Probabilities) Special Case: Let \( A, B \) be events with \( P(B) > 0 \). Then \( P(A) = P(A\mid B)P(B) + P(A\mid B^c)P(B^c) \).

Proof: The sets \( A \cap B \) and \( A \cap B^c \) are disjoint, and \( (A \cap B) \cup (A \cap B^c) = A \). Therefore \( P(A) = P[(A \cap B) \cup (A \cap B^c)] = P(A \cap B) + P(A \cap B^c) = P(A\mid B)P(B) + P(A\mid B^c)P(B^c) \).

Theorem 3.2. Law of Total Probability (Rule of Average Conditional Probabilities): Let \( A \) be an event. Let \( B_1, \ldots \) be a disjoint collection of events such that \( P(B_i) > 0 \) for each \( i \) and \( A \subset (\cup B_i) \). Then \( P(A) = \sum P(A\mid B_i)P(B_i) \).

Proof: The sets \( A \cap B_i \) are disjoint, and \( \cup (A \cap B_i) = A \). Therefore \( P(A) = P(\cup (A \cap B_i)) = \sum P(A \cap B_i) = \sum P(A\mid B_i)P(B_i) \).

3.3 Bayes’ Rule

Sometimes we know \( P(A\mid B) \) and we want to find \( P(B\mid A) \). This can be done by using a result known as Bayes’ Rule.
Theorem 3.3. Bayes’ Rule, Special Case: Let $A$, $B$ be events with $P(A) > 0$ and $P(B) > 0$.

Then

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Proof: $P(B|A) = P(A \cap B)/P(A)$. Now $P(A \cap B) = P(A|B)P(B)$, and

$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$.

Theorem 3.4. Bayes’ Rule, General Case: Let $A$ be an event. Let $B_1, ..., B_n$ be a disjoint collection of events such that $P(A) > 0$, $P(B_i) > 0$ for each $i$ and $A \subset (\cup B_i)$. Then for each $j$,

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum P(A|B_i)P(B_i)}$$

Proof: $P(B_j|A) = P(A \cap B_j)/P(A)$. Now $P(A \cap B_j) = P(A|B_j)P(B_j)$, and by the Law of Total Probability, $P(A) = \sum P(A|B_i)P(B_i)$.

Following is a famous application of Bayes’ Rule in the field of medicine. It explains why medical tests are often repeated when they give a positive result.

Example 3.10. One percent of the people in a certain population have a certain disease. A test designed to detect the disease gives a positive signal with probability 0.99 for someone who has the disease, and a negative signal with probability 0.95 for someone who does not have the disease. (In medical terms, the test has sensitivity 0.99 and specificity 0.95). A person chosen at random tests positive. What is the probability that this person has the disease?

Solution: Let $D$ be the event that a randomly chosen person has the disease, let $+$ be the event that the test gives a positive signal, and let $-$ be the event that the test gives a negative signal. We are given that $P(D) = 0.01$, $P(+) | D) = 0.99$, and $P(− | D^c) = 0.95$.

We want to find $P(D|+)$. Using Bayes’ Rule,

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$
We know the values of $P(+ | D)$ and $P(D)$. Now $P(+ | D^c) = 1 - P(- | D^c) = 1 - 0.95 = 0.05$, and $P(D^c) = 1 - P(D) = 1 - 0.01 = 0.99$. Using Bayes’ Rule,

$$P(D | +) = \frac{(0.99)(0.01)}{(0.99)(0.01) + (0.05)(0.01)} = \frac{1}{6}$$

Take a moment to contemplate this result. The test gives the correct result 99% of the time for diseased people, and 95% of the time for non-diseased people. Yet only $1/6$ of the positive results correspond to a diseased person. This is due to the fact that only 1% of the people have the disease. Even though the probability of a positive result is low for a person without the disease, there are so many people without the disease that they account for most of the positive results.

3.4 Probabilities of Intersections of Events

Proposition 3.1 on page 18 provided a method of computing the probability of the intersection of two events. The next theorem extends this to intersections of arbitrary numbers of events.

Theorem 3.5. Let $A_1, ..., A_n$ be events with $P(A_1 \cap \cdots \cap A_{n-1}) > 0$. Then

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \cdots P(A_n | A_1 \cap \cdots A_{n-1})$$

Proof: By induction. True for $n = 2$. Now let $k$ be an integer such that $P(A_1 \cap \cdots \cap A_k) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2) \cdots P(A_k | A_1 \cap \cdots A_{k-1})$.

Let $B = A_1 \cap \cdots \cap A_k$. Then $P(A_1 \cap \cdots \cap A_{k+1}) = P(A_{k+1} \cap B) = P(B)P(A_{k+1} | B)$.

Example 3.11. An urn contains 8 red balls and 6 green balls. Three balls are drawn without replacement. Find the probability that three red balls are withdrawn.

Solution: Let $R_1$ be the event that the first ball drawn is red, let $R_2$ be the event that the second ball drawn is red, and let $R_3$ be the event that the third ball drawn is red. We want to find $P(R_1 \cap R_2 \cap R_3)$. Now

$$P(R_1 \cap R_2 \cap R_3) = P(R_1)P(R_2 | R_1)P(R_3 | R_1 \cap R_2)$$

$P(R_1) = 8/14$, $P(R_2 | R_1) = 7/13$, and $P(R_3 | R_1 \cap R_2) = 6/12$. Therefore

$$P(R_1 \cap R_2 \cap R_3) = (8/14)(7/13)(6/12) = 2/13$$
Theorem 3.6. Let $A_1, A_2, B$ be events with $P(A_1 \cap B) > 0$. Then $P(A_2 \cap A_1|B) = P(A_1|B)P(A_2|A_1 \cap B)$.

Proof:

$$P(A_2 \cap A_1|B) = \frac{P(A_2 \cap A_1 \cap B)}{P(B)} = \frac{P(A_1 \cap B)}{P(B)} \frac{P(A_2 \cap A_1 \cap B)}{P(A_1 \cap B)}$$

Example 3.12. If it rains today, the probability of rain tomorrow is 0.4. If it rains both today and tomorrow, the probability of rain the day after tomorrow is 0.5. Assume it rains today. Find the probability that it rains both tomorrow and the day after tomorrow.

Solution: Let $R_1$ denote the event that it rains today, $R_2$ the event that it rains tomorrow, and $R_3$ the event that it rains the day after tomorrow. We want to find $P(R_3 \cap R_2 | R_1)$.

We know that $P(R_3 | R_2 \cap R_1) = 0.5$ and $P(R_2 | R_1) = 0.4$.

Now $P(R_3 \cap R_2 | R_1) = P(R_2 | R_1)P(R_3 | R_2 \cap R_1) = (0.4)(0.5) = 0.2$.

3.5 Independence

One of the most important ideas in probability is the notion of independence. Intuitively, two events are independent if the occurrence of one does not affect the probability of the other. This intuition is not well reflected in the definition we present, but it will become clear shortly.

Definition 3.2. Two events $A$ and $B$ are independent if $P(A \cap B) = P(A)P(B)$.

Example 3.13. Two dice are rolled: $A =$ first die comes up 3, $B =$ sum is 7. Show that $A$ and $B$ are independent.

Solution: We must show that $P(A \cap B) = P(A)P(B)$. Now $P(A) = P(B) = 1/6$. The event $A \cap B$ is the event that the first die is 3 and the second die is 4, which is one of 36 equally likely outcomes for the roll of two dice. Therefore $P(A \cap B) = 1/36 = P(A)P(B)$.

Note that if $P(A) = 0$ or $P(B) = 0$ then $A$ and $B$ are independent.
**Theorem 3.7.** Let \( A \) and \( B \) be events with \( P(B) > 0 \). Then \( A \) and \( B \) are independent if and only if \( P(A|B) = P(A) \). Similarly, if \( P(A) > 0 \) then \( A \) and \( B \) are independent if and only if \( P(B|A) = P(B) \).

**Proof:** Assume \( P(B) > 0 \). Now

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

If \( A \) and \( B \) are independent then

\[
\frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)
\]

If \( P(A|B) = P(A) \) then

\[
P(A) = \frac{P(A \cap B)}{P(B)}
\]

so \( P(A \cap B) = P(A)P(B) \).

**Remark 3.1.** We often model successive experiments as independent. For example, successive coin tosses, successive die rolls, responses of successively sampled individuals.

**Example 3.14.** Two dice are rolled. Let \( A = \text{odd total} \), \( B = \text{first die comes up 1} \), \( C = \text{sum is 7} \). Which pairs of events are independent?

**Solution:** We check \( A \) and \( B \). \( P(A) = 1/2 \) and \( P(B) = 1/6 \). Now \( A \cap B \) is the event \( \{(1,2), (1,4), (1,6)\} \). So \( P(A \cap B) = 3/36 = 1/12 = P(A)P(B) \). So \( A \) and \( B \) are independent.

We check \( A \) and \( C \). \( P(A) = 1/2 \) and \( P(C) = 1/6 \). Now \( A \cap C = C \). So \( P(A \cap C) = P(C) = 1/6 \neq P(A)P(C) \).

Finally we check \( B \) and \( C \). \( P(B) = 1/6 \) and \( P(C) = 1/6 \). The event \( B \cap C \) is \( \{(1,6)\} \). So \( P(B \cap C) = 1/36 = P(B)P(C) \). So \( B \) and \( C \) are independent.

**Theorem 3.8.** If \( A \) and \( B \) are independent, so are \( A^c \) and \( B^c \), \( A \) and \( B^c \), and \( A^c \) and \( B^c \).

**Proof:** We show the result for \( A \) and \( B^c \). The rest are similar.

\[
P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(B^c)
\]
**Definition 3.3.** Let $A_1, \ldots, A_n$ be events. $A_1, \ldots, A_n$ are pairwise independent if $A_i$ and $A_j$ are independent whenever $i \neq j$.

**Definition 3.4.** Let $A_1, \ldots, A_n$ be events. $A_1, \ldots, A_n$ are mutually independent if for every subcollection $A_{i_1}, \ldots, A_{i_k}$ of events, $P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})$.

**Remark 3.2.** Pairwise independence does not imply mutual independence. For example, if two dice are rolled, let $A_1 =$ first die comes up odd, $A_2 =$ second die comes up odd, $A_3 =$ sum of two dice is odd. The probabilities are $P(A_1) = P(A_2) = P(A_3) = 1/2$. It can be checked that $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$, and $P(A_2 \cap A_3) = P(A_2)P(A_3)$. However, $A_1 \cap A_2 \cap A_3 = \emptyset$, so $P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)$.

Often independent experiments are performed, and we need to compute the probability that a certain result happens at least once. Often the best way to approach such problems is to compute the probability that the result doesn’t happen at all, and subtract from 1. The following example illustrates the method.

**Example 3.15.** A fair coin is tossed five times. What is the probability that it comes up heads at least once?

**Solution:** The key is to realize that the complement of coming up heads at least once is never coming up heads. Therefore

$$P(\text{at least one head}) = 1 - P(\text{no heads}) = 1 - P(\text{five tails}) = 1 - (1/2)^5 = 31/32.$$ 

In Example 3.10 we found that even for a medical test that is quite accurate, most positive results are false positives. We said at the time that this problem could be handled by repeating positive tests. The next example illustrates this.

**Example 3.16.** Same as Example 3.10, but now assume that each person is tested twice, and the results of the tests are independent conditional on disease status i.e., $P(+ + \mid D) = P(+)P(+)P(+ \mid D)c$ and $P(+ + \mid Dc) = P(+)P(+)P(+)P(Dc)$. If a person tests positive twice, what is the probability that the person has the disease?
Solution: In Example 3.10, we are given $P(+ | D) = 0.99$, $P(D) = 0.01$, $P(- | D^c) = 0.95$. By independence, $P( + + | D) = (0.99)^2 = 0.9801$. Also $P(+ | D^c) = 1 - P(- | D^c) = 0.05$, so $P(+ + | D^c) = (0.05)^2 = 0.0025$. We need to find $P(D | ++)$. Using Bayes’ Rule,

$$P(D | ++) = \frac{P(+ + | D)P(D)}{P(+ + | D)P(D) + P(+ + | D^c)P(D^c)} = \frac{(0.9801)(0.01)}{(0.9801)(0.01) + (0.0025)(0.99)} = 0.7984$$

We see that a patient who tests positive twice has the disease with probability $\approx 0.8$. In Example 3.10 we saw that a patient who tests positive once has the disease with a probability of only $1/6$.

Example 3.17. A system contains two components, A and B, connected in series as shown in the diagram below:

![Series Connection Diagram]

The system will function only if both components function. The probability that A functions is given by $P(A) = 0.98$, and the probability that B functions is given by $P(B) = 0.95$. Assume that A and B function independently. Find the probability that the system functions.

Solution: $P(\text{system functions}) = P(A \cap B) = P(A)P(B) = (0.98)(0.95) = 0.931$.

The next example illustrates the computation of the reliability of a system consisting of two components connected in parallel.

Example 3.18. A system contains two components, C and D, connected in parallel as shown in the diagram below:

![Parallel Connection Diagram]

The system will function if either C or D functions. The probability that C functions is 0.90, and the probability that D functions is 0.85. Assume C and D function independently. Find the probability that the system functions.
Solution:
\[ P(\text{system functions}) = P(C \cup D) = P(C) + P(D) - P(C \cap D). \]
Now \[ P(C \cap D) + P(C)P(D) = (0.90)(0.85) = 0.765. \] Therefore
\[ P(\text{system functions}) = 0.90 + 0.85 - 0.765 = 0.985 \]

Example 3.19. Refer to Example 3.18.

a) If the probability that C fails is 0.08, and the probability that D fails is 0.12, find the probability that the system functions.

b) If both C and D have probability \( p \) of failing, what must the value of \( p \) be so that the probability that the system functions is 0.99?

c) If three components are connected in parallel, function independently, and each has probability \( p \) of failing, what must the value of \( p \) be so that the probability that the system functions is 0.99?

d) If components function independently, and each component has probability 0.5 of failing, what is the minimum number of components that must be connected in parallel so that the probability that the system functions is at least 0.99?

Solution:

a) This can be solved in the same way as Example 3.18. We will show an alternative method.
\[ P(C \cup D) = 1 - P(C^c \cap D^c) = 1 - P(C^c)P(D^c) = 1 - (0.08)(0.12) = 0.9904. \]

b) By the method of part a), \( P(C \cup D) = 1 - p^2 = 0.99 \). Solve for \( p \) to obtain \( p = 0.1 \).

c) \( P(\text{at least one component functions}) = 1 - P(\text{all three fail}) = 1 - p^3 = 0.99 \). Solve for \( p \) to obtain \( p = 0.2154 \).

d) Let \( n \) be the required number of components. Then \( P(\text{at least one component functions}) = 1 - P(\text{all fail}) = 1 - (0.5)^n = 0.99 \). Solve for \( n \) to obtain \( n = 6.64 \). Round up to \( n = 7 \).

Example 3.20. A system consists of four components connected as shown below.
Assume A, B, C, and D function independently. If the probabilities that A, B, C, and D fail are 0.10, 0.05, 0.10, and 0.20 respectively, what is the probability that the system functions?

**Solution:** The system functions if A and B both function, or if either C or D function. Therefore \( P(\text{system functions}) = P((A \cap B) \cup (C \cup D)) \).

Now \( P((A \cap B) \cup (C \cup D)) = P(A \cap B) + P(C \cup D) - P((A \cap B) \cap (C \cup D)) \).

We compute \( P(A \cap B) = P(A)P(B) = (0.90)(0.95) = 0.855 \)

We compute \( P(C \cup D) = P(C) + P(D) - P(C \cap D) = 0.9 + 0.8 - (0.9)(0.8) = 0.98 \) [Note that \( P(C \cap D) = P(C)P(D) \).]

Now \( P((A \cap B) \cap (C \cup D)) = P(A \cap B)P(C \cup D) = (0.855)(0.98) = 0.8379 \).

Finally, \( P((A \cap B) \cup (C \cup D)) = 0.855 + 0.98 - 0.8379 = 0.9971 \).

## 4 Discrete Random Variables

**Definition 4.1.** Let \( \Omega \) be a sample space with probability function \( P \). A random variable \( X \) is a function with domain \( \Omega \) and range \( \mathbb{R} \).

**Example 4.1.** Roll two dice. The sample space is \( \Omega = \{(i,j)|i = 1,\ldots,6; j = 1,\ldots,6\} \). Define \( X(\omega) = i + j \).

**Example 4.2.** Toss a coin twice. The sample space is \( \Omega = \{HH,HT,TH,TT\} \). Define \( X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0 \).

A random variable creates a probability function on the real line, as follows:

**Definition 4.2.** Let \( \Omega \) be a sample space with probability function \( P \). Let \( X \) be a random variable defined on \( \Omega \). Let \( A \subset \mathbb{R} \). Define \( P(X \in A) = P(\{\omega \in \Omega|X(\omega) \in A\}) \).

## 4.1 Probability mass functions

We begin this section by presenting some useful notation.
**Definition 4.3.** Let $A$ be a set of real numbers. The indicator function on $A$ is the function

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

**Example 4.3.** The function

$$f(x) = \begin{cases} 
2x^{-3} & x < 1 \\
0 & x \geq 1
\end{cases}$$

can be written $f(x) = 2x^{-3}I_{(1,\infty)}(x)$.

**Example 4.4.** The function

$$f(x) = \begin{cases} 
\left(\frac{1}{2}\right)^x & x \in \{1,2,\ldots\} \\
0 & \text{otherwise}
\end{cases}$$

can be written $f(x) = (1/2)^xI_{\{1,2,\ldots\}}(x)$.

**Definition 4.4.** A random variable $X$ is **discrete** if there exists a countable set $S$ of points such that $P(X = x) > 0$ if $x \in S$ and $P(X \in S) = 1$. The points in $S$ are called the **mass points** of $X$.

Note that if $X$ is discrete, then $P(X = x) = 0$ if $x$ is not a mass point of $X$.

**Definition 4.5.** Let $X$ be a discrete random variable. The function $p_X(x) = P(X = x)$ is called the **probability mass function** of $X$. Note that $p_X(x) > 0$ if $x$ is a mass point of $X$, and $p_X(x) = 0$ if $x$ is not a mass point of $X$.

Sometimes $p_X(x)$ is called the **distribution** of $X$.

**Example 4.5.** Toss two fair coins. Let $X$ be the number of heads. Then $p_X(0) = 1/4$, $p_X(1) = 1/2$, $p_X(2) = 1/4$, and $p_X(x) = 0$ for other values of $x$. 
The probability mass function can be represented by a graph in which a vertical line is drawn at each of the possible values of the random variable. The heights of the lines are equal to the probabilities of the corresponding values. The physical interpretation of this graph is that each line represents a mass equal to its height. The following figure presents a graph of the probability mass function of the random variable $X$ that represents the number of heads in two fair coin tosses.

![Probability mass function for a random variable representing the number of heads in two fair coin tosses.]

Example 4.6. A coin has probability $p$ of coming up heads. The coin is tossed until a head appears. Let $X$ be the number of tosses up to and including the first head. Then

$$p_X(x) = p(1 - p)^{x-1}I_{\{1,2,...\}}(x)$$

Example 4.7. An random variable that is not discrete:
A shot hits a target. Let $X$ be the distance from the bull’s eye.

4.2 Expectation of Discrete Random Variables

A die has the number 1 painted on three of its faces, the number 2 on two faces, and the number 3 on one face. The die is rolled a large number $n$ of times. What, approximately, do you expect the average of the $n$ rolls to be?

The possible values for the roll of a die are 1, 2, and 3. The probabilities are $P(X = 1) = 1/2$, $P(X = 2) = 1/3$, and $P(X = 3) = 1/6$.

As the experiment is repeated again and again, the proportion of times that each number is observed will approach its probability. Therefore in a large number $n$ of rolls, we expect
that approximately $n/2$ will be 1s, $n/3$ will be 2s, and $n/6$ will be 3s. The the average will be approximately

$$\frac{(n/2)(1) + (n/3)(2) + (n/6)(3)}{n} = \frac{1(1/2) + 2(1/3) + 3(1/6)}{n} = 1P(X = 1) + 2P(X = 2) + 3P(X = 3)$$

**Definition 4.6.** Let $X$ be discrete with mass points $x_1, \ldots$ and pmf $p_X(x)$. The expectation, or expected value, or mean of $X$ is $E(X) = \sum_i x_i p_X(x_i) = \sum_i x_i P(X = x_i)$ if the sum converges.

**Remark 4.1.** If an experiment that produces a value of a random variable $X$ is repeated over and over again, we can think of $E(X)$ as the long-run average of the outcomes of the experiment.

**Example 4.8.** A certain industrial process is brought down for recalibration whenever the quality of the items produced falls below specifications. Let $X$ represent the number of times the process is brought down during a week, and assume that $X$ has the following probability mass function.

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>0.35</td>
<td>0.25</td>
<td>0.20</td>
<td>0.15</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Find the mean of $X$.

**Solution**

The mean is

$$E(X) = 0(0.35) + 1(0.25) + 2(0.20) + 3(0.15) + 4(0.05) = 1.30$$

The population mean has an important physical interpretation. It is the horizontal component of the center of mass of the probability mass function, that is, it is the point on the horizontal axis at which the graph of the probability mass function would balance if supported there. The following figure illustrates this property for the probability mass function of the number of flaws in a wire, where the population mean is $\mu = 0.66$. 
The graph of a probability mass function will balance if supported at the population mean.

**Remark 4.2.** \( E(X) \) is the center of mass of the pdf of \( X \).

**Theorem 4.1.** Let \( A \) be any event and let \( I_A \) be the indicator of \( A \). In other words, \( I_A = 1 \) if \( A \) occurs and \( I_A = 0 \) if \( A \) does not occur. Then \( E(I_A) = P(A) \).

**Proof:** \( E(I_A) = 1P(I_A = 1) + 0P(I_A = 0) = P(I_A = 1) = P(A) \).

Sometimes we need to compute the expectation of a function of a random variable. Here is an example.

Let \( X \) have the following probability mass function:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( P(X = x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Now let \( Y = X^2 \). We will find \( E(Y) \) by using the definition of expectation. First we find the probability mass function of \( Y \).

<table>
<thead>
<tr>
<th>( y )</th>
<th>( P(Y = y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
</tbody>
</table>
The expectation is $E(Y) = 0(0.1) + 1(0.5) + 4(0.4) = 2.1$.

Here is an easier way, that does not require the computation of the probability mass function of $Y$.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X = x)$</th>
<th>$y = x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>4</td>
</tr>
</tbody>
</table>

Notice that $E(Y) = (-1)^2(0.2) + 0^2(0.1) + 1^2(0.3) + 2^2(0.4) = 2.1$.

This leads to the following theorem.

**Theorem 4.2.** Let $X$ be a discrete random variable with mass points $x_1$, ... and let $Y = g(X)$ be a function of $X$. Then $E(Y) = \sum_i g(x_i)P(X = x_i)$.

**Proof:** The proof is straightforward but tedious. It is omitted.

**Example 4.9.** Let $X$ be the number that turns up on the roll of a fair die. Find $E(|X - 3|)$.

**Solution:** $E(|X - 3|) = \sum_{x=1}^{6} |x-3|P(X = x) = |1-3|(1/6)+|2-3|(1/6)+\cdots+|6-3|(1/6) = 1.5$.

**Theorem 4.3.** Let $X$ be a discrete random variable and let $a$ and $b$ be constants. Then $E(aX + b) = aE(X) + b$.

**Proof:** $E(aX + b) = \sum_x (ax + b)P(X = x) = a \sum_x xP(X = x) + b = aE(X) + b$.

**Theorem 4.4.** Let $X$ be a discrete random variable. Let $g_1(X)$ and $g_2(X)$ be functions of $X$. Then $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$.

**Proof:**

$$
E[g_1(X) + g_2(X)] = \sum_x [g_1(x) + g_2(x)]P(X = x) \\
= \sum_x [g_1(x)P(X = x) + g_2(x)P(X = x)] \\
= \sum_x g_1(x)P(X = x) + \sum_x g_2(x)P(X = x) \\
= E[g_1(X)] + E(g_2(X))
$$
Theorem 4.5. Let $X$ be a discrete random variable whose mass points are non-negative integers. Then $E(X) = \sum_{x=1}^{\infty} P(X \geq x)$.

Proof: First notice that $E(X) = 0P(X = 0) + 1P(X = 1) + \cdots = \sum_{y=1}^{\infty} yP(X = y)$. Now

$$\sum_{x=1}^{\infty} P(X \geq x) = \sum_{x=1}^{\infty} \sum_{y=x}^{\infty} P(X = y) = \sum_{y=1}^{\infty} \sum_{x=1}^{y} P(X = y) = \sum_{y=1}^{\infty} yP(X = y) = E(X).$$

Example 4.10. A coin has probability $p$ of landing heads. It is tossed repeatedly. Let $X$ be the number of tosses before the first tail appears. Find $E(X)$.

$P(X \geq x) = P(\text{First}\ x\ \text{tosses are heads}) = p^x$. Therefore $E(X) = \sum_{x=1}^{\infty} p^x = \frac{p}{1 - p}$.

Theorem 4.6. Markov’s inequality: If $X$ is a nonnegative random variable, then for all $a > 0$, $P(X \geq a) \leq \frac{E(X)}{a}$.

Proof:

$$E(X) = \sum_{x=0}^{\infty} xP(X = x) \geq \sum_{x=a}^{\infty} xP(X = x) \geq \sum_{x=a}^{\infty} aP(X = x) = a \sum_{x=a}^{\infty} P(X = x) = aP(X \geq a)$$

Example 4.11. On an exam 80% of the class scored 70 or more. What is the minimum possible value for the mean score?

Solution: Let $X$ be the exam score of a randomly chosen student. We know that $P(X \geq 70) = 0.8$. By Markov’s inequality, $0.8 \leq E(X)/70$. Solving for $E(X)$ we obtain $E(X) \geq 70(0.8) = 56$. The mean can be no less than 56.
Example 4.12. In a certain town the mean household income is $40,000. What is the maximum possible proportion of households with incomes of $100,000 or more?

Solution: Let $X$ be the income of a randomly chosen household. We know that $E(X) = 40,000$. By Markov’s inequality, $P(X \geq 100,000) \leq \frac{40,000}{100,000} = 0.4$.

4.3 Variance and Standard Deviation of Discrete Random Variables

Definition 4.7. Let $X$ be a random variable with $E(X) = \mu$. The variance of $X$ is $V(X) = E[(X - \mu)^2]$, if the expectation exists.

It follows that if $X$ is discrete then $V(X) = \sum_x(x - \mu)^2 P(X = x)$

Theorem 4.7. Let $X$ be a random variable with $E(X) = \mu$. Let $a$ be any constant. Then the quantity $E[(X - a)^2]$ is minimized when $a = \mu$.

Proof: $E[(X - a)^2] = E[(X - \mu + \mu - a)^2] = E[(X - \mu)^2] + 2(\mu - a)E[X - \mu] + E[(\mu - a)^2] = E[(X - \mu)^2] + (\mu - a)^2$. This is minimized when $(\mu - a)^2 = 0$, or when $\mu = a$.

Definition 4.8. Let $X$ be a random variable such that $V(X)$ exists. The standard deviation of $X$ is $\sqrt{V(X)}$.

There is an alternative form for the variance that is often easier to calculate.

Theorem 4.8. $V(X) = E(X^2) - [E(X)]^2$.

Proof: Let $\mu = E(X)$. Now

\[
V(X) = E[(X - \mu)^2] = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2
\]

Corollary: If $E(X) = 0$ then $V(X) = E(X^2)$. 

34
Example 4.13. Let $X$ be the number on the roll of a fair die. Find $V(X)$.

Theorem 4.9. Let $X$ be a random variable, and $a$ a constant. Then $V(aX) = a^2V(X)$.

Proof: First note that $E(aX) = aE(X)$. Now

$$V(aX) = E[(aX)^2] - [E(aX)]^2 = E(a^2X^2) - [aE(X)]^2 = a^2E(X^2) - a^2[E(X)]^2 = a^2V(X)$$

Theorem 4.10. Let $X$ be a random variable, and $b$ a constant. Then $V(X + b) = V(X)$.

Proof: Let $Y = X + b$. We show that $V(Y) = V(X)$.

Now $E(Y) = E(X) + b$, so $Y - E(Y) = (X + b) - (E(X) + b) = X - E(X)$.

Theorem 4.11. Let $X$ be a random variable, and $b$, $a$ a constant.

Then $V(aX + b) = a^2V(X)$.

Proof: $V(aY + b) = V(aY) = a^2V(Y)$.

Corollary: Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$. Let $Y = (X - \mu)/\sigma$. Then $E(Y) = 0$ and $V(Y) = E(Y^2) = 1$.

Proof: $Y = (1/\sigma)X - \mu/\sigma$. Therefore $E(Y) = (1/\sigma)E(X) - \mu/\sigma = \mu/\sigma - \mu/\sigma = 0$, and $V(Y) = (1/\sigma)^2V(X) = 1$.

Theorem 4.12. Chebychev’s Inequality: Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$. Let $k > 0$. Then $P(|X - \mu| > k\sigma) < \frac{1}{k^2}$.

Proof: Let $Z = \frac{X - \mu}{\sigma}$. Then $E(Z) = 0$ and $E(Z^2) = 1$.

Now $P(|X - \mu| > k\sigma) = P(Z > k) = P(Z^2 > k^2) < \frac{E(Z^2)}{k^2} = \frac{1}{k^2}$.

5 Common Parametric Families of Discrete Distributions

In this section we describe some commonly used families of discrete distributions.
5.1 The Discrete Uniform Distribution

Let $N_0 < N_1$ be integers. A random variable $X$ has the **discrete uniform distribution** on the integers $N_0, \ldots, N_1$ if the pmf of $X$ is

$$p_X(x) = \frac{1}{N_1 - N_0 + 1} I_{\{N_0, \ldots, N_1\}}(x)$$

It can be verified that

$$E(X) = \frac{N_0 + N_1}{2}, \quad V(X) = \frac{(N_1 - N_0 + 1)^2 + 1}{12}$$

Most commonly $N_0 = 0$. Then

$$E(X) = \frac{N_1}{2}, \quad V(X) = \frac{N_1^2 + 1}{12}$$

5.2 The Bernoulli Distribution

Let $0 < p < 1$. A random variable $X$ has the **Bernoulli distribution** with parameter $p$ if the pmf of $X$ is

$$f(x; p) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

The pmf may also be written

$$f(x; p) = p^x(1 - p)^{1-x} I_{\{0,1\}}(x)$$

It is easy to verify that

$$E(X) = p, \quad V(X) = p(1 - p)$$

A random variable with a Bernoulli distribution is often called a **Bernoulli trial**. If $X = 1$, the trial is said to result in success, if $X = 0$, the trial is said to end in failure. The probability $p$ is sometimes called the “success probability.” The toss of a coin is the archetypal Bernoulli trial.
5.3 The Binomial Distribution

Let \( n \) be a positive integer, and let \( 0 < p < 1 \). A random variable \( X \) has the **binomial distribution** with parameters \( n \) and \( p \) if the pmf of \( X \) is

\[
f(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x} I_{\{0,1,...,n\}}(x)
\]

If \( X \) is distributed binomial with parameters \( n \) and \( p \), we will write \( X \sim \text{Bin}(n,p) \).

Intuitively, a binomial random variable counts the number of successes in \( n \) independent Bernoulli trials, all of which have the same success probability. To see this, imagine that \( n \) independent Bernoulli trials \( X_1, ..., X_n \) are conducted. Let \( X \) be the number that result in success. Let \( 0 \leq x \leq n \). We compute \( P(X = x) \). First we compute

\[
P(X_1 = 1, X_2 = 1, ..., X_x = 1, X_{x+1} = 0, ..., X_n = 0) = p^x (1 - p)^{n-x}.
\]

Now there are \( \binom{n}{x} \) ways to choose the \( x \) trials that will be successes, and all of them have probability \( p^x (1 - p)^{n-x} \). Therefore \( P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \).

It may be verified that

\[
E(X) = np \quad V(X) = np(1 - p)
\]

Intuitively, we can think of \( X \) as the sum of \( n \) independent Bernoulli trials, all with the same success probability, that is, \( X = X_1 + \cdots + X_n \). Note that \( E(X) = E(X_1) + \cdots E(X_n) \) and \( V(X) = V(X_1) + \cdots V(X_n) \).

5.4 The Hypergeometric Distribution

An urn contains \( N \) balls, \( M \) of which are red, and \( N - M \) of which are green. \( K \) balls are drawn without replacement. Let \( X \) be the number of red balls drawn. Find \( P(X = x) \).

\[
P(X = x) = \frac{\text{number of combinations of } K \text{ balls that contain } x \text{ red balls}}{\text{number of combinations of } K \text{ balls chosen from } N}
\]

The number of combinations of \( K \) balls chosen from \( N \) is \( \binom{N}{K} \). To compute the numerator, notice that the numerator consists of two combinations; one of \( x \) red balls chosen from \( M \),
and one of \( K - x \) green balls chosen from \( N - M \). The number of ways to choose \( x \) red balls from \( M \) red balls is \( \binom{M}{x} \), and the number of ways to choose \( K - x \) green balls from \( N - M \) red balls is \( \binom{N-M}{K-x} \). The total number of ways to make both these choices is 
\[
\binom{M}{x} \binom{N-M}{K-x}.
\]

Therefore
\[
P(X = x) = \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}
\]

We now define the hypergeometric distribution.

**Definition 5.1.** Let \( N, M, K \) be positive integers with \( M \leq N \) and \( K \leq N \). A random variable \( X \) has the hypergeometric distribution with parameters \( N, M, \) and \( K \) if the pmf of \( X \) is

\[
p_X(x) = \begin{cases} 
\frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}} & \text{if } x \text{ is an integer, } \max[0, K - (N - M)] \leq x \leq \min(K, M) \\
0 & \text{otherwise}
\end{cases}
\]

If \( X \) is distributed hypergeometric with parameters \( N, M, \) and \( K \), we will write \( X \sim H(N, M, K) \).

It can be shown that \( E(X) = K \frac{M}{N} \), and \( V(X) = K \frac{M}{N} \left( 1 - \frac{M}{N} \right) \frac{N - K}{N - 1} \).

We can compare the hypergeometric distribution to the binomial distribution. If the balls are drawn with replacement, then \( X \sim \text{Bin}(K, M/N) \). The mean and variance of \( X \) are then \( E(X) = K(M/N) \), and \( V(X) = K(M/N)(1 - M/N) \). Note that the mean number of successes is the same whether or not the draws are made with replacement. The variance is smaller when the draws are made without replacement. The variance without replacement is found by multiplying the variance with replacement by the factor \( (N - K)/(N - 1) \). This factor is sometimes called the **finite population correction factor**.
5.5 The Poisson Distribution

Definition 5.2. The random variable $X$ has the Poisson distribution with parameter $\lambda > 0$ if the pmf of $X$ is

$$f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!} I_{\{0,1,...\}}(x)$$

If $X$ is distributed Poisson with parameter $\lambda$, we will write $X \sim \text{Poisson}(\lambda)$.

Note that the Poisson($\lambda$) distribution puts a probability on an integer $x$ that is proportional to the $x$th term in the Taylor series expansion of $e^{-\lambda}$.

Theorem 5.1. Let $X \sim \text{Poisson}(\lambda)$. Then $E(X) = \lambda$ and $V(X) = \lambda$.

Proof:

$$E(X) = \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} = \lambda \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^{x-1}}{(x-1)!} = \lambda \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} = \lambda(1) = \lambda$$

To compute the variance we first compute $E(X^2)$.

$$E(X^2) = \sum_{x=0}^{\infty} x^2 e^{-\lambda} \frac{\lambda^x}{x!}$$

$$= \sum_{x=2}^{\infty} x(x-1)e^{-\lambda} \frac{\lambda^x}{x!} + \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!}$$

$$= \sum_{x=2}^{\infty} e^{-\lambda} \frac{\lambda^x}{(x-2)!} + \lambda$$

$$= \lambda^2 \sum_{x=0}^{\infty} e^{-\lambda} \frac{\lambda^x}{x!} + \lambda$$

$$= \lambda^2 + \lambda$$

Now $V(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

One use of the Poisson distribution is as an approximation to the binomial distribution. If $X \sim \text{Bin}(n, p)$ with $n$ large and $p$ small, then $X$ is distributed approximately Poisson with parameter $\lambda = np$. The precise statement of the result is as follows:
**Theorem 5.2.** Let \( \lambda > 0 \). For each integer \( n \), let \( X_n \) be a random variable distributed \( \text{Bin}(n, \lambda/n) \). Let \( x \) be a nonnegative integer. Then \( \lim_{n \to \infty} P(X_n = x) = e^{-\lambda} \lambda^x / x! \).

**Proof:** 

\[
P(X_n = x) = \binom{n}{x} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x}.
\]

We therefore must show that \( \lim_{n \to \infty} \binom{n}{x} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x} = e^{-\lambda} \frac{\lambda^x}{x!} \). The proof relies on the fact that \( \lim_{n \to \infty} \left( 1 + \frac{1}{n} \right) = e \), and for any constant \( \lambda \), \( \lim_{n \to \infty} \left( 1 + \frac{\lambda}{n} \right) = e^\lambda \).

\[
\lim_{n \to \infty} \frac{n(n-1) \cdots (n-x+1)}{n^x} \left( \frac{1}{n} \right)^x \left( \frac{1 - \lambda}{n} \right)^{n-x} = \lim_{n \to \infty} \binom{n}{x} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x} \frac{\lambda^x}{x!}
\]

\[
= (1)(1)e^{-\lambda} \frac{\lambda^x}{x!}
\]

One of the first applications of the Poisson distribution was to characterize the distribution of the number of particles in a sampled volume of suspension. The argument, first presented by Student, is given below.

A large volume of suspension contains particles, thoroughly mixed, at a concentration of \( \lambda \) per unit volume. Let \( N \) be the total volume of suspension, so that there are a total of \( N\lambda \) particles. One unit of volume is withdrawn. Let \( X \) be the number of particles withdrawn. Think of each particle as a Bernoulli trial, with success occurring if the particle is withdrawn. Since the proportion of the total volume withdrawn is \( 1/N \), each particle has probability \( 1/N \) of being withdrawn. The distribution of \( X \) is therefore \( \text{Bin}(N\lambda, 1/N) \). Since \( 1/N \) is small, \( X \sim \text{Poisson}(\lambda) \) approximately.

**Example 5.1.** Grandma makes chocolate chip cookies in batches of 100. She adds 300 chips to the dough. You get one cookie. What is the probability that your cookie has no chocolate chips?

**Solution:** Your cookie is 1% of the batch, so the probability that any given chip ends up in
your cookie is 0.01. Now let $X$ be the number of chips in your cookie. $X \sim \text{Bin}(300, 0.01)$ so $X \sim \text{Poisson}(3)$ approximately. $P(X = 0) = e^{-3} = 0.0498$.

**Example 5.2.** Refer to Example 5.1. What is the probability that your cookie contains fewer than three chips?

**Solution:** Let $X$ be the number of chips in your cookie.

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = e^{-3} + e^{-3} \frac{3^1}{1!} + e^{-3} \frac{3^2}{2!} = 0.4239.$$ 

**Example 5.3.** Grandma makes chocolate chip cookies in batches of 100. How many chips should she put in the dough so that the probability that cookie contains no chips is 0.02?

**Solution:** Let $n$ be the required number of chips, and let $X$ be the number of chips in your cookie. Then $X \sim \text{Bin}(n, 0.01)$, so $X \sim \text{Poisson}(0.01n)$ approximately. $P(X = 0) = e^{-0.01n} = 0.02$, so $n = 100 \log 50 = 391.2$. Round to 392.

Consider a radioactive mass containing a large number, $n$, of atoms. Let $p$ be the probability that any given atom emits a particle during a given unit of time. Choose the unit of time small enough so that $p$ is small. Let $X$ be the number of particles emitted during one unit of time. We will find the distribution of $X$.

Think of each atom as a Bernoulli trial, with success probability $p$. Then $X \sim \text{Bin}(n, p)$. The mean number of particles emitted is $E(X) = np$. Let $\lambda = np$. Then because $p$ is small, $X \sim \text{Poisson}(\lambda)$.

**Example 5.4.** A radioactive mass emits particles at a mean rate of 4 per second. Find the probability that 5 particles are emitted during a given second.

**Solution:** Let $X$ be the number of particles emitted during a given second. Then $X \sim \text{Poisson}(4)$. $P(X = 5) = e^{-4} \frac{4^5}{5!} = 0.1563$.

**Example 5.5.** Refer to Example 5.4. Find the probability that the mass emits 10 particles during a three-second period.
Solution: The mean number of particles emitted during a time unit of three seconds is $3(4) = 12$. Let $X$ be the number of particles emitted in 3 seconds. Then $X \sim \text{Poisson}(12)$, so $P(X = 10) = e^{-10}\frac{10^{12}}{10!} = 0.1048$.

Example 5.5 illustrates that for a radioactive mass emitting particles at a mean rate of $\lambda$ per time unit, the number of particles emitted in $t$ time units will be distributed $\text{Poisson}(\lambda t)$. Another fact of note about radioactive decay is that the numbers of events in disjoint intervals are independent. This is true because the atoms act independently of each other.

The process of radioactive decay has two important stochastic properties: The number of events occurring in a time interval of length $t$ is distributed $\text{Poisson}(\lambda t)$, and the numbers of events occurring in disjoint time intervals are independent. We call a process that satisfies these two properties a **Poisson process**.

**Definition 5.3.** Events are occurring in time or space. Let $X$ be the number of events that occur in an interval of length $t$. If there exists $\lambda > 0$ such that $X \sim \text{Poisson}(\lambda t)$ for all $t$, then the events are said to follow a Poisson process with rate $\lambda$.

**Example 5.6.** Traffic accidents occur at a certain intersection according to a Poisson process with rate 1.2 per month. What is the probability that there are exactly three accidents in a two-month period?

**Solution:** Let $X$ be the number of accidents occurring in a two-month period. Then $X \sim \text{Poisson}(2.4)$. $P(X = 3) = e^{-2.4}\frac{2.4^3}{3!} = 0.2090$.

**Example 5.7.** Refer to Example 5.6. What is the probability that there are 2 accidents in June and 1 accident in September?

**Solution:** Let $X$ be the number of accidents in June and let $Y$ be the number of accidents in September. Then $X$ and $Y$ are both distributed $\text{Poisson}(1.2)$. Because the numbers of events in disjoint time intervals are independent, $X$ and $Y$ are independent. Therefore $P(X = 2 \cap Y = 1) = P(X = 2)P(Y = 1) = \left(e^{-1.2}\frac{1.2^2}{2!}\right)\left(e^{-1.2}\frac{1.2^1}{1!}\right) = (0.2612)(0.2177) = 0.0569$.

Example 5.8 presents a Poisson process in space.
Example 5.8. The number of plants of a certain species in a forest has a Poisson distribution with a mean of 10 plants per acre. What is the probability that there will be exactly 12 plants in a circle with radius 100 ft$^2$? (Note that 1 acre = 43,560 ft$^2$).

Solution: The area of the circle is $10,000\pi = 31,416$ ft$^2 = 0.721$ acres. Let $X$ be the number of plants in the circle. $X \sim \text{Poisson}(7.21)$ so $P(X = 12) = e^{-7.21} \frac{7.21^{12}}{12!} = 0.0304$.

Acquired Traits are not Inherited: A Theory Proved by Using the Poisson Distribution

The Poisson distribution was instrumental in settling a fundamental scientific controversy that had lasted for nearly a century. This was the controversy over whether acquired characteristics can be inherited.

The process of evolution, over time, produces organisms that are adapted to their environments. The process of adaptation occurs through genetic mutations that produce traits that enhance survival. There are potentially two mechanisms by which this could occur. First, it is possible that environmental forces may cause mutations which render organisms more adaptable. Acquired characteristics are thereby inherited. Alternatively, it is possible that mutations occur at random, and those organisms who happen to get favorable mutations are more likely to survive.

Charles Darwin argued for the second hypothesis. Other prominent biologists, such as Lamarck, argued for the first. The question was finally resolved in 1943, when biologists Salvador Luria and Max Delbrück conducted a series of experiments and used the Poisson distribution to prove that Darwin was right; that acquired characteristics are not inherited. The experiments are described below. For this and related work, Luria and Delbrück, along with their colleague Alfred Hershey, were awarded the Nobel Prize in 1969.

Luria and Delbrück worked with bacteria. It was well known that if a colony of bacteria is exposed to a virus, a few bacteria will survive. These bacteria are “resistant” to the virus. It is also the case that the descendants of resistant bacteria are resistant. The two theories regarding the mechanism of resistance were:

1) Acquired immunity: Under this hypothesis, a few bacteria survive the virus due to random physiologic variations. Their interaction with the virus causes a hereditary change, which confers resistance to their offspring. This is the theory that says that the environment causes mutations that aid adaptability.
2) *Mutation to immunity:* The resistant bacteria had inherited a genetic mutation which conferred resistance. This mutation was not caused by interaction with the virus. This it the theory that says that mutations are random, and those who happen to get a favorable mutation are the ones who survive to reproduce.

The basic question, therefore, is whether the interaction with the virus confers heritable immunity on those bacteria that survive (acquired immunity), or whether these bacteria had inherited their immunity before the exposure (mutation to immunity). Luria and Debrück conducted two types of experiments.

The first experiment was a control. It involved plating random samples of bacteria, approximately 10 ml in volume, onto an agar plate that contained a large quantity of virus. This killed nearly all the bacteria. After 24 hours or so, small colonies of bacteria appeared. Each colony consisted of the descendants of a resistant bacterium that had survived the virus. By counting the number of colonies, the number of resistant bacteria in the original sample could be determined.

Here are the numbers of resistant bacteria found in a series of control experiments:

Control: 14, 15, 13, 21, 15, 14, 26, 16, 20, 13

In the second set of experiments, the bacteria to be plated were obtained in a different way. Instead of selecting a random sample of bacteria, a small number of founders (50–500) was used to grow a colony. When the colony grew to 10 ml in volume, it was plated onto an agar plate that had been treated with virus, just as in the control experiment. Here are the results of four sets of these experiments.

Experiment 1: 10, 18, 125, 10, 14, 27, 3, 17, 17
Experiment 2: 29, 41, 17, 20, 31, 30, 7, 17
Experiment 3: 30, 10, 40, 45, 183, 12, 173, 23, 57, 51
Experiment 4: 6, 5, 10, 8, 24, 13, 165, 15, 6, 10

The only difference between the control and the treatment was that in the treatment experiment, all the bacteria were descendants of a few founders, while in the control experiment, the bacteria could be considered to be a large sample from the general population of bacteria.

Means and variances of the numbers of resistant bacteria were computed for each experiment. Here are the results:
<table>
<thead>
<tr>
<th>Experiment</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>16.7</td>
<td>18.23</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>26.78</td>
<td>1401</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>24</td>
<td>114.57</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>62.4</td>
<td>3958.7</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>26.2</td>
<td>2410</td>
</tr>
</tbody>
</table>

In the control experiment, the mean and variance were similar. This would be expected. Each bacterium in the sample has a very small probability of being resistant, so the number of resistant bacteria in the sample has a binomial distribution with a large \( n \) and a small \( p \). Therefore the number of resistant bacteria will have a Poisson distribution, so the mean should be equal to the variance.

Under the assumption of acquired immunity, the number of bacteria in the treatment experiment should also have followed a Poisson distribution. The reason is that under this assumption, immunity does not occur until exposure to the virus. Therefore none of the founders, or their descendants, would have been resistant. As in the control experiment, then, each bacterium would have had a very small probability of surviving, which would cause the number of survivors to follow a Poisson distribution.

However, when the colonies of bacteria were grown from a small number of founders, the mean was always much smaller than the variance. Therefore it can be concluded that the number of resistant bacteria in such a colony did not follow a Poisson distribution. We conclude that the acquired immunity hypothesis is false.

The results of the treatment experiments are consistent with the mutation to immunity hypothesis. Because mutations to immunity are very rare, in most cases none of the founders will be resistant, and no mutations to resistance will occur for many generations. In these cases the total number of resistant bacterial will be small. Occasionally, however, a mutation will occur in a relatively early generation, in which case the total number of bacteria will be large. The result is that the variance of the number of resistant bacteria is much larger than the mean.

### 5.6 The Geometric Distribution

A coin has probability \( p \) of landing heads. The coin is tossed repeatedly. Let \( X \) be the number of tosses up to and including the first head. Find the pmf of \( X \).
\[ P(X = x) = P(x - 1 \text{ tails followed by a head}) = p(1 - p)^{x-1}. \]

**Definition 5.4.** A random variable \( X \) has the geometric distribution with parameter \( p \), \( 0 < p < 1 \), if the pmf of \( X \) is

\[
f(x; p) = p(1 - p)^{x-1} I_{\{1, 2, \ldots\}}(x)
\]

We write \( X \sim \text{Geom}(p) \).

We interpret \( X \) as the number of trials up to and including the first success in an independent sequence of Bernoulli trials.

We note that \( P(X > x) = P(x \text{ failures in a row}) = (1 - p)^x \), and \( P(X \geq x) = P(x - 1 \text{ failures in a row}) = (1 - p)^{x-1} \).

**Theorem 5.3.** Let \( X \sim \text{Geom}(p) \). Then \( E(X) = \frac{1}{p} \) and \( V(X) = \frac{1 - p}{p^2} \).

**Proof:**

\[
E(X) = \sum_{x=1}^{\infty} xp(1 - p)^{x-1} = p \sum_{x=1}^{\infty} x(1 - p)^{x-1}
\]

\[
= -p \sum_{x=1}^{\infty} \frac{d}{dp} (1 - p)^x = -p \frac{d}{dp} \sum_{x=1}^{\infty} (1 - p)^x = -p \frac{d}{dp} \left( \frac{1 - p}{p} \right) = \frac{1}{p}
\]

To compute the variance we first compute \( E(X^2) \).

\[
E(X^2) = \sum_{x=1}^{\infty} x^2 p(1 - p)^{x-1} = p(1 - p) \sum_{x=1}^{\infty} x(x - 1)(1 - p)^{x-2} + \sum_{x=1}^{\infty} xp(1 - p)^{x-1}
\]

\[
= p(1 - p) \sum_{x=1}^{\infty} \frac{d^2}{dp^2} (1 - p)^x + \frac{1}{p} = p(1 - p) \frac{d^2}{dp^2} \sum_{x=1}^{\infty} (1 - p)^x + \frac{1}{p}
\]

\[
= p(1 - p) \frac{d^2}{dp^2} \left( \frac{1 - p}{p} \right) + \frac{1}{p} = p(1 - p) \left( \frac{2}{p^3} \right) + \frac{1}{p} = \frac{2(1 - p)}{p^2} + \frac{1}{p}
\]

\[
V(X) = \frac{2(1 - p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1 - p}{p^2}.
\]
**Theorem 5.4. Memoryless Property:** Let $X \sim \text{Geom}(p)$. Then for any positive integers $t$ and $s$, $P(X > t + s \mid X > t) = P(X > s)$.

**Proof:** For any positive integer $x$, $P(X > x) = (1 - p)^x$. Now

\[
P(X > t + s \mid X > t) = \frac{P(X > t + s \cap X > t)}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} = \frac{(1 - p)^{t+s}}{(1 - p)^t} = p(1 - p)^s.
\]

The converse of the theorem is also true.

**Theorem 5.5.** Let $X$ be a random variable whose mass points are positive integers, and for which $P(X > t + s \mid X > t) = P(X > s)$ for all positive integers $t$ and $s$. Then $X \sim \text{Geom}(p)$, where $p = P(X = 1)$.

**Proof:**

\[
P(X > t + s \mid X > t) = \frac{P(X > t + s)}{P(X > t)}.\]

Therefore $P(X > s) = \frac{P(X > t + s)}{P(X > t)}$, so $P(X > t + s) = P(X > t)P(X > s)$.

We must show that for any positive integer $n$, $P(X > n) = (1 - p)^n$. Since $P(X > 1) = 1 - p$, this is true for $n = 1$. Now assume $P(X = k) = (1 - p)^k$.

Then $P(X > k + 1) = P(X > k)P(X > 1) = (1 - p)^k(1 - p) = (1 - p)^{k+1}$.

### 5.7 The Negative Binomial Distribution

Let $r$ be a positive integer. A coin that lands heads with probability $p$ is tossed until $r$ heads are obtained. Let $X$ be the number of tosses up to and including the $r$th head. Find the pmf of $X$.

If $X = x$, then there were $x - r$ tails. The last toss was a head, and the first $x - 1$ tosses consisted of $r - 1$ heads and $x - r$ tails in any order. The probability of any particular sequence of $r$ heads and $x - r$ tails is $p^r(1 - p)^{x-r}$. The number of such sequences in which the last toss is a head is \(\binom{x-1}{r-1}\). It follows that
Definition 5.5. A random variable $X$ has the negative binomial distribution with parameters $r$ and $p$ if the pmf of $X$ is

$$f(x; r, p) = \binom{x - 1}{r - 1} p^r (1 - p)^{x-r} I_{\{r,r+1,r+2,...\}}$$

We write $X \sim \text{NB}(r, p)$. $r$ must be a positive integer, and $0 < p < 1$.

If $X \sim \text{NB}(r, p)$, then $X$ can be thought of as a sum of $r$ Geom($p$) random variables.

Theorem 5.6. Let $X \sim \text{NB}(r, p)$. Then $E(X) = \frac{r}{p}$ and $V(X) = \frac{r(1-p)}{p^2}$.

6 Continuous Random Variables

Definition 6.1. A random variable $X$ is said to be absolutely continuous if there exists a function $f(x)$ such that

$$f(x) \geq 0 \text{ for all } x$$

and, for all $-\infty \leq a \leq b \leq \infty$

$$P(a \leq X \leq b) = \int_a^b f(t) \, dt$$

A function $f_X(x)$ satisfying these conditions is said to be a probability density function (pdf) for the random variable $X$.

Example 6.1. Let $X$ be continuous with pdf

$$f(x) = 3e^{-3x} I_{(0, \infty)}(x)$$

Find a) $P(X \leq 2)$ b) $P(X > 0.5)$ c) $P(X < -3)$ d) $P(1 < X \leq 2)$

Solution:

a) $P(X \leq 2) = \int_{-\infty}^{2} f(x) \, dx = \int_{0}^{2} 3e^{-3x} \, dx = 1 - e^{-6}$

b) $P(X > 0.5) = \int_{0.5}^{\infty} 3e^{-3x} \, dx = e^{-1.5}$.

c) $P(X < -3) = \int_{-\infty}^{-3} f(x) \, dx = \int_{-\infty}^{-3} 0 \, dx = 0$.

d) $P(1 < X \leq 2) = \int_{1}^{2} 3e^{-3x} \, dx = e^{-3} - e^{-6}$.  

48
Theorem 6.1. If \( f_X(x) \) is the pdf of a random variable \( X \), then \( \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \).

Proof: \( \int_{-\infty}^{\infty} f_X(x) \, dx = \lim_{x \to \infty} \int_{-\infty}^{x} f_X(t) \, dt = \lim_{x \to \infty} F_X(x) = 1 \).

Theorem 6.2. Let \( X \) be a continuous random variable. Then \( P(X = x) = 0 \) for all \( x \).

Proof: \( P(X = x) = \int_{-\infty}^{a} f(x) \, dx = 0 \).

Corollary: Let \( X \) be a continuous random variable, and let \( a < b \).
Then \( P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_{a}^{b} f_X(x) \, dx \).

Definition 6.2. Let \( X \) be a random variable. The cumulative distribution function (cdf) of \( X \) is \( F(x) = P(X \leq x) \).

If \( X \) is continuous with pdf \( f_X(x) \), then \( F(x) = \int_{-\infty}^{x} f_X(t) \, dt \).

Theorem 6.3. Let \( X \) be continuous with cdf \( F(x) \). Then the function \( f(x) = F'(x) \) is a pdf of \( X \).

Remark 6.1. The derivative of \( F(x) \) will exist everywhere or almost everywhere. The pdf \( f(x) \) can be defined arbitrarily at any points where \( F'(x) \) does not exist.

Definition 6.3. Let \( X \) be continuous with pdf \( f_X(x) \). The expectation of \( X \) is \( E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx \), if the integral exists.

Remark 6.2. The expectation may also be called the expected value, or the mean.

Example 6.2. Let \( X \) have pdf \( f(x) = 3e^{-3x}I_{(0,\infty)}(x) \). Find \( E(X) \).

Solution: \( E(X) = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{0}^{\infty} 3xe^{-3x} \, dx = 1/3 \).

Definition 6.4. Let \( X \) be continuous with cdf \( F(x) \). Let \( 0 < q < 1 \). The \( q \)th quantile of \( X \) is the number \( x_q \) such that \( P(X \leq x_q) = q \). Equivalently, \( \int_{-\infty}^{x_q} f(x) \, dx = q \).

The median of \( X \) is the number \( x_m \) such that \( P(X \leq x_m) = 0.5 \).
Example 6.3. Let $X$ be continuous with pdf

$$f(x) = 3e^{-3x}I_{(0,\infty)}(x)$$

a) Find the median of $X$.  

b) Find the 0.9 quantile of $X$.

Solution:

a) The median $x_m$ satisfies $P(X \leq x_m) = 0.5$. Therefore we solve $\int_{-\infty}^{x_m} f(x) \, dx = 0.5$.

$$\int_{-\infty}^{x_m} f(x) \, dx = \int_{0}^{x_m} 3e^{-3x} \, dx = 1 - e^{-3x_m} = 0.5.$$  

We find that $x_m = 0.231$.

b) The 0.9 quantile $x_{0.9}$ (also called the 90th percentile) solves $P(X \leq x_{0.90}) = 0.9$. Therefore we solve $\int_{-\infty}^{x_{0.90}} f(x) \, dx = 0.9$.

$$\int_{-\infty}^{x_{0.90}} f(x) \, dx = \int_{0}^{x_{0.90}} 3e^{-3x} \, dx = 1 - e^{-3x_{0.90}} = 0.9.$$  

We find that $x_{0.90} = 0.768$.

Definition 6.5. Let $X$ be continuous with pdf $f_X(x)$ and mean $\mu = E(X)$. The variance of $X$ is

$$V(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) \, dx$$

if the integral exists.

Note that an alternative expression is $V(X) = E(X^2) - E(X)^2$.

7 Common Parametric Families of Continuous Distributions

In this section we describe some commonly used families of continuous distributions.

7.1 The Uniform Distribution

Definition 7.1. A random variable $X$ has the uniform distribution on $(a,b)$ if the pdf of $X$ is

$$f_X(x) = \frac{1}{b-a}I_{(a,b)}(x)$$

We write $X \sim U(a,b)$.

We show that $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$.
The mean is found by computing the following integral.

$$E(X) = \int_a^b x \frac{1}{b-a} \, dx = \frac{a+b}{2}$$

Finding the variance is a bit more complicated. First we find $E(X^2)$.

$$E(X^2) = \int_a^b x^2 \frac{1}{b-a} \, dx = \frac{b^3-a^3}{3(b-a)}$$

Now $V(X) = E(X^2) - E(X)^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}$.

Often $a = 0$ and $b = 1$. The $U(0,1)$ distribution is called the **standard uniform distribution**.

### 7.2 The Normal Distribution

**Definition 7.2.** A random variable $X$ has the **normal distribution** with parameters $\mu$ and $\sigma$ if the pdf of $X$ is

$$f(x; \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We write $X \sim N(\mu, \sigma^2)$. $\mu$ may be any real number, and $\sigma > 0$.

**Remark 7.1.** The normal distribution is sometimes called the **Gaussian distribution**.

We now show that this pdf is in fact a pdf.

We show that $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = 1$. Make the substitution $u = \frac{x-\mu}{\sigma}$. Then

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du$$

Let $A = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du$. Then

$$A^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} \, dv = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{u^2+v^2}{2}} \, du \, dv$$

We show that $A^2 = 1$. Since clearly $A > 0$ this implies $A = 1$. Transform to polar coordinates: $u = r \cos \theta$, $v = r \sin \theta$. Then

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{u^2+v^2}{2}} \, du \, dv = \int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} e^{-\frac{r^2}{2}} \, r \, dr \, d\theta = \int_0^{2\pi} \frac{1}{2\pi} d\theta \left( e^{-r^2} \bigg|_0^\infty \right) = \int_0^{2\pi} \frac{1}{2\pi} \, d\theta = 1$$
Theorem 7.1. Let \( X \sim N(\mu, \sigma^2) \). Then \( E(X) = \mu \) and \( V(X) = \sigma^2 \).

Proof: \( E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \, dx \). Make the substitution \( z = \frac{x - \mu}{\sigma} \). Then
\[
E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (\sigma z + \mu) e^{-z^2/2} \, dz = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-z^2/2} \, dz + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz = 0 + \mu(1) = \mu
\]

\( V(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \, dx \). Make the substitution \( z = \frac{x - \mu}{\sigma} \). Then
\[
V(X) = \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 e^{-z^2/2} \, dz = \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z^2 e^{-z^2/2} \, dz = \sigma^2(1) = \sigma^2
\]

Let \( X \sim N(\mu, \sigma^2) \). It can be shown that \( E[(X - \mu)^3] = 0 \) and \( E[(X - \mu)^4] = 3\sigma^4 \).

For any normal distribution \( P(|X - \mu| \leq \sigma) \approx 0.68 \), \( P(|X - \mu| \leq 2\sigma) \approx 0.95 \), and \( P(|X - \mu| \leq 3\sigma) \approx 0.997 \).

Definition 7.3. If \( Z \sim N(0, 1) \), then \( Z \) is said to have a **standard normal distribution**.

Definition 7.4. The pdf of a standard normal random variable is denoted \( \phi(x) \). The cdf is denoted \( \Phi(x) \). So \( \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \) and \( \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \, dy \).

Theorem 7.2. Let \( Z \sim N(0,1) \), and let \( a \) and \( b \) be real numbers with \( b \neq 0 \). Then \( a + bZ \sim N(a,b^2) \).

Proof: Let \( Y = a + bX \). First we assume \( b > 0 \).
\[
f_Y(y) = \frac{d}{dy} P(a + bX \leq y) = \frac{d}{dy} P \left( X \leq \frac{y-a}{b} \right) = \frac{d}{dy} \int_{-\infty}^{\frac{y-a}{b}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz
\]
\[
= \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{y-a}{2b}\right)^2} \frac{d}{dy} \frac{y-a}{b} = \frac{1}{b\sqrt{2\pi}} e^{-\left(\frac{y-a}{2b}\right)^2}
\]

which is the pdf of \( N(a,b^2) \). The proof for the case \( b < 0 \) is similar.

Corollary: Let \( X \sim N(\mu, \sigma^2) \). Then \( a + bX \sim N(a + b\mu, b^2 \sigma^2) \).

Proof: By the theorem, \( X = \mu + \sigma Z \) where \( Z \sim N(0,1) \). So \( a + bX = a + b\mu + b\sigma Z \). The result follows.

Corollary: Let \( X \sim N(\mu, \sigma^2) \). Then \( \frac{X - \mu}{\sigma} \sim N(0,1) \).

Proof: \( \frac{X - \mu}{\sigma} = \frac{-\mu}{\sigma} + \frac{1}{\sigma}X \). The result follows.
Example 7.1. Heights are normally distributed with mean 68” and standard deviation 2”.
(a) What proportion of the population is more than 70” tall?
(b) What proportion is between 67” and 71”?

Solution:
(a) Converting 70 to standard units, we have \( z = \frac{70 - 68}{2} = 1 \). Consulting the table or an electronic device, we find that the probability that a standard normal random variable is greater than 1 is 0.1587.

(b) Converting to standard units, we have \( z_1 = \frac{67 - 68}{2} = -0.5 \), and \( z_2 = \frac{71 - 68}{2} = 1.5 \). Consulting the table or an electronic device, we find that the probability that a standard normal random variable is less than 1.5 is 0.9332 and the probability that it is less than \(-0.5\) is 0.3085. The probability that it is between \(-0.5\) and 1.5 is therefore 0.9332 - 0.3085 = 0.6247.

Example 7.2. IQ scores are normally distributed with mean 100 and standard deviation 15.
(a) Find the 85th percentile of the scores. (b) In a certain school, students with IQ scores in the top 20% are put in a special class. What proportion of students in the special class have scores above 120?

Solution:
(a) Consulting a table or electronic device, we find that the \( z \)-score corresponding to the 85th percentile is \( z = 1.04 \). The IQ score corresponding to the 85th percentile is therefore 100 + 1.04(15) = 115.60.

(b) First we find the proportion of the whole class with scores over 120. Converting to standard units, we have \( z = \frac{120 - 100}{15} = 1.33 \) Consulting a table or electronic device, we find that the probability that a standard normal random variable is greater than 1.33 is 0.0918. So the proportion of the entire class with scores over 120 is 0.0918.

Now only 0.2 of the whole class is in the special class. So the proportion of the special class with scores over 120 is \( \frac{0.0918}{0.2} = 0.459 \).
7.3 The Exponential Distribution

Definition 7.5. A random variable $X$ has the exponential distribution with parameter $\lambda > 0$ if the pdf of $X$ is

$$f(x; \lambda) = \lambda e^{-\lambda x} I_{(0, \infty)}(x)$$

We note that the cdf of $X$ is $F(x; \lambda) = (1 - e^{-\lambda x}) I_{(0, \infty)}(x)$.

We write $X \sim \text{Exp} (\lambda)$.

It can be shown that if $X \sim \text{Exp} (\lambda)$ then $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$.

The exponential distribution is often used to model waiting times. It is in a sense the continuous analog of the geometric distribution, because it too satisfies the lack of memory property.

Theorem 7.3. Lack of Memory Property: Let $X \sim \text{Exp}(\lambda)$. Then for all $s, t > 0$,

$$P(X > t + s | X > s) = P(X > t)$$

Proof: $P(X > x) = 1 - F_X(x) = e^{-\lambda x}$. Therefore

$$P(X > t + s | X > s) = \frac{P(X > t + s)}{P(X > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

The exponential distribution is essentially the only continuous distribution with the lack of memory property, as the following theorem shows.

Theorem 7.4. Let $X$ be a nonnegative continuous random variable with the property that for any nonnegative numbers $s$ and $t$, $P(X > t + s | X > s) = P(X > t)$. Then $X \sim \text{Exp}(\lambda)$ where $\lambda = -\log P(X > 1)$.

Proof: $P(X > t) = P(X > t + s | X > s) = \frac{P(X > t + s)}{P(X > s)}$,

so $P(X > t + s) = P(X > t)P(X > s)$. Define $G(x) = P(X > x)$. Note that for $s, t > 0$, $G(s + t) = G(s)G(t)$. To prove the theorem, we will show that $G(x) = e^{-\lambda x}$ for all $x > 0$. We assume that $G$ is differentiable.

Note that $G(0) = 1$ and define $\lambda = -G'(0)$. Now let $x > 0$.

$$G'(x) = \lim_{h \to 0} \frac{G(x + h) - G(x)}{h} = \lim_{h \to 0} \frac{G(x)G(h) - G(x)}{h} = \lim_{h \to 0} G(x) \frac{G(h) - 1}{h}$$

$$= G(x) \lim_{h \to 0} \frac{G(h) - 1}{h} = G(x)G'(0) = -\lambda G(x)$$
Therefore $G'(x) = -\lambda G(x)$ for all $x > 0$, so $G(x) = e^{-\lambda x}$.

### 7.4 The Poisson Process

Events occur at random times. Let $X$ be the number of events that occur in a time interval of length $t$. If there exists $\lambda > 0$ such that $X \sim \text{Poisson}(\lambda t)$ for all $t$, then the events are said to follow a Poisson process with rate $\lambda$.

**Definition 7.6.** Events follow a Poisson process if the following conditions are satisfied:
1) The number of events in any time interval of length $t$ is distributed $\text{Poisson}(\lambda t)$ (stationary increments).
2) The numbers of events in disjoint time intervals are independent (independent increments).

**Example 7.3.** Radioactive decay: It has been experimentally verified that the decay events in a large radioactive mass follow a Poisson process.

**Theorem 7.5.** Let $W$ be the waiting time from an arbitrary time until the next event occurs in a Poisson process with rate $\lambda$. Then $W \sim \text{Exp}(\lambda)$.

**Proof:**
Let $0$ represent the time at which we begin waiting for the next event. Let $t > 0$. Let $X$ be the number of events that occur in $(0, t]$. $X \sim \text{Poisson}(\lambda t)$. Now $P(W > t) = P(X = 0) = e^{-\lambda t}$. So $P(W \leq t) = 1 - e^{-\lambda t}$ so the pdf of $W$ is $d/dt(1 - e^{-\lambda t}) = \lambda e^{-\lambda t}$ which is the pdf of $\text{Exp}(\lambda)$.

**Example 7.4.** Particles are emitted by a radioactive mass at a mean rate of one every three seconds. Find the probability that more than 5 seconds elapse between events.

**Solution:**
The rate of the Poisson process is $\lambda = 1/3$. Therefore the waiting time $W$ between events is distributed $\text{Exp}(1/3)$. So $P(W > 5) = e^{-5(1/3)} = 0.1889$.

### 7.5 The Gamma Distribution

First we define the gamma function.
Definition 7.7. For $t > 0$, define $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x} \, dx$

The gamma function satisfies a recursion relation, shown below.

Theorem 7.6. For any $t > 0$, $\Gamma(t+1) = t\Gamma(t)$.

Proof: $\Gamma(t+1) = \int_0^\infty x^t e^{-x} \, dx = \left[ -x^t e^{-x} \right]_0^\infty + \int_0^\infty t x^t e^{-x} \, dx = t \int_0^\infty x^{t-1}e^{-x} \, dx = t\Gamma(t)$.

Note that $\Gamma(1) = \int_0^\infty e^{-x} \, dx = 1$. It follows from the theorem that $\Gamma(n) = (n-1)!$ for any positive integer $n$.

It can be shown that $\Gamma(1/2) = \sqrt{\pi}$.

We can now define the gamma distribution.

Definition 7.8. A random variable $X$ has the gamma distribution with parameters $r > 0$ and $\lambda > 0$ if the pdf of $X$ is

$$f(x; r, \lambda) = \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1}e^{-\lambda x} I_{(0,\infty)}(x)$$

We write $X \sim \Gamma(r, \lambda)$.

Note that the $\Gamma(1, \lambda)$ distribution is the same as Exp($\lambda$). So the exponential distribution is a special case of a gamma distribution.

Theorem 7.7. Let $X \sim \Gamma(r, \lambda)$. Then $E(X) = r/\lambda$ and $V(X) = r/\lambda^2$.

Proof:

$$E(X) = \int_0^\infty x \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1}e^{-\lambda x} \, dx = \frac{r}{\lambda} \int_0^\infty \frac{\lambda}{r\Gamma(r)} (\lambda x)^{r-1}e^{-\lambda x} \, dx = \frac{r}{\lambda}$$

$$E(X^2) = \int_0^\infty x^2 \frac{\lambda}{\Gamma(r)} (\lambda x)^{r-1}e^{-\lambda x} \, dx = \frac{r(r+1)}{\lambda^2} \int_0^\infty \frac{\lambda}{r(r+1)\Gamma(r)} (\lambda x)^{r+1}e^{-\lambda x} \, dx$$

$$= \frac{r^2 + r}{\lambda^2} \int_0^\infty \frac{\lambda}{\Gamma(r+2)} (\lambda x)^{r+2}e^{-\lambda x} \, dx = \frac{r^2 + r}{\lambda^2}$$

Now $V(X) = E(X^2) - E(X)^2 = \frac{r^2 + r}{\lambda^2} - \frac{r^2}{\lambda^2} = \frac{r}{\lambda^2}$. 

56
The family of gamma distributions contains the exponential distributions as a sub-family. Another important sub-family is defined below.

**Definition 7.9.** Let $r$ be a positive integer. The $\Gamma(r/2, 1/2)$ distribution is known as the *chi-square distribution* with $r$ degrees of freedom.

We write $X \sim \chi^2_r$. Note that if $X \sim \chi^2_r$, then $E(X) = r$ and $V(X) = 2r$.

**Theorem 7.8.** Events occur in a Poisson process with rate $\lambda$. Let $r$ be a positive integer. Let $W$ be the waiting time from an arbitrary starting point until $r$ events have occurred. Then $W \sim \Gamma(r, \lambda)$.

**Proof:** Let 0 represent the time at which we begin waiting. Let $t > 0$. Let $X$ be the number of events in $(0, t]$. Now $P(W \leq t) = P(X \geq r) = 1 - P(X < r) = 1 - \sum_{k=0}^{r-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!}$.

Now the pdf of $W$ is

$$f(t) = \frac{d}{dt} P(W \leq t) = -\sum_{k=0}^{r-1} \frac{d}{dt} e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

$$= \sum_{k=0}^{r-1} \lambda e^{-\lambda t} \frac{(\lambda t)^k}{k!} - \sum_{k=0}^{r-1} e^{-\lambda t} \frac{k\lambda t^{k-1}}{k!}$$

$$= \sum_{k=0}^{r-1} e^{-\lambda t} \frac{\lambda^{k+1} t^k}{k!} - \sum_{k=1}^{r-1} e^{-\lambda t} \frac{\lambda^k t^{k-1}}{(k-1)!}$$

$$= \sum_{k=0}^{r-1} e^{-\lambda t} \frac{\lambda^{k+1} t^k}{k!} - \sum_{k=0}^{r-2} e^{-\lambda t} \frac{\lambda^k t^{k+1}}{k!}$$

$$= e^{-\lambda t} \frac{\lambda^r t^{r-1}}{(r-1)!}$$

which is the pdf of $\Gamma(r, \lambda)$.

**Corollary:** Let $X_1, \ldots, X_r$ be i.i.d. Exp($\lambda$). Then $X_1 + \cdots + X_r \sim \Gamma(r, \lambda)$.

### 7.6 The Beta Distribution

We first define the beta function.
Definition 7.10. For $a > 0$ and $b > 0$ the function $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} \, dx$ is called the beta function.

We state without proof the fact that $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$.

Definition 7.11. The random variable $X$ has the beta distribution with parameters $a$ and $b$ if the pdf of $X$ is

$$f(x; a, b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)$$

It can be shown that if $X$ has the beta distribution with parameters $a$ and $b$, then $E(X) = a/(a + b)$ and $V(X) = ab/(a + b + 1)(a + b)^2$.

The beta distribution is often used to model probabilities.

7.7 The Lognormal Distribution

Definition 7.12. A random variable $X$ has the lognormal distribution with parameters $\mu$ and $\sigma^2$ if $\log X \sim N(\mu, \sigma^2)$.

We write $X \sim LN(\mu, \sigma^2)$.

Equivalently, we can say that if $X \sim N(\mu, \sigma)$ then $e^X \sim LN(\mu, \sigma^2)$.

The pdf of $X$ is

$$f(x, \mu, \sigma) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} I_{(0,\infty)}(x)$$

It can be shown that if $X \sim LN(\mu, \sigma^2)$, then $E(X) = e^{\mu+\sigma^2/2}$ and $V(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

Note that $\mu$ and $\sigma^2$ are not the mean and variance of the lognormal distribution. They are the mean and variance of the underlying normal distribution.

8 The Central Limit Theorem

8.1 Sample Mean and Sample Variance
Definition 8.1. Let $X_1, \ldots, X_n$ be independent random variables, all with the same distribution. Then $X_1, \ldots, X_n$ are said to be independent and identically distributed. $X_1, \ldots, X_n$ is also said to be a random sample.

Definition 8.2. Let $X_1, \ldots, X_n$ be i.i.d. The quantity $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is called the sample mean.

Definition 8.3. Let $X_1, \ldots, X_n$ be i.i.d. The quantity $s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$ is called the sample variance.

The following algebraic identity is useful.

Theorem 8.1. Let $x_1, \ldots, x_n$ be any real numbers. Let $\bar{x} = \frac{x_1 + \cdots + x_n}{n}$. Let $a$ be any real number. Then $\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$.

Proof:

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - a)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2\sum_{i=1}^{n} (x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n} (\bar{x} - a)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + 2(\bar{x} - a)\sum_{i=1}^{n} (x_i - \bar{x}) + n(\bar{x} - a)^2$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

The algebraic identity remains true, of course, when $x_1, \ldots, x_n$ are replaced with random variables $X_1, \ldots, X_n$.

Corollary: Let $X_1, \ldots, X_n$ be i.i.d. with $E(X_i) = \mu$.

Then $\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$.

In particular, $\sum_{i=1}^{n} (X_i - \mu)^2 \geq \sum_{i=1}^{n} (X_i - \bar{X})^2$. 

59
**Theorem 8.2.** Let $X_1, \ldots, X_n$ be i.i.d. with $E(X_i) = \mu$ and $V(X) = \sigma^2$. Then $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$.

**Proof:**

\[
E(\bar{X}) = E \left[ \frac{1}{n} (X_1 + \cdots + X_n) \right] = \frac{1}{n} [E(X_1) + \cdots + E(X_n)] = \frac{1}{n} n \mu = \mu.
\]

\[
V(\bar{X}) = V \left[ \frac{1}{n} (X_1 + \cdots + X_n) \right] = \frac{1}{n^2} [V(X_1) + \cdots + V(X_n)] = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}.
\]

**Theorem 8.3.** Let $X_1, \ldots, X_n$ be i.i.d. with $E(X_i) = \mu$ and $V(X) = \sigma^2$. Then $E(s^2) = \sigma^2$.

**Proof:**

\[
E(s^2) = E \left( \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} \right)
\]

\[
= \frac{1}{n-1} E \left[ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right]
\]

\[
= \frac{1}{n-1} E \left[ \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \right]
\]

\[
= \frac{1}{n-1} \sum_{i=1}^{n} E[(X_i - \mu)^2] - \frac{n}{n-1} E[(\bar{X} - \mu)^2]
\]

\[
= \frac{1}{n-1} n \sigma^2 - \frac{n}{n-1} \frac{\sigma^2}{n}
\]

\[
= \sigma^2
\]

**8.2 Statement of the Central Limit Theorem**

**Definition 8.4.** Recall the definition of the normal probability density function. Let $\mu$ be any real number and let $\sigma > 0$. The normal probability density function (normal curve) with parameters $\mu$ and $\sigma$ is the function

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

**Definition 8.5.** Recall the definition of the standard normal probability density function. The normal curve with $\mu = 0$ and $\sigma = 1$ is called the standard normal curve, and is denoted $\phi(x)$.

\[
\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]
Definition 8.6. The standard normal cumulative distribution function is \( \Phi(x) = \int_{-\infty}^{x} \phi(z) \, dz \).

The Central Limit Theorem is the most important theorem in both probability and statistics. It says that if we take a sum or average of a large enough sample, this sum or average will be approximately normally distributed, no matter what the distribution the sample came from. Following is a formal statement of the theorem.

Theorem 8.4. The Central Limit Theorem: Let \( X_1, \ldots \) be i.i.d. with mean \( \mu \) and variance \( \sigma^2 \). For each \( n \), let \( Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \). Then, for any \( a, b \) with \( -\infty \leq a < b \leq \infty \),

\[
\lim_{n \to \infty} P(a \leq Z_n \leq b) = \int_{a}^{b} \phi(x) \, dx.
\]

Note that for all \( n \), \( E(Z_n) = 0 \) and \( V(Z_n) = 1 \). The Central Limit Theorem says that in addition, the distribution of \( Z_n \) approaches the standard normal as \( n \to \infty \).

The practical implications of the Central Limit Theorem are as follows: For sufficiently large \( n \):

1) \( Z_n \approx N(0, 1) \)
2) \( \bar{X}_n \approx N(\mu, \sigma^2/n) \). This follows because \( \bar{X} = \frac{\sigma}{\sqrt{n}} Z_n + \mu \).
3) Let \( S_n = X_1 + \cdots + X_n \). Then \( S_n = n\bar{X}_n \), so \( S_n \approx N(n\mu, n\sigma^2) \).

The practical importance of the Central Limit Theorem is that the distribution of a sample mean can be approximated with the normal distribution, so long as the sample size is large enough. It is natural to ask how large the sample size must be so that the normal approximation is good. The answer is that the necessary sample size depends on the amount of skewness \( \mu_3/\sigma^3 \). The larger the skewness, the larger the sample size is needed. When the sample is from a symmetric distribution (skewness = 0) a small sample is sufficient. For example the distribution of the mean of 12 uniform random variables is virtually indistinguishable from the normal distribution. For most distributions, a sample size of 30 is large enough for the normal approximation to be adequate. The following figure illustrates this.
The leftmost plot in each row is the distribution (probability density function or probability mass function) of a random variable. The two plots to its right are the distributions of the sample mean (solid line) for samples of size 5 and 30 respectively, with the normal curve (dashed line) superimposed. **Top row:** Since the original distribution is nearly symmetric, the normal approximation is good even for a sample size as small as 5. **Middle row:** The original distribution is somewhat skewed. Even so, the normal approximation is reasonably close even for a sample of size 5, and very good for a sample of size 30. **Bottom row:** The original distribution is highly skewed. The normal approximation is not good for a sample size of 5, but is reasonably good for a sample of size 30. Note that two of the original distributions are continuous, and one is discrete. The Central Limit Theorem holds for both continuous and discrete distributions.
We will prove the Central Limit Theorem later, after we have covered moment generating functions. Now we present some applications:

Example 8.1. Students at a certain university have a mean age of 20.2 years, with a standard deviation of 2.1 years. Find the probability that the mean age of a random sample of 144 students is less than 19.7 years.

Solution: Let $X_i$ be the age of the $i$th sampled student. Then $E(X_i) = 20.2$ and $\sigma = 2.1$. It follows that $\sqrt{\frac{144}{2.1}}(\bar{X} - 20.2) \approx N(0, 1)$.

Now $P(\bar{X} < 19.7) = P \left( \frac{\sqrt{144}(\bar{X} - 20.2)}{2.1} < \frac{\sqrt{144}(19.7 - 20.2)}{2.1} \right) = P(Z < -2.29) = 0.0110$.

Example 8.2. It is claimed that the breaking strength (in kg/mm) for a certain type of fabric has mean 2.0 and standard deviation 0.4. A random sample of 80 pieces of fabric is drawn. The sample mean is 1.91.

(a) Assuming the claim is true, find the probability that a sample of size 80 will have a mean less than or equal to 1.91.

(b) Assuming the claim is true, how large a sample is needed so that $P(\bar{X} < 1.91) = 0.01$?

Solution:

(a) Let $X_i$ be the breaking strength of the $i$th piece of fabric. Then $\mu = E(X_i) = 2.0$ and $\sigma = SD(X_i) = 0.4$. It follows that $\bar{X} \approx N(2, 0.4^2/80)$.

Now $P(\bar{X} < 1.91) = P \left( \frac{\sqrt{80}(\bar{X} - 2.0)}{0.4} < \frac{\sqrt{80}(1.91 - 2.0)}{0.4} \right) = P(Z < -2.01) = 0.0222$.

(b) $P(\bar{X} < 1.91) = P \left( \frac{\sqrt{n}(\bar{X} - 2.0)}{0.4} < \frac{\sqrt{n}(1.91 - 2.0)}{0.4} \right) = P \left( Z < \frac{\sqrt{n}(1.91 - 2.0)}{0.4} \right) = 0.01$. Now $P(Z < -2.33) = 0.01$, so $\frac{\sqrt{n}(1.91 - 2.0)}{0.4} = -2.33$. Solve to find $n = 107.24$. Round up to $n = 108$.

8.3 Normal Approximation to the Binomial

Let $Y_1, \ldots, Y_n$ be i.i.d. Bernoulli with success probability $p$. $E(Y_i) = p$, $V(Y_i) = p(1-p)$. Let $X = Y_1 + \cdots + Y_n$. Then $X \sim Bin(n, p)$. From the Central Limit Theorem we conclude:
1) If $n$ is sufficiently large, $X \approx N(np, np(1-p))$.

2) Let $\hat{p} = \bar{Y} = X/n$ be the proportion of successes in the sample. Then $\hat{p} \approx N(p, p(1-p)/n)$.

The normal approximation is good if $np > 10$ and $n(1-p) > 10$.

**Example 8.3.** A fair coin is tossed 100 times. Find the probability of obtaining more than 55 heads. Find the probability of obtaining less than 55 heads. Find the probability of obtaining exactly 55 heads.

Note the continuity correction. For $P(X > 55)$ we compute $P(X > 55.5)$. For $P(X < 55)$, we compute $P(X < 54.5)$. For $P(X = 55)$, we compute $P(54.5 < X < 55.5)$.

**Solution:** Let $X$ be the number of heads. Then $X \sim \text{Bin}(100, 0.5)$. Now $E(X) = 50$, $V(X) = 25$. $P(X > 55.5) = P \left( \frac{X - 50}{5} > \frac{55.5 - 50}{5} \right) = P(Z > 1.1) = 0.1357$.

Similarly for $P(X < 55)$, we obtain $P(X < 55) = P(Z < 0.9) = 0.8159$. For $P(X = 55)$, we obtain $P(0.9 < Z < 1.1) = 0.0484$.

### 8.4 Normal Approximation to the Poisson

Let $\lambda > 10$. For each integer $n > 10$, let $X_n \sim \text{Bin}(n, \lambda/n)$. Because $\lambda > 10$, $X_n$ is approximately normal for large $n$. Now $X_n \approx N(\lambda, \lambda(1-\lambda/n))$. As $n \to \infty$, $\lambda(1-\lambda/n) \to \lambda$, so $X_n \to N(\lambda, \lambda)$. However, we also know that $X_n \to \text{Poisson}(\lambda)$. From the Central Limit Theorem we conclude:

Let $X \sim \text{Poisson}(\lambda)$. Then for sufficiently large $\lambda$ (i.e. $\lambda > 10$), $X \approx N(\lambda, \lambda)$.

**Example 8.4.** Grandma is making chocolate chip cookies. She accidentally dumps the whole bag of chocolate chips into the dough, so that there are an average of 50 chips per cookie. What is the probability that your cookie contains more than 40 chips?

**Solution:**

Let $X$ be the number of chips in your cookie. Then $X \sim \text{Poisson}(50)$, so $X \approx N(50, 50)$. $P(X > 40) = P \left( \frac{X - 50}{\sqrt{50}} > \frac{40 - 50}{\sqrt{50}} \right) = P(Z > -1.41) = 0.9207$. 

64
9 Jointly Distributed Random Variables

9.1 Jointly Discrete Random Variables

Definition 9.1. The random variables $X$ and $Y$ are said to be jointly discrete if the ordered pair $(X, Y)$ can take on only a countable number of values with positive probability.

Definition 9.2. Let $X, Y$ be jointly discrete. The joint probability mass function is $p(x, y) = P(X = x, Y = y)$.

Definition 9.3. Let $X$ and $Y$ be jointly discrete. The probability mass function of $X$, $p_X(x)$, is called the \textbf{marginal pmf} of $X$. Similarly, $p_Y(y)$ is called the marginal pmf of $Y$.

Example 9.1. Two fair three-sided dice are rolled. Let $X$ be the number on the first die and let $Y$ be the larger of the two numbers. Find the joint pmf of $X$ and $Y$.

Solution:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2/9</th>
<th>1/9</th>
<th>1/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Note that the numbers in the margins are the marginal mass functions.

Definition 9.4. Discrete random variables that have the same distribution (pmf) are said to be \textbf{identically distributed}. So discrete random variables $X$ and $Y$ are identically distributed if $P(X = v) = P(Y = v)$ for every real number $v$.

Proposition 9.1. If $X$ and $Y$ are identically distributed, and $g$ is a function, then $g(X)$ and $g(Y)$ are identically distributed.

Definition 9.5. Random variables $X$ and $Y$ are equal if $P(X = Y) = 1$.

Theorem 9.1. If $X$ and $Y$ are equal, then $X$ and $Y$ are identically distributed. But if $X$ and $Y$ are identically distributed, they are not necessarily equal.
Example 9.2. Let $X$ be a random variable with pmf $P(X = 0) = P(X = 1) = 1/2$. Let $Y = 1 - X$. Then $P(X = 0) = P(Y = 0) = 1/2$ and $P(X = 1) = P(Y = 1) = 1/2$, so $X$ and $Y$ have the same distribution. However, $X \neq Y$. When $X = 0$, $Y = 1$ and when $X = 1$ $Y = 0$.

Theorem 9.2. Let $X$ and $Y$ be jointly discrete with joint pmf $p(x, y)$. The the marginal pmf of $X$ is $p_X(x) = \sum_y p(x, y)$, where the sum is taken over the mass points of $Y$. Similarly, $p_Y(y) = \sum_x p(x, y)$.

Proof: $p_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y p(x, y)$.

Remark 9.1. If $X$ and $Y$ are jointly discrete, then the joint pmf determines the marginal pmfs, as shown above. However, the marginal pmfs do not determine the joint pmf.

Example 9.3. Two dice are rolled. Let $X$ be the number on the top of the first die, let $Y$ be the number on the top of the second die, and let $Z$ be the number on the bottom of the first die. Then $X$, $Y$, and $Z$ have the same marginals, each putting mass $1/6$ on the points $1, 2, 3, 4, 5, 6$. But the joint pmf of $X$ and $Y$ differs from the joint pmf of $X$ and $Z$.

<table>
<thead>
<tr>
<th>Joint pmf of $X$ and $Y$</th>
<th>Joint pmf of $X$ and $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>Total</td>
<td>1/6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Joint pmf of $X$ and $Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Definition 9.6. Let $X_1,\ldots,X_n$ be discrete random variables, each defined on the same sample space. The random variables are said to be jointly distributed. The vector $(X_1,\ldots,X_n)$ is called a random vector. The joint probability mass function of $X_1,\ldots,X_n$ is $p(x_1,\ldots,x_n) = P(X_1 = x_1,\ldots,X_n = x_n)$. 66
9.2 The Multinomial Distribution

The multinomial distribution is an extension of the binomial distribution. A multinomial trial is an experiment that results in one of $k$ outcomes, with probabilities $p_1, ..., p_k$. Clearly $p_1 + \cdots + p_k = 1$. Now assume that $n$ independent multinomial trials, each with the same probabilities, are carried out, and let $X_i, i = 1, ..., k$ be the number of trials that result in outcome $k$. Then $X_1, ..., X_k$ are jointly distributed with joint pmf

$$f(x_1, ..., x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

for $x_1, ..., x_k$ non-negative integers with $x_1 + \cdots + x_k = n$.

It is important to note that if $(X_1, ..., X_k) \sim MN(n, p_1, ..., p_k)$, then each $X_i$ is marginally binomial with parameters $n$ and $p_i$.

Example 9.4. A fair die is rolled 10 times. Find the probability that 2 of the rolls are 1, 1 is a 2, none are 3, 5 are 4, 1 is 5 and 1 is 6.

Solution:

$$P(X_1 = 2, X_2 = 1, X_3 = 0, X_4 = 5, X_5 = 1, X_6 = 1) = \frac{10!}{2! 1! 0! 5! 1! 1!} \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^5 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 = 0.00025006$$

9.3 Jointly Continuous Random Variables

Definition 9.7. The random variables $X$ and $Y$ are jointly continuous if there exists a nonnegative function $f(x, y)$ such that for all $x$ and $y$,

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u, v) \, du \, dv$$

The function $f(x, y)$ is a joint pdf for $X$ and $Y$. The function $F(x, y) = P(X \leq x, Y \leq y)$ is the joint cdf for $X$ and $Y$.

We state the following without proof:

1) It follows from the definition that if $f(x, y)$ is a joint pdf, then

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$$
2) The joint pdf can be used to compute probabilities. For example, if \( X \) and \( Y \) are jointly continuous, then

\[
P(a \leq X \leq b, c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x, y) \, dy \, dx.
\]

3) Let \( X \) and \( Y \) be jointly continuous. Then a joint pdf for \( X \) and \( Y \) is

\[
f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).
\]

**Definition 9.8.** Let \( X \) and \( Y \) be jointly continuous. The pdf of \( X \), \( f_X(x) \), is called the marginal pdf of \( X \). Similarly for the pdf of \( Y \).

**Theorem 9.3.** Let \( X \) and \( Y \) be jointly continuous with joint pdf \( f(x, y) \). The marginal pdf of \( X \) is \( f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, dy \), and the marginal pdf of \( Y \) is \( f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx \).

**Proof:**

\[
f_X(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} P(X \leq x) = \frac{d}{dx} P(X \leq x, Y \leq \infty)
\]

\[
= \frac{d}{dx} \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y) \, dy \, du = \int_{-\infty}^{\infty} f(x, y) \, dy
\]

**Example 9.5.** Let \( X \) and \( Y \) be jointly distributed with joint pdf \( f(x, y) = (x+y)I_{(0,1)}(x)I_{(0,1)}(y) \). Find the marginal pdfs \( f_X(x) \) and \( f_Y(y) \).

**Solution:** \( f_X(x) = (x + 1/2)I_{(0,1)}(x) \), \( f_Y(y) = (y + 1/2)I_{(0,1)}(y) \).

**Example 9.6.** Let \( X \) and \( Y \) have joint pdf \( f(x, y) = xe^{-(x+xy)}I_{(0,\infty)}(x)I_{(0,\infty)}(y) \). Find \( P(X > 1, Y > 2) \).

**Solution:**

The pdf is positive for positive values of \( x \) and \( y \). We integrate the pdf over the region where \( x > 1 \) and \( y > 2 \).

\[
P(X > 1, Y > 2) = \int_{1}^{\infty} \int_{2}^{\infty} xe^{-(x+xy)} \, dy \, dx = \frac{1}{3e^3}
\]

**Example 9.7.** Let \( X \) and \( Y \) have joint pdf \( f(x, y) = 4xyI_{(0,1)}(x)I_{(0,1)}(y) \). Find \( P(Y < 2X) \).

**Solution:**

The pdf is positive on the unit square: \( 0 < x < 1, 0 < y < 1 \). We need to compute the integral of the pdf over the region of the unit square where \( y < 2x \). It is easier to compute the integral over the region \( y > 2x \) and subtract from 1.

\[
P(Y > 2X) = \int_{0}^{1/2} \int_{2x}^{1} 4xy \, dy \, dx = \frac{3}{16}
\]

Therefore \( P(Y < 2X) = 1 - \frac{3}{16} = \frac{13}{16} \).
Example 9.8. Let $X$ and $Y$ have joint pdf $f(x, y) = 8xyI_{(0,1)}(x)I_{(0,x)}(y)$. Find $P(Y > X^2)$.

Solution:
The pdf is positive on the triangle bounded by $0 < x < 1$ and $0 < y < x$. We compute the integral of the pdf over the region of this triangle where $y < x^2$.

$$P(Y > X^2) = \int_0^1 \int_{x^2}^x 8xy \, dy \, dx = \frac{1}{3}$$

9.4 Conditional Distributions

Sometimes we have two jointly distributed random variables and we know the value of one of them. This knowledge can affect the probability distribution of the other random variable. The distribution of a random variable, given that we know the value of another random variable, is called a conditional distribution.

Definition 9.9. Let $X$ and $Y$ be jointly discrete random variables. The conditional distribution of $Y$ given $X = x$ is

$$p(y|x) = P(Y = y | X = x) = \frac{P(Y = y \cap X = x)}{P(X = x)}$$

For now, we will think of $x$ as fixed, so $p(y|x)$ is a function of $y$.

Theorem 9.4. Let $X$, $Y$ be jointly discrete. Then for all $x, y$,

$$P(X = x, Y = y) = P(X = x)P(Y = y | X = x).$$

We now extend this idea to continuous random variables.

Definition 9.10. Let $X$ and $Y$ be jointly continuous with joint pdf $f(x, y)$. The conditional pdf of $Y$ given $X = x$ is $f(y|x) = \frac{f(x, y)}{f_X(x)}$, defined for all $x$ for which $f_X(x) > 0$. $f(y|x)$ is undefined if $f_X(x) = 0$.

Similarly, $f(x|y) = \frac{f(x, y)}{f_Y(y)}$, defined for all $y$ for which $f_Y(y) > 0$.

Definition 9.11. The conditional cumulative distribution function of $Y$ given $X = x$ is $F(y | x) = \int_{-\infty}^y f(v|x) \, dv$
Example 9.9. Let $X$ and $Y$ be jointly distributed with joint pdf $f(x, y) = (x+y)I_{(0,1)}(x)I_{(0,1)}(y)$. Find $f(y|x)$ and $F(y|x)$.

Solution: We know that $f_X(x) = (x + 1/2)I_{(0,1)}(x)$. So $f(y|x)$ is defined only for $0 < x < 1$. Let $0 < x < 1$. Then

$$f(y|x) = \frac{(x + y)I_{(0,1)}(y)}{x + 1/2} = \frac{x + y}{x + 1/2}I_{(0,1)}(y)$$

Now if $0 < y < 1$,

$$F(y|x) = \int_{-\infty}^{y} \frac{x + u}{x + 1/2}I_{(0,1)}(u) du = \int_{0}^{y} \frac{x + u}{x + 1/2} du = \frac{xy + y^2/2}{x + 1/2}$$

Theorem 9.5. Let $X$ and $Y$ be jointly continuous. Then $F_Y(y) = \int_{-\infty}^{\infty} F_{Y|X}(y|x)f_X(x) dx$.

Proof:

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(u) du = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(x, u) dx du = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(u|x)f_X(x) dx du$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(u|x)f_X(x) du dx = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{y} f(u|x) du \right] f_X(x) dx = \int_{-\infty}^{\infty} F(y|x)f_X(x) dx$$

Definition 9.12. Random variables $X_1, ..., X_n$ are jointly discrete if the vector $(X_1, ..., X_n)$ can take on only a finite or countably infinite number of values with positive probability. The joint pmf is $p(x_1, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n)$.

Definition 9.13. Random variables $X_1, ..., X_n$ are jointly continuous if there exists a function $f(x_1, ..., x_n)$ such that $f(x_1, ..., x_n) \geq 0$ for all $x_1, ..., x_n$ and $P(X_1 \leq x_1, ..., X_n \leq x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, ..., u_n) du_1 \cdots du_n$. The function $f(x_1, ..., x_n)$ is the joint pdf. The function $F(x_1, ..., x_n) = P(X_1 \leq x_1, ..., X_n \leq x_n)$ is called the joint cdf.

The joint pmf can be obtained from the joint cdf through differentiation:

$$f(x_1, ..., x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F(x_1, ..., x_n)$$

9.5 Independent Random Variables

Sometimes knowledge of the value of one random variable does not affect the probability distribution of the other. When this occurs we say that the random variables are independent.
Definition 9.14. Let $X$ and $Y$ be random variables. $X$ and $Y$ are independent if for any sets of real numbers $A$ and $B$, $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.

If $X$ and $Y$ are jointly discrete, then $X$ and $Y$ are independent if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$ for any real numbers $x$ and $y$. In other words, the joint pmf is the product of the marginal pmfs.

If $X$ and $Y$ are jointly continuous with joint pdf $f(x,y)$, then $X$ and $Y$ are independent if and only if $f(x,y) = f_X(x)f_Y(y)$ for any real numbers $x$ and $y$. In other words, the joint pdf is the product of the marginal pdfs. In fact, $X$ and $Y$ are independent if $f(x,y) = g(x)h(y)$ for any functions $g$ and $h$.

Theorem 9.6. Let $X$ and $Y$ be jointly discrete. $X$ and $Y$ are independent if and only if for any real numbers $x$ and $y$ with $P(X = x) \neq 0$, $P(Y = y|X = x) = P(Y = y)$; and if $P(Y = y) \neq 0$, $P(X = x|Y = y) = P(X = x)$.

Proof: Assume $X$ and $Y$ are independent. Then

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

Now assume $P(Y = y) = P(Y = y|X = x)$. Then

$$P(X = x, Y = y) = P(Y = y|X = x)P(X = x) = P(Y = y)P(X = x)$$

Theorem 9.7. Let $X$ and $Y$ be jointly continuous. $X$ and $Y$ are independent if and only if for any real numbers $x$ and $y$ with $f_X(x) \neq 0$, $f(y|x) = f_Y(y)$; and if $f_Y(y) \neq 0$, $f(x|y) = f_X(x)$.

Proof: Assume $X$ and $Y$ are independent. Then

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{f_X(x)f_Y(y)}{f_X(x)} = f_Y(y)$$

Now assume $f_Y(y) = f(y|x)$. Then

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

Theorem 9.8. Let $X$ and $Y$ be independent, and let $g$ and $h$ be any functions. Then $g(X)$ and $h(Y)$ are independent.

Proof: Omitted
Example 9.10. Let $X$, $Y$ be independent. Then so are $X^2$ and $e^Y$, $\sin X$ and $|Y|$, etc.

Definition 9.15. Let $X_1, \ldots, X_n$ be jointly discrete. $X_1, \ldots, X_n$ are independent if $P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$ for all real numbers $x_1, \ldots, x_n$.

Definition 9.16. Let $X_1, \ldots, X_n$ be jointly continuous. $X_1, \ldots, X_n$ are independent if $f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$ for all real numbers $x_1, \ldots, x_n$.

Theorem 9.9. Let $X_1, \ldots, X_n$ be independent random variables. Let $g_1, \ldots, g_n$ be any functions. Then $g_1(X_1), \ldots, g_n(X_n)$ are independent.

Proof: Omitted

Example 9.11. Let $X$ and $Y$ be jointly distributed with joint pdf $f(x, y) = 4xyI_{(0,1)}(x)I_{(0,1)}(y)$. $X$ and $Y$ are independent. To see this, note that $f(x, y)$ can be factored into a product of a function of $x$ and a function of $y$: $f(x, y) = [2xI_{(0,1)}(x)][2yI_{(0,1)}(y)]$.

Example 9.12. Let $X$ and $Y$ be jointly distributed with joint pdf $f(x, y) = (x+y)I_{(0,1)}(x)I_{(0,1)}(y)$. Are $X$ and $Y$ independent?

Solution: $f_X(x) = (x+1/2)I_{(0,1)}(x)$, $f_Y(y) = (y+1/2)I_{(0,1)}(y)$. Since $f(x, y) \neq f_X(x)f_Y(y)$, $X$ and $Y$ are not independent.

Example 9.13. A penny is tossed 5 times and a nickel is tossed 6 times. Both are fair coins. Find the probability that the penny comes up heads three times and the nickel comes up heads four times.

Solution:

Let $X$ be the number of times the penny comes up heads and let $Y$ be the number of times the nickel comes up heads. We must find $P(X = 3 \cap Y = 4)$.

$X$ and $Y$ are independent, with $X \sim \text{Bin}(5, 0.5)$ and $Y \sim \text{Bin}(6, 0.5)$. Because $X$ and $Y$ are independent, $P(X = 3 \cap Y = 4) = P(X = 3)P(Y = 4)$.

$P(X = 3) = \frac{5!}{3!2!}(0.5)^3(0.5)^2 = 10/32$, and $P(Y = 4) = \frac{6!}{4!2!}(0.5)^4(0.5)^2 = 15/64$.

Therefore $P(X = 3 \cap Y + 4) = (10/32)(15/64) = 150/2048 = 0.0732$.
9.6 Computing Expectations from Joint Distributions

**Theorem 9.10.** Let $X$ and $Y$ be jointly distributed. Then

- $E(X) = \sum_x \sum_y x P(X = x, Y = y)$ if $X$ and $Y$ are jointly discrete.
- $E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) \, dy \, dx$ if $X$ and $Y$ are jointly continuous.

Similarly, $E(Y) = \sum_x \sum_y y P(X = x, Y = y)$ or $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) \, dy \, dx$.

**Proof:**
Discrete: $\sum_x \sum_y x P(X = x, Y = y) = \sum_x x \sum_y P(X = x, Y = y) = \sum_x x P(X = x) = E(X)$.
Continuous: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) \, dy \, dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x, y) \, dy \, dx = \int_{-\infty}^{\infty} x f_X(x) \, dx$.

**Theorem 9.11.** Let $X$ and $Y$ be jointly distributed, and let $g(X,Y)$ be a function. Then

- $E(g(X,Y)) = \sum_x \sum_y g(x, y) P(X = x, Y = y)$ if $X$ and $Y$ are jointly discrete.
- $E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) \, dx \, dy$ if $X$ and $Y$ are jointly continuous.

**Proof:** Omitted.

**Example 9.14.** Let $X$ and $Y$ be jointly distributed with joint pdf $f(x, y) = (x+y)I_{(0,1)}(x)I_{(0,1)}(y)$. Find $E(X)$, $E(Y)$, $E(X+Y)$, $E(XY)$.

**Solution:** $E(X) = E(Y) = 7/12$, $E(X+Y) = 7/6$, $E(XY) = 1/3$.
Note that $E(X+Y) = E(X) + E(Y)$, but $E(XY) \neq E(X)E(Y)$.

**Example 9.15.** Two three-sided dice are rolled. Let $X$ be the number on the first die and let $Y$ be the number on the second die. Find $E(X)$, $E(Y)$, $E(X+Y)$, $E(XY)$.

**Solution:** $E(X) = E(Y) = 2$, $E(X+Y) = 4$, $E(XY) = 4$.
Note that $E(X+Y) = E(X) + E(Y)$, and $E(XY) = E(X)E(Y)$ as well.

**Theorem 9.12.** Let $X_1, \ldots, X_n$ be jointly distributed, and assume that $E(X_i)$ exists for each $i$. Let $c_1, \ldots, c_n$ be constants. Then

$$E(c_1 X_1 + \cdots + c_n X_n) = c_1 E(X_1) + \cdots + c_n E(X_n)$$

**Proof:** Follows directly from properties of sums and integrals.
Example 9.16.
1) Let $Y \sim \text{Bin}(n,p)$. Then $Y$ can be considered to be the sum $X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent Bernoulli trials with success probability $p$. It follows that $E(Y) = E(X_1) + \cdots + E(X_n) = np$.

2) There are $N$ balls in an urn, with $M$ of them red and $N - M$ green. $K$ balls are drawn without replacement. Let $X_i = 1$ if the $i$th ball is red and 0 otherwise. Let $Y = X_1 + \cdots + X_n$ be the total number of red balls drawn. Each $X_i$ is a Bernoulli trial with success probability $M/N$ (the trials are not independent, however). It follows that $E(Y) = E(X_1) + \cdots + E(X_n) = np$. The fact that $X_1, \ldots, X_n$ are not independent does not matter.

The fact that the joint pmf or pdf of independent random variables factors into the product of the marginals makes expectations of products of independent random variables easy to compute, as the following theorem shows.

**Theorem 9.13.** Let $X$ and $Y$ be independent random variables. Then $E(XY) = E(X)E(Y)$.

**Proof:** We prove the result in the continuous case. The discrete case is similar.

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_X(x)f_Y(y) \, dx \, dy = \int_{-\infty}^{\infty} x f_X(x) \, dx \int_{-\infty}^{\infty} y f_Y(y) \, dy = E(X)E(Y)$$

### 9.7 Conditional Expectation

**Definition 9.17.** Let $X$ and $Y$ be jointly distributed. The **conditional expectation** of $Y$ given $X = x$ is

$$E(Y|X = x) = \begin{cases} \sum_y yP(Y = y|X = x) & \text{if } X \text{ and } Y \text{ are jointly discrete} \\ \int_{-\infty}^{\infty} yf_Y(y|X = x) \, dy & \text{if } X \text{ and } Y \text{ are jointly continuous} \end{cases}$$

Notice that $E(Y|X = x)$ is a function of $x$. 
Definition 9.18. The function \( h(x) = E(Y|X = x) \) is called the regression curve of \( Y \) on \( X \).

Example 9.17. Let \( X \) and \( Y \) be independent binomial random variables, both distributed Bin\((n, p)\). Let \( Z = X + Y \). Find \( E(X|Z = m) \).

Example 9.18. Two fair three-sided dice are rolled. Let \( X \) be the number on the first die and let \( Y \) be the larger of the two numbers. Find \( E(Y|X = 1) \), \( E(Y|X = 2) \), and \( E(Y|X = 3) \). Also find the unconditional expectation \( E(Y) \).

Solution: Below are the joint pmf of \( X \) and \( Y \) and the conditional pmf of \( Y \) given \( X \).

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>( f_X(x) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/9</td>
<td>1/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2/9</td>
<td>1/9</td>
<td>1/3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>1/3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( y )</th>
<th>1/9</th>
<th>3/9</th>
<th>5/9</th>
</tr>
</thead>
</table>

\[
E(Y|X = 1) = 1P(Y = 1|X = 1) + 2P(Y = 2|X = 1) + 3P(Y = 3|X = 1) = 1(1/3) + 2(1/3) + 3(1/3) = 2
\]

\[
E(Y|X = 2) = 1P(Y = 1|X = 2) + 2P(Y = 2|X = 2) + 3P(Y = 3|X = 2) = 1(0) + 2(2/3) + 3(1/3) = 7/3
\]

\[
E(Y|X = 3) = 1P(Y = 1|X = 3) + 2P(Y = 2|X = 3) + 3P(Y = 3|X = 3) = 1(0) + 2(0) + 3(1) = 3
\]

Using the marginal pmf \( f_Y(y) \), we find that

\[
E(Y) = 1(1/9) + 2(3/9) + 3(5/9) = 22/9
\]
Example 9.19. Let $X$ and $Y$ be jointly distributed with joint pdf $f(x, y) = (x + y)I_{(0,1)}(x)I_{(0,1)}(y)$. Find $E(Y|X = x)$.

Solution: We have previously computed $f(y|x) = \frac{x + y}{x + 1/2}I_{(0,1)}(y)$. Therefore

$$E(Y|X = x) = \int_0^1 y \frac{x + y}{x + 1/2} dy = \frac{3x + 2}{6x + 3}$$

The conditional expectation $E(Y|X = x)$ is a function of $x$. If we substitute $X$ for $x$, formally composing this function with the random variable $X$, the result is a random variable denoted $E(Y|X)$. So in Example 9.19, $E(Y|X) = \frac{3X + 2}{6X + 3}$.

Since $E(Y|X)$ is a random variable, we can find its expectation. In Example 9.19, the marginal pdf of $X$ is $f_X(x) = (x + 1/2)I_{(0,1)}(x)$, so

$$E[E(Y|X)] = \int_0^1 \frac{3x + 2}{6x + 3}(x + 1/2) dx = 7/12$$

Note that $E[E(Y|X)] = E(Y)$.

Now we will compute $E[E(Y|X)]$ for the distribution in Example 9.18. Again, since $E(Y|X)$ is a function of $X$, we use the marginal pmf of $X$.

$$E[E(Y|X)] = E(Y|X = 1)P(X = 1) + E(Y|X = 2)P(X = 2) + E(Y|X = 3)P(X = 3)$$

$$= 2(1/3) + (7/3)(1/3) + 3(1/3) = 22/9$$

Once again, $E[E(Y|X)] = E(Y)$. The following theorem shows that this is true in general.


Proof: We prove the result in the continuous case. The discrete case is similar.

$$E[E(Y|X)] = \int_{-\infty}^{\infty} E(Y|X = x)f_X(x) dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(y|x) dyf_X(x) dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(y|x)f_X(x) dy dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf(x, y) dy dx = E(Y)$$
Example 9.20. Each day, ten items are sampled from an assembly line and checked for defects. Each item has probability \( p \) of being defective. The value of \( p \) varies from day to day, and is distributed uniformly on \((0, 0.2)\). Let \( Y \) be the number of defective items found on a given day. Find \( E(Y) \).

Solution:
Given \( p, Y \sim \text{Bin}(10, p) \), so \( E(Y \mid p) = 10p \). Now \( p \sim U(0, 0.2) \), so
\[
E(p) = \frac{0 + 0.2}{2} = 0.1.
\]
Therefore
\[
E(Y) = E(E(Y \mid p)) = E(10p) = 10E(p) = 10(0.1) = 1
\]

Theorem 9.15. Let \( X \) and \( Y \) be jointly distributed. Let \( g(X, Y) \) be any function of \( X \) and \( Y \). The conditional expectation of \( g(X, Y) \) given \( X = x \) is
\[
E(g(X, Y) \mid X = x) = \begin{cases} 
\sum_y g(x, y) P(Y = y \mid X = x) & \text{if } X \text{ and } Y \text{ are jointly discrete} \\
\int_{-\infty}^{\infty} g(x, y) f(y \mid x) dy & \text{if } X \text{ and } Y \text{ are jointly continuous}
\end{cases}
\]

Definition 9.19. Let \( X \) and \( Y \) be jointly distributed. The conditional variance of \( Y \) given \( X \) is
\[
V(Y \mid X) = E(Y^2 \mid X) - [E(Y \mid X)]^2.
\]

Theorem 9.16. Analysis of Variance:
\[
V(Y) = E[V(Y \mid X)] + V[E(Y \mid X)].
\]

Proof:
\[
E[V(Y \mid X)] = E[E(Y^2 \mid X)] - E[E(Y \mid X)^2] = E(Y^2) - E(Y)^2 + E(Y)^2 - E[E(Y \mid X)^2] = E(Y) - (E[E(Y \mid X)^2] - E[E(Y \mid X)^2]) = V(Y) - V[E(Y \mid X)]
\]

10 Covariance and Correlation

Definition 10.1. Let \( X \) and \( Y \) be jointly distributed with means \( \mu_X = E(X) \) and \( \mu_Y = E(Y) \). The covariance of \( X \) and \( Y \) is
\[
\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]
\]
if the expectation exists.

An important special case of the covariance is:
Proposition 10.1. For any random variable $X$, $\text{Cov}(X, X) = V(X)$.

Theorem 10.1. $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

Proof: Let $\mu_X = E(X)$ and $\mu_Y = E(Y)$. Then

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY - \mu_XY - \mu_YX + \mu_X\mu_Y)$$

$$= E(XY) - \mu_XE(Y) - \mu_YE(X) + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y$$

Definition 10.2. Let $X$ and $Y$ be jointly distributed with means $\mu_X = E(X)$ and $\mu_Y = E(Y)$, standard deviations $\sigma_X > 0$ and $\sigma_Y > 0$. The correlation between $X$ and $Y$ is

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

$\rho_{X,Y}$ is sometimes called the correlation coefficient.

Definition 10.3. Random variables $X$ and $Y$ are uncorrelated if $\text{Cov}(X, Y) = 0$.

Remark 10.1. It follows that $X$ and $Y$ are uncorrelated if and only if $\rho_{X,Y} = 0$, or $X$ or $Y$ is constant.

Remark 10.2. It immediately follows that $X$ and $Y$ are uncorrelated if and only if $E(XY) = E(X)E(Y)$.

Theorem 10.2. If $X$ and $Y$ are independent, then $X$ and $Y$ are uncorrelated, since then $E(XY) = E(X)E(Y)$.

The converse of the theorem is false. There are many examples of random variables that are uncorrelated but not independent. Here are two.

Example 10.1. $(X, Y) = (-1, 0), (0, 1), (1, 0)$ each with probability $1/3$. 

78
Example 10.2. Let \( U \sim U(0, \pi) \). Let \( X = \cos(U) \) and \( Y = \sin(U) \). Then \( E(X) = 0 \) so \( \text{Cov}(X, Y) = E(XY) = \frac{1}{\pi} \int_0^\pi \cos u \sin u \, du = 0 \). But \( X \) and \( Y \) are not independent; in fact \( Y = \sqrt{1 - X^2} \).

In general, let \( X \sim U(-a, a) \), and let \( Y = g(X) \) be an even function of \( X \). Then \( \text{Cov}(X, Y) = E(XY) = \frac{1}{2a} \int_{-a}^a xg(x) \, dx = 0 \) since \( xg(x) \) is an odd function.

Example 10.3. Let \( X \) and \( Y \) be jointly distributed with joint pdf \( f(x, y) = (x+y)I_{(0,1)}(x)I_{(0,1)}(y) \). Find \( \text{Cov}(X, Y) \) and \( \rho_{X,Y} \).

Solution: We have previously computed \( E(X) = E(Y) = 7/12 \), \( E(XY) = 1/3 \). So \( \text{Cov}(X, Y) = 1/3 - (7/12)(7/12) = -1/144 \).

To compute the correlation we first compute \( E(X^2) = E(Y^2) = 5/12 \).
Then \( V(X) = V(Y) = 5/12 - (7/12)^2 = 11/144 \).
So \( \rho_{X,Y} = (-1/144)/\sqrt{(11/144)(11/144)} = -1/11 \).

Theorem 10.3. Cauchy-Schwartz Inequality: Let \( X \) and \( Y \) be random variables such that \( E(X^2) \) and \( E(Y^2) \) exist. Then \( E(XY)^2 \leq E(X^2)E(Y^2) \).

Proof: Define \( h(t) = E[(tX - Y)^2] \).
Then \( h(t) \geq 0 \) for all \( t \), and \( h(t) = t^2 E(X^2) - 2t E(XY) + E(Y^2) \).
Now substitute \( E(XY) \) for \( t \). We conclude that \( \frac{E(XY)^2}{E(X^2)} - 2 \frac{E(XY)^2}{E(X^2)} + E(Y^2) \geq 0 \).
But then \( \frac{E(XY)^2}{E(X^2)} \leq E(Y^2) \), so \( E(XY)^2 \leq E(X^2)E(Y^2) \).

The Cauchy-Schwartz inequality has an important corollary.

Corollary: Let \( X \) and \( Y \) be random variables such that \( E(X^2) \) and \( E(Y^2) \) exist. Then \( -1 \leq \rho_{X,Y} \leq 1 \). Furthermore, \( \rho_{X,Y} = \pm 1 \) if and only if there exist constants \( a \) and \( b \), \( b \neq 0 \), such that \( Y = a + bX \) with probability 1.
Proof: Let \( \mu_X = E(X) \) and \( \mu_Y = E(Y) \). Let \( U = X - \mu_X \) and \( V = Y - \mu_Y \). Then \( V(X) = E(U^2) \), \( V(Y) = E(V^2) \), and \( \text{Cov}(X,Y) = E(UV) \), and \( \rho_{X,Y}^2 = \frac{E(UV)^2}{E(U^2)E(V^2)} \). It follows from the Cauchy-Schwartz inequality that \( \rho_{X,Y}^2 \leq 1 \).

Now \( \rho_{X,Y}^2 = 1 \) iff \( \frac{E(UV)^2}{E(U^2)E(V^2)} = 1 \) iff \( E(V^2) - \frac{E(UV)^2}{E(U^2)} = 0 \) iff

\[
\frac{E(UV)^2}{E(U^2)E(V^2)} E(U^2) = 2 \frac{E(UV)^2}{E(U^2)} E(U^2) + E(V^2) = 0 \] iff

\[
E \left( \left( \frac{E(UV)}{E(U^2)} U - V \right)^2 \right) = 0 \] iff

\[ V = \frac{E(UV)}{E(U^2)} U \text{ with probability 1} \] iff

\[ Y - \mu_Y = \frac{E(UV)}{E(U^2)} (X - \mu_X) \text{ with probability 1} \] iff

\[ Y = \frac{E(UV)}{E(U^2)} X + \mu_Y - \frac{E(UV)}{E(U^2)} \mu_X \text{ with probability 1.} \]

Note that the constant \( b \) is equal to \( \frac{E(UV)}{E(U^2)} = \frac{\text{Cov}(X,Y)}{V(X)} \). It follows that \( \rho_{X,Y} = 1 \) if \( b > 0 \) and \( \rho_{X,Y} = -1 \) if \( b < 0 \).

There are several important results concerning the variance and covariance of linear combinations of random variables.

**Theorem 10.4.** Let \( X \) and \( Y \) be random variables, and let \( a \) and \( b \) be constants. Then \( \text{Cov}(aX, bY) = ab\text{Cov}(X,Y) \).

Proof: \( \text{Cov}(aX, bY) = E(abXY) - E(aX)E(bY) = ab[E(XY) - E(X)E(Y)] = ab\text{Cov}(X,Y) \).

**Theorem 10.5.** Let \( X_1, X_2, \) and \( Y \) be random variables. Then \( \text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) \).

Proof:

\[
\text{Cov}(X_1 + X_2, Y) = E[(X_1 + X_2)Y] - E(X_1 + X_2)E(Y) = E(X_1Y) + E(X_2Y) - [E(X_1) + E(X_2)]E(Y) = E(X_1Y) + E(X_2Y) - E(X_1)E(Y) - E(X_2)E(Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) \]
Theorem 10.6. Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be random variables, and let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be constants. Then
\[
\text{Cov} \left( \sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} b_j Y_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \text{Cov}(X_i, Y_j)
\]

Proof: Theorems 10.4 and 10.5 and induction.

Theorem 10.7. Let $X$ be a random variable, and let $a$ be a constant. Then
\[
\text{Cov}(X, a) = 0.
\]

Proof:
\[
\text{Cov}(X, a) = E(Xa) - E(X)E(a) = aE(X) - aE(X) = 0.
\]

The next theorem states that the correlation coefficient is unchanged by linear functions, except possibly for the sign.

Theorem 10.8. Let $X$ and $Y$ be jointly distributed, and let $a, b, c, d$ be constants. Then
\[
\rho_{aX+b,cY+d} = \pm \rho_{X,Y}, \text{ with the sign the same as the sign of the product } ac.
\]

Proof:
\[
\rho_{aX+b,cY+d} = \frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{V}(aX + b)V(cY + d)}}. \text{ Now}
\]
\[
\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y) + c\text{Cov}(b, Y) + a\text{Cov}(X, d) + \text{Cov}(b, d) = ac\text{Cov}(X, Y).
\]
Also, $V(aX + b) = a^2V(X)$ and $V(cY + d) = c^2V(Y)$, so
\[
\sqrt{\text{V}(aX + b)V(cY + d)} = |ac|\sqrt{\text{V}(X)V(Y)}.
\]
Therefore,
\[
\frac{\text{Cov}(aX + b, cY + d)}{\sqrt{\text{V}(aX + b)V(cY + d)}} = \frac{ac\text{Cov}(X, Y)}{|ac|\sqrt{\text{V}(X)V(Y)}} = \frac{ac}{|ac|}\rho_{X,Y}.
\]

A special case of Theorem 10.6 provides a formula for the variance of a sum.

Corollary: Let $X$ and $Y$ be random variables. Then
\[
V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y).
\]

Proof:
\[
V(X + Y) = \text{Cov}(X + Y, X + Y) = \text{Cov}(X, X) + \text{Cov}(X, Y) + \text{Cov}(Y, X) + \text{Cov}(Y, Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)
\]

Corollary: Let $X$ and $Y$ be random variables. Then
\[
V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y).
\]
Proof:

\[ V(X - Y) = \text{Cov}(X - Y, X - Y) \]
\[ = \text{Cov}(X, X) + \text{Cov}(X, -Y) + \text{Cov}(-Y, X) + \text{Cov}(-Y, -Y) \]
\[ = \text{Cov}(X, X) - \text{Cov}(X, Y) - \text{Cov}(Y, X) + \text{Cov}(Y, Y) \]
\[ = V(X) + V(Y) + 2\text{Cov}(X, Y) \]

**Corollary:** Let \( X \) and \( Y \) be independent variables. Then \( V(X + Y) = V(X) + V(Y) \) and \( V(X - Y) = V(X) + V(Y) \).

Following is the general case for the variance of a linear combination.

**Corollary:** Let \( X_1, \ldots, X_n \) be random variables, and let \( a_1, \ldots, a_n \) be constants. Then

\[ V(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 V(X_i) + \sum_{i=1}^n \sum_{j \neq i} a_i a_j \text{Cov}(X_i, X_j) \]

**Proof:**

\[ V(\sum_{i=1}^n a_i X_i) = \text{Cov} \left( \sum_{i=1}^n a_i X_i, \sum_{j=1}^n a_j X_j \right) \]
\[ = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j) \]
\[ = \sum_{i=1}^n a_i^2 V(X_i) + \sum_{i=1}^n \sum_{j \neq i} a_i a_j \text{Cov}(X_i, X_j) \]

**Remark 10.3.** Since \( \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) \),

\[ V(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{i=1}^n \sum_{j < i} a_i a_j \text{Cov}(X_i, X_j) \]

**Corollary:** Let \( X_1, \ldots, X_n \) be pairwise uncorrelated random variables. Then \( V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) \).

A somewhat weaker version of the corollary is useful.

**Corollary:** Let \( X_1, \ldots, X_n \) be independent random variables. Then \( V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) \).
11 Distributions of Functions of Random Variables

11.1 Distributions of Functions of Discrete Random Variables

Theorem 11.1. Let $X$ be a random variable, and let $g: \mathbb{R} \to \mathbb{R}$ be a function. Then $g(X)$ is a random variable.

Proof: Let $\Omega$ be the domain of $X$. Then clearly $g(X): \Omega \to \mathbb{R}$.

Given the distribution of a random variable $X$, it is useful to determine the distribution of a function of $X$. The method is best illustrated by example.

Example 11.1. Let $X$ be a discrete random variable with pmf $p_X(x) = P(X = x) = 1/5$ for $x = -2, -1, 0, 1, 2$, and $p_X(x) = 0$ for other values of $x$. Let $Y = X^2$. Find the pmf of $Y$.

Solution: The mass points of $Y$ are 0, 1, 4. The probabilities of these mass points are easily computed.

$P(Y = 0) = P(X^2 = 0) = P(X = 0) = 1/5$

$P(Y = 1) = P(X^2 = 1) = P(X = -1 \cup X = 1) = P(X = -1) + P(X = 1) = 2/5$

$P(Y = 4) = P(X^2 = 2) = P(X = -2 \cup X = 2) = P(X = -2) + P(X = 2) = 2/5$

The example illustrates the following theorem.

Theorem 11.2. Let $X$ be a discrete random variable with pmf $f_X(x)$. Let $Y = g(X)$ be a function of $X$. Then $Y$ is discrete with pmf $f_Y(y) = \sum_{x | g(x) = y} f_X(x)$.

11.2 Distributions of Functions of Continuous Random Variables

We illustrate the method for determining the distribution of a differentiable function of a continuous random variable. The method can be described as follows. Let $Y = g(X)$. Given the pdf of $X$ find the pdf of $Y$ as follows:

1) Express the cdf of $Y$, $F_Y(y)$ in terms of the cdf of $X$.
2) Differentiate to obtain $f_Y(y)$.

The following example illustrates the method.
Example 11.2. Let $X$ be a random variable with pdf $f_X(x) = 2x^{-3}I_{(1,\infty)}(x)$. Find the pdf of $Y = \log X$.

Solution: $F_Y(y) = P(Y \leq y) = P(\log X \leq y) = P(X \leq e^y) = F_X(e^y)$.

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(e^y) = f_X(e^y)e^y = f_X(e^y)e^y.$$ 

$$f_X(e^y)e^y = 2(e^y)^{-3}I_{(1,\infty)}(e^y)e^y = 2e^{-3y}e^yI_{(1,\infty)}(e^y) = 2e^{-2y}I_{(1,\infty)}(e^y) = 2e^{-2y}I_{(0,\infty)}(y).$$

Example 11.3. Let $X$ be a random variable with pdf $f_X(x) = \frac{1}{2}I_{(-1,1)}(x)$. Let $Y = X^2$. Find the pdf of $Y$.

Solution: $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = \begin{cases} P(-\sqrt{y} \leq X \leq \sqrt{y}) & y \geq 0 \\ 0 & y < 0 \end{cases}$

For $y < 0$, $f_Y(y) = \frac{d}{dy}0 = 0$.

For $y \geq 0$, $F_Y(y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$.

We find the pdf $f_Y(y)$ by differentiating the cdf $F_Y(y)$.

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}[F_X(\sqrt{y}) - F_X(-\sqrt{y})]$$

$$= \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}) = \frac{1}{4\sqrt{y}}I_{(-1,1)}(\sqrt{y}) + \frac{1}{4\sqrt{y}}I_{(-1,1)}(-\sqrt{y})$$

$$= \frac{1}{2\sqrt{y}}I_{(0,1)}(\sqrt{y})$$

Following is a general theorem that justifies the method. It is essentially the change of variable theorem for univariate calculus.

**Theorem 11.3.** Let $X$ have pdf $f_X(x)$ where $f_X(x)$ is continuous, and let $Y = g(X)$ where $g$ is strictly monotone and continuously differentiable. Then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right|$$

**Proof:** Case 1: $g$ is increasing. It follows that $g^{-1}$ is increasing and $\frac{d}{dy}g^{-1}(y) > 0$ for all $y$.

Now
\[ F_Y(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \text{, so} \]

\[ f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|. \]

Case 2: \( g \) is decreasing. It follows that \( g^{-1} \) is decreasing and \( \frac{d}{dy} g^{-1}(y) < 0 \) for all \( y \). Now

\[ F_Y(y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)), \text{ so} \]

\[ f_Y(y) = - \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \left( - \frac{d}{dy} g^{-1}(y) \right) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|. \]

Example 11.4. Let \( X \sim U(0, 1) \). Find the pdf of \( Y = -\log X \).

Solution:

We first find the cdf of \( Y \) in terms of the cdf of \( X \).

\[ F_Y(y) = P(Y \leq y) = P(-\log X \leq y) = P(\log X \geq -y) = P(X \geq e^{-y}) = 1 - F_X(e^{-y}) \]

We differentiate \( F_Y(y) \) to obtain \( f_Y(y) \):

\[ f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} [1 - F_X(e^{-y})] = e^{-y} f_X(e^{-y}) \]

The pdf of \( X \) is \( f_X(x) = I_{(0,1)}(x) \). Therefore \( f_Y(y) = e^{-y} I_{(0,1)}(e^{-y}) = e^{-y} I_{(0,\infty)}(y) \)

We recognize this as the exponential density with \( \lambda = 1 \). Thus \( Y = -\log X \sim \text{Exp}(1) \).

Example 11.5. Let \( X \) and \( Y \) have joint pdf \( f(x,y) = 3x I_{(0,1)}(x) I_{(0,x)}(y) \). Find the pdf of \( Z = X - Y \).

Solution:

We find the cdf \( F_Z(z) = P(Z \leq z) \). First, note that \( X \) and \( Y \) are between 0 and 1, and that \( X > Y \). Therefore \( 0 < X - Y < 1 \). It follows that \( F_Z(z) = 0 \) for \( z \leq 0 \) and \( F_Z(z) = 1 \) for \( z \geq 1 \).

Now let \( 0 < z < 1 \).

\[ F_Z(z) = P(Z \leq z) = P(X - Y \leq z) = 1 - P(X - Y > z) = 1 - \int_z^1 \int_0^{x-z} 3x \, dy \, dx \]

Now:

\[ \int_z^1 \int_0^{x-z} 3x \, dy \, dx = \int_z^1 3x^2 - 3zx \, dx = 1 - \frac{3z - z^3}{2} \]
Therefore, for \( 0 < z < 1 \), the cdf of \( Z \) is

\[
F_Z(z) = 1 - \left( 1 - \frac{3z - z^3}{2} \right) = \frac{3z - z^3}{2}
\]

The pdf is

\[
f_Z(z) = \frac{d}{dz} F_Z(z) = 0 \quad \text{for } z < 0
\]

\[
f_Z(z) = \frac{d}{dz} \left( \frac{3z - z^3}{2} \right) = \frac{3 - 3z^2}{2} \quad \text{for } 0 < z < 1
\]

\[
f_Z(z) = \frac{d}{dz} 1 = 0 \quad \text{for } z > 1
\]

### 11.3 Distributions of Sums

Let \( X \) and \( Y \) be random variables with joint pmf or pdf \( f(x, y) \). The general problem is to find the distribution of the sum \( X + Y \). The main theorem in this regard is the convolution formula.

**Theorem 11.4.** Let \( X \) and \( Y \) be continuous random variables with joint pdf \( f_{X,Y}(x, y) \). Let \( Z = X + Y \). Then

\[
f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) \, dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y) \, dy
\]

**Proof:**

\[
F_Z(z) = P(X + Y \leq z) = \int_{-\infty}^{z-x} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \, dx
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_{X,Y}(x, u-x) \, du \, dx = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} f_{X,Y}(x, u-x) \, dx \right] \, du
\]

Now differentiate:

\[
f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{d}{dz} \left[ \int_{-\infty}^{\infty} f_{X,Y}(x, u-x) \, dx \right] \, du = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) \, dx
\]

The proof of the second equality is similar.

When \( X \) and \( Y \) are independent the joint density factors, which leads to the following special case of the convolution formula.

**Corollary:** Let \( X \) and \( Y \) be independent continuous random variables and let \( Z = X + Y \). Then

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) \, dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \, dy
\]
Example 11.6. Let \( X \) and \( Y \) be independent, each distributed \( \text{Exp}(\lambda) \). Find the distribution of \( Z = X + Y \).

Solution:
First note that \( P(Z > 0) = 1 \). So \( f_Z(z) = 0 \) for \( z \leq 0 \). Now let \( z > 0 \).

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, dx
= \int_{-\infty}^{\infty} \lambda e^{-\lambda x} I_{(0,\infty)}(x) \lambda e^{-\lambda(z-x)} I_{(0,\infty)}(z-x) \, dx
= \int_{0}^{z} \lambda^2 e^{-\lambda z} \, dx
= \lambda^2 z e^{-\lambda z}
\]

The pdf of \( Z = X + Y \) is \( f_Z(z) = \lambda^2 z e^{-\lambda z} I_{(0,\infty)}(z) \), which is the pdf of \( \Gamma(2, \lambda) \).

12 Moment Generating Functions

12.1 Basic Results

Definition 12.1. Let \( X \) be a random variable. The moment generating function of \( X \) is the function \( m_X(t) = E(e^{tx}) \).

If \( X \) is discrete then \( m_X(t) = \sum e^{tx_i} P(X = x_i) \).

If \( X \) is continuous then \( m_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx \).

Note that for any random variable \( X \), \( m_X(0) = 1 \). \( m_X(t) \) may not exist for other values of \( t \).

Theorem 12.1. Let \( X \) be a random variable with moment generating function \( m_X(t) \). If \( m_X(t) \) exists in a neighborhood of 0, then for every positive integer \( r \), \( E(X^r) \) exists and \( E(X^r) = \frac{d^r}{dt^r} m_X(0) \).

Proof: \( \frac{d^r}{dt^r} m_X(t) = \frac{d^r}{dt^r} \sum e^{tx} P(X = x) = \sum x^r e^{tx} P(X = x) \).

So \( \frac{d^r}{dt^r} m_X(0) = \sum x^r P(X = x) = E(X^r) \).

\[
\frac{d^r}{dt^r} m_X(t) = \frac{d^r}{dt^r} \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx = \int_{-\infty}^{\infty} \frac{d^r}{dt^r} e^{tx} f_X(x) \, dx = \int_{-\infty}^{\infty} x^r e^{tx} f_X(x) \, dx.
\]
\[
\frac{d^r}{dt^r} m_X(0) = \int_{-\infty}^{\infty} x^r f_X(x) \, dx = E(X^r).
\]

**Theorem 12.2.** Let \(X\) and \(Y\) be random variables whose moment generating functions \(m_X(t)\) and \(m_Y(t)\) exist. If \(m_X(t) = m_Y(t)\) for all \(t\), then \(X\) and \(Y\) have the same distribution.

**Proof:** Omitted.

**Example 12.1.** Let \(X \sim \text{Exp}(\lambda)\) so the pdf of \(X\) is \(f_X(x) = \lambda e^{-\lambda x} I_{(0, \infty)}(x)\). Let \(t < \lambda\). Find the moment generating function of \(X\).

**Solution:**

\[
m_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, dx = \int_0^\infty \lambda e^{-(\lambda-t)x} \, dx = \frac{\lambda}{\lambda-t}.
\]

We can now compute the moments of this distribution. \(m_X'(t) = \lambda/(\lambda - t)^2\), \(E(X) = m_X'(0) = 1/\lambda\), \(E(X^2) = m_X''(0) = 2/\lambda^2\). In general, \(m_X^n(0) = n!/\lambda^n\).

**Example 12.2.** Let \(X \sim \Gamma(r, \lambda)\). Find the moment generating function of \(X\).

**Solution:**

\[
m_X(t) = \int_0^\infty e^{tx} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \, dx
\]

\[
= \int_0^\infty \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} \, dx
\]

\[
= \left( \frac{\lambda}{\lambda-t} \right)^r \int_0^\infty \frac{(\lambda-t)^r}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} \, dx
\] (Note that the integrand is the \(\Gamma(r, \lambda-t)\) pdf)

\[
= \left( \frac{\lambda}{\lambda-t} \right)^r
\]

**Example 12.3.** Let \(X \sim \text{Poisson}(\lambda)\). Find the moment generating function of \(X\).

**Solution:**

\[
m_X(t) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} = \sum_{x=0}^{\infty} e^{-\lambda} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda + \lambda e^t} \sum_{x=0}^{\infty} e^{-\lambda e^t} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda + \lambda e^t} = e^{-\lambda(1-e^t)}
\]
Example 12.4. Let $Z \sim N(0, 1)$. Find the moment generating function of $Z$.

Solution:

\[
m_Z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz
\]

\[
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz)} \, dz
\]

\[
= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2} \, dz
\]

(Note that the integrand is the $N(t, 1)$ pdf)

Example 12.5. Let $X \sim N(\mu, \sigma^2)$. Find the moment generating function of $X$.

Solution:

Let $Z \sim N(0, 1)$, and let $X = \mu + \sigma Z$.

Now $m_X(t) = E(e^{tX}) = E(e^{t\mu + t\sigma Z}) = e^{t\mu} E(e^{t\sigma Z}) = e^{t\mu} m_Z(t\sigma) = e^{t\mu} e^{\sigma^2 t^2/2} = e^{t\mu + \sigma^2 t^2/2}$

Moment generating functions can be used to find distributions of sums of independent random variables, as the following theorem shows.

**Theorem 12.3.** Let $X$ and $Y$ be independent random variables with moment generating functions $m_X(t)$ and $m_Y(t)$.

The moment generating function of $X + Y$ is $m_{X+Y}(t) = m_X(t)m_Y(t)$.

**Proof:**

\[
m_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = m_X(t)m_Y(t).
\]

We can use this result to find the distribution for many sums.

**Theorem 12.4.** Let $X$ and $Y$ be independent, with $X \sim \Gamma(r, \lambda)$ and $Y \sim \Gamma(s, \lambda)$.

Then $X + Y \sim \Gamma(r + s, \lambda)$.

**Proof:**

\[
m_{X+Y}(t) = m_X(t)m_Y(t) = \left(\frac{\lambda}{\lambda-t}\right)^r \left(\frac{\lambda}{\lambda-t}\right)^s = \left(\frac{\lambda}{\lambda-t}\right)^{r+s},
\]

which is the mgf of $\Gamma(r + s, \lambda)$. 

89
Corollary:
Let \( X_1, \ldots, X_n \) be independent, with \( X_i \sim \Gamma(r_i, \lambda) \). Then \( X_1 + \cdots + X_n \sim \Gamma(r_1 + \cdots + r_n, \lambda) \).

Corollary:
Let \( X_1, \ldots, X_n \) be independent, each distributed \( \text{Exp}(\lambda) \). Then \( X_1 + \cdots + X_n \sim \Gamma(n, \lambda) \).

**Theorem 12.5.** Let \( X \) and \( Y \) be independent, with \( X \sim \text{Poisson}(\lambda) \) and \( Y \sim \text{Poisson}(\theta) \). Then \( X + Y \sim \text{Poisson}(\lambda + \theta) \).

**Proof:**
\[
m_{X+Y}(t) = m_X(t)m_Y(t) = e^{-\lambda(1-e^t)} e^{-\theta(1-e^t)} = e^{-(\lambda+\theta)(1-e^t)},
\]
which is the mgf of \( \text{Poisson}(\lambda + \theta) \).

Corollary: Let \( X_1, \ldots, X_n \) be independent, with \( X_i \sim \text{Poisson}(\lambda_i) \). Then \( X_1 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_n) \).

**Theorem 12.6.** Let \( X \) and \( Y \) be independent, with \( X \sim N(\mu_1, \sigma_1^2) \) and \( Y \sim N(\mu_2, \sigma_2^2) \). Then \( X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \).

**Proof:**
\[
m_{X+Y}(t) = m_X(t)m_Y(t) = e^{t\mu_1 + \sigma_1^2 t^2/2}e^{t\mu_2 + \sigma_2^2 t^2/2} = e^{t(\mu_1 + \mu_2) + (\sigma_1^2 + \sigma_2^2) t^2/2},
\]
which is the mgf of \( N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \).

Corollary: Let \( X_1, \ldots, X_n \) be independent, with \( X_i \sim N(\mu_1, \sigma_1^2) \). Then \( X_1 + \cdots + X_n \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2) \).

### 12.2 Proof of the Central Limit Theorem

We will prove the Central Limit Theorem under the somewhat restrictive condition that the moment generating function of the \( X_i \) exists.

We will also state the following facts without proof:

1) Let \( X_1, \ldots \) be random variables with cdfs \( F_1(x), \ldots \) and moment generating functions \( m_1(t), \ldots \). If there exists a random variable \( Y \) with cdf \( F_Y(t) \) and mgf \( m_Y(t) \) such that \( \lim_{n \to \infty} m_n(t) = m_Y(t) \), then \( \lim_{n \to \infty} F_n(t) = F_Y(t) \).

2) Let \( \varepsilon_1, \ldots \) be a sequence such that \( \lim_{n \to \infty} \varepsilon_n = 0 \). Then for any real number \( x \),
\[ \lim_{n \to \infty} \left( 1 + \frac{x}{n} + \frac{\varepsilon_n}{n} \right)^n = e^x. \]

First assume that \( \mu = 0 \) and \( \sigma^2 = E(X^2) = 1 \). Then \( Z_n = \sqrt{n} \bar{X}_n \). Let \( m_n(t) \) denote the moment generating function of \( Z_n \). We will show that \( \lim_{n \to \infty} m_n(t) = e^{t^2/2} \), which is the mgf of \( N(0,1) \).

\[
    m_n(t) = E(e^{tZ}) = E(e^{t\sqrt{n}\bar{X}}) = E\left( e^{\frac{t}{\sqrt{n}} \sum_{i=1}^n X_i} \right) \\
    = E \left( \prod_{i=1}^n e^{\frac{t}{\sqrt{n}} X_i} \right) = \prod_{i=1}^n E(e^{\frac{t}{\sqrt{n}} X_i}) = E \left( e^{\frac{t}{\sqrt{n}} X} \right)^n
\]

Now \( E \left( e^{\frac{t}{\sqrt{n}} X} \right)^n = E \left( 1 + \frac{tX}{\sqrt{n}} + \frac{t^2X^2}{2n} + \frac{t^3X^3}{6n^{3/2}} \right) \), for some \( t^* \) with \( 0 < t^* < t \).

It follows that \[ E(e^{\frac{t}{\sqrt{n}} X})^n = \left( 1 + \frac{tE(X)}{\sqrt{n}} + \frac{t^2E(X^2)}{2n} + \frac{t^3E(X^3)}{6n^{3/2}} \right)^n = \left[ 1 + \frac{t^2}{2n} + \frac{t^3E(X^3)/\sqrt{n}}{6n} \right]^n.
\]

We conclude that \( \lim_{n \to \infty} m_n(t) = \lim_{n \to \infty} \left[ 1 + \frac{t^2}{2n} + o\left( \frac{1}{n} \right) \right]^n = e^{t^2/2}. \)

This establishes the result in the case that \( \mu = 0 \) and \( \sigma^2 = 1 \). To prove the general result, let \( Y_i = \frac{X_i - \mu}{\sigma} \). Then \( E(Y_i) = 0 \), and \( V(Y_i) = 1 \). Now \( Z_n = \sqrt{n} Y_n \), so it follows from the special case that the cdf of \( Z_n \) converges to \( N(0,1) \).

### 13 Sampling from the Normal Distribution

Recall that the distribution of a linear combination of normally distributed random variables is normal. From this we can compute the distribution of the mean of a normal random sample.

**Proposition 13.1.** Let \( X_1, \ldots, X_n \) be i.i.d. \( N(\mu, \sigma^2) \). Then \( \bar{X} \sim N(\mu, \sigma^2/n) \).

### 13.1 The Chi-Square Distribution

Recall:
1. The gamma function is \( \Gamma(r) = \int_0^\infty x^{r-1}e^{-x} \, dx \). It is defined for all \( r > 0 \).

2. A random variable \( X \) has the gamma distribution with parameters \( r \) and \( \lambda \) if the pdf of \( X \) is
\[
f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} \text{I}_{(0,\infty)}(x)
\]

3. Let \( k \) be a positive integer. The gamma distribution with parameters \( r = k/2 \) and \( \lambda = 1/2 \) is called the Chi-square distribution with \( k \) degrees of freedom, denoted \( \chi^2_k \).

4. If \( X \sim \chi^2_n \), then \( E(X) = n \) and \( V(X) = 2n \).

**Theorem 13.1.** Let \( Z \sim N(0, 1) \). Then \( Z^2 \sim \chi^2_1 \).

**Proof:** The \( \chi^2_1 \) distribution is the \( \Gamma(1/2, 1/2) \) distribution, whose pdf is
\[
f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2} \text{I}_{(0,\infty)}(x).
\]
We show that this is the pdf of \( Z^2 \).

Let \( x > 0 \). The cdf of \( Z^2 \) evaluated at \( x \) is
\[
F(x) = P(Z^2 \leq x) = P(Z \leq \sqrt{x}) - P(Z \leq -\sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x})
\]
where \( \Phi(x) \) is the cdf of \( N(0,1) \).

Therefore the pdf of \( Z^2 \) is
\[
f(x) = F'(x) = \frac{1}{2\sqrt{x}} \phi(\sqrt{x}) + \frac{1}{2\sqrt{x}} \phi(-\sqrt{x}) = \frac{1}{2\sqrt{x}} \sqrt{\frac{1}{\pi}} e^{-x/2} = \frac{1}{\sqrt{2\pi x}} e^{-x/2}.
\]

When \( x \leq 0 \), \( F_Z(x) = 0 \), so \( f(x) = 0 \). Therefore the pdf of \( Z^2 \) is
\[
f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2} \text{I}_{(0,\infty)}(x),
\]
which is the pdf of \( \chi^2_1 \).

Recall that if \( X_1, ..., X_n \) are independent random variables, with \( X_i \sim \Gamma(r_i, \lambda) \), then \( X_1 + \cdots + X_n \sim \Gamma(r_1 + \cdots + r_n, \lambda) \). The following proposition is immediate.

**Proposition 13.2.** Let \( X_1, ..., X_n \) are independent random variables, with \( X_i \sim \chi^2_{k_i} \) then \( X_1 + \cdots + X_n \sim \chi^2_{k_1 + \cdots + k_n} \).

**Proposition 13.3.** Let \( Z_1, ..., Z_n \) be i.i.d. \( N(0,1) \). Then \( Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n \).
Proposition 13.4. Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$.

Proof: The random variables $\frac{X_i - \mu}{\sigma}$ are i.i.d. $N(0, 1)$.

The following theorem provides an important characterization of the normal distribution.

Theorem 13.2. Let $Z_1, ..., Z_n$ be i.i.d. $N(0, 1)$. Then $\bar{Z}$ and $\sum_{i=1}^{n} (Z_i - \bar{Z})^2$ are independent, and $\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sim \chi_{n-1}^2$.

Proof: Omitted.

Corollary: Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $\bar{X}$ and $\sum_{i=1}^{n} (X_i - \bar{X})^2$ are independent, and $\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$.

Proof: Let $X_i = \sigma Z_i + \mu$ where $Z_1, ..., Z_n$ are i.i.d. $N(0, 1)$. Then $\bar{X} = \sigma \bar{Z} + \mu$ and $\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} = \sum_{i=1}^{n} (Z_i - \bar{Z})^2$.

Corollary: Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$. Let $\bar{X}$ and $s^2$ denote the sample mean and sample variance, respectively. Then $\bar{X}$ and $s^2$ are independent.

Corollary: Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$. Then $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$.

13.2 The $F$-distribution

Definition 13.1. Let $U$ and $V$ be independent, with $U \sim \chi_m^2$ and $V \sim \chi_n^2$. Then the random variable $\frac{U/m}{V/n}$ has the $F$ distribution with $m$ and $n$ degrees of freedom, denoted $F_{m,n}$.

Note that if $X \sim F_{m,n}$ then $1/X \sim F_{n,m}$.

Theorem 13.3. Let $X_1, ..., X_m$ be i.i.d. $N(\mu_X, \sigma^2)$ and let $Y_1, ..., Y_n$ be i.i.d. $N(\mu_Y, \sigma^2)$. Let $s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{m-1}$ and $s_Y^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$ be the sample variances of the $X$s and the $Y$s, respectively. Then the random variable $\frac{s_X^2}{s_Y^2} \sim F_{m-1,n-1}$. 

93
Proof: \( \frac{(m-1)s_X^2}{\sigma^2} \sim \chi^2_{m-1}, \frac{(n-1)s_Y^2}{\sigma^2} \sim \chi^2_{n-1} \), and these two random variables are independent. It follows that \( \frac{s_X^2/\sigma^2}{s_Y^2/\sigma^2} = \frac{s_X^2}{s_Y^2} \sim F_{m-1,n-1} \).

13.3 The Student’s t distribution

Definition 13.2. Let \( Z \sim N(0,1) \), and let \( U \sim \chi^2_n \), where \( Z \) and \( U \) are independent. Then the random variable \( T = \frac{Z}{\sqrt{U/n}} \) has the Student’s t distribution with \( n \) degrees of freedom, denoted \( t_n \).

Theorem 13.4. Let \( X_1, ..., X_n \) be i.i.d. \( N(\mu, \sigma^2) \). Let \( s^2 \) denote the sample variance. Then the random variable \( \sqrt{n}(\bar{X} - \mu)/s \) has the Student’s t distribution with \( n-1 \) degrees of freedom.

Proof: The random variable \( \frac{\sqrt{n}(\bar{X} - \mu)}{s} \) is distributed \( N(0,1) \), and is independent of \( \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \), which is distributed \( \chi^2_{n-1} \).

Therefore \( \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\sigma^2}} \sim t_{n-1} \).

But \( \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)\sigma^2}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}} \sim \frac{\sqrt{n}(\bar{X} - \mu)}{s} \).

Note that the assumption of normality is crucial in the above theorem.

14 Asymptotics

In many cases it is difficult to make a precise evaluation of a situation for a specific sample size. In many of these cases, however, precise results are available in the limit as the sample size approaches infinity. Such results are known as asymptotic results, and they form the basis of much of modern statistical theory. Following is one of the two most basic such results.
14.1 Convergence in Probability

There are several modes of convergence that are important. The first is called **convergence in probability**.

**Definition 14.1.** Let $X_1, \ldots, X_n$ be an infinite sequence of random variables. Let $c$ be a constant. The sequence $Y_1, \ldots, Y_n$ converges in probability to $c$ if $\forall \varepsilon > 0, \lim_{n \to \infty} P(|Y_n - c| > \varepsilon) = 0$. We write $Y_n \overset{P}{\to} c$.

Equivalently, $Y_n \overset{P}{\to} Z$ if $\forall \varepsilon > 0, \lim_{n \to \infty} P(|Y_n - Z| < \varepsilon) = 1$

Note that convergence of a sequence of real numbers is a special case of convergence in probability:

**Example 14.1.** Let $X_1, \ldots$ be i.i.d. $N(\mu, 1)$. Let $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$ be the average of the first $n$ Xs. Then $\bar{X}_n \sim N(\mu, 1/n)$. We show that $\bar{X} \overset{P}{\to} \mu$.

First note that $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$. Now let $\varepsilon > 0$.

$$P(|\bar{X} - \mu| < \varepsilon) = P(-\varepsilon < \bar{X} - \mu < \varepsilon)$$

$$= P(-\sqrt{n}\varepsilon < \sqrt{n}(\bar{X} - \mu) < \sqrt{n}\varepsilon)$$

$$= \Phi(\sqrt{n}\varepsilon) - \Phi(-\sqrt{n}\varepsilon)$$

So $\lim_{n \to \infty} P(|\bar{X} - \mu| < \varepsilon) = \lim_{n \to \infty} \Phi(\sqrt{n}\varepsilon) - \Phi(-\sqrt{n}\varepsilon) = 1 - 0 = 1$.

14.2 The Weak Law of Large Numbers

**Theorem 14.1.** The Weak Law of Large Numbers: Let $X_1, \ldots, X_n$ be independent and identically distributed with $E(X_i) = \mu$. For each positive integer $n$, let $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$ be the average of the first $n$ of the $X_i$. Then for all $\varepsilon > 0$, $\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$. In other words $\bar{X}_n \overset{P}{\to} \mu$.

**Proof:** We prove the theorem under the slightly more restrictive condition that $V(X) = \sigma^2 < \infty$. The proof follows immediately from Markov’s inequality.

For any $\varepsilon > 0$, $P(|\bar{X}_n - \mu| > \varepsilon) = P([X_n - \mu]^2 > \varepsilon^2] \leq \frac{E[(X_n - \mu)^2]}{\varepsilon^2} \leq \frac{\sigma^2}{n\varepsilon^2}$.

Therefore $\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) \leq \lim_{n \to \infty} \frac{\sigma^2}{n\varepsilon^2} = 0$. 

95
The usefulness of the weak law of large numbers is that it tells us that for a reasonably large sample size, the sample mean probably close to $\mu$. Therefore $\bar{X}$ is a good estimator for $\mu$ when $\mu$ is unknown.

It is natural to ask how large the sample size must be before one can be reasonably sure, in practice, that $\bar{X}$ is close to $\mu$. The necessary sample size depends on the variance $\sigma^2$. The larger $\sigma^2$ is, the larger the sample size needs to be.

Convergence in probability behaves in some ways like convergence of sequences of real numbers. We now explore the extent of the similarities.

**Theorem 14.2.** Let $X_1, \ldots$ be a sequence of random variables, and let $c$ be a constant with $X_n \xrightarrow{P} c$. Let $k$ be any constant. Then $kX_n \xrightarrow{P} kc$.

**Proof:**
Let $\varepsilon > 0$. $\lim_{n \to \infty} P(|kX_n - kc| > \varepsilon) = \lim_{n \to \infty} P(|X_n - c| > \varepsilon/|k|) = 0$.

**Theorem 14.3.** Let $X_1, \ldots$ and $Y_1, \ldots$ be sequences of random variables, and $a$ and $b$ constants with $X_n \xrightarrow{P} a$ and $Y_n \xrightarrow{P} b$. Then $X_n + Y_n \xrightarrow{P} a + b$.

**Proof:**
Let $\varepsilon > 0$.

\[
P(|X_n + Y_n - (a - b)| > \varepsilon) < P(|X_n - a| + |Y_n - b| > \varepsilon)
\]
\[
< P(|X_n - a| > \varepsilon/2 \cap |Y_n - b| > \varepsilon/2)
\]
\[
< P(|X_n - a| > \varepsilon/2)
\]

Therefore $\lim_{n \to \infty} P(|X_n + Y_n - (a - b)| > \varepsilon) \leq \lim_{n \to \infty} P(|X_n - a| > \varepsilon/2) = 0$.

**Corollary:**
Let $Y_n \xrightarrow{P} b$, let $a_n \to a$, and $Z_n \xrightarrow{P} c$, where $a, b, c$ are constants. Then $a_nY_n + Z_n \xrightarrow{P} ab + c$. 

96
Example 14.2. Let \( X_1, \ldots, \) be i.i.d., \( E(X_i) = \mu \). Let \( W_n = \frac{X_1 + \ldots + X_n}{n} + \frac{X_1}{\sqrt{n}} \). Show that \( W_n \xrightarrow{P} \mu \).

Solution: Let \( a_n = n/(n - 1), Z_n = X_1/\sqrt{n} \). Then \( W_n = a_n \bar{X}_n + Z_n \). Now \( a_n \to 1, Z_n \xrightarrow{P} 0 \), and \( \bar{X}_n \xrightarrow{P} \mu \), so \( W_n \xrightarrow{P} 1 \cdot \mu + 0 = \mu \).

The corollary can be used to prove the following important result concerning the usual estimator of variance.

Theorem 14.4. Let \( X_1, \ldots, \) be i.i.d., \( V(X_i) = \sigma^2 \). Let \( s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1} \). Then \( s^2 \xrightarrow{P} \sigma^2 \).

Proof:

\[
s^2 = \frac{1}{n - 1} \sum_{i=1}^{n}(X_i - \bar{X})^2
= \frac{1}{n - 1} \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2
= \frac{1}{n - 1} \sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{n - 1}(\bar{X} - \mu)^2.
\]

For each \( i \), let \( Y_i = (X_i - \mu)^2 \).

Then \( E(Y_i) = \sigma^2, \bar{Y} = \frac{1}{n} \sum_{i=1}^{n}(X_i - \mu)^2, \) and \( s^2 = \frac{1}{n - 1} \bar{Y} + \frac{n}{n - 1}(\bar{X} - \mu)^2 \).

Since \( \frac{n}{n - 1} \to 1 \), and by the weak law of large numbers \( \bar{Y} \xrightarrow{P} \sigma^2, \) and \( \bar{X} - \mu \xrightarrow{P} 0 \), it follows that \( s^2 \xrightarrow{P} \sigma^2 \).

Theorem 14.5. Let \( X_1, \ldots, \) be a sequence of random variables, and \( c \) a constant with \( X_n \xrightarrow{P} c \). Let \( g \) be any continuous function. Then \( g(X_n) \xrightarrow{P} g(c) \).

Proof: Let \( \varepsilon > 0 \). Let \( \delta > 0 \) be such that \( |X_n - c| < \delta \Rightarrow |g(X_n) - g(c)| < \varepsilon \).

Then \( \lim_{n \to \infty} P(\{|g(X_n) - g(c)| < \varepsilon\}) \geq \lim_{n \to \infty} P(|X_n - c| < \delta) = 1 \).

Example 14.3. Let \( X_1, \ldots, \) be i.i.d., \( E(X_i) = \mu, V(X_i) = \sigma^2 \). By the weak law of large numbers, we know that \( \bar{X} \xrightarrow{P} \mu \). It follows that \( g(\bar{X}) \xrightarrow{P} g(\mu) \) for any continuous function \( g \); e.g., \( \bar{X}^2 \xrightarrow{P} \mu^2, \log |\bar{X}| \xrightarrow{P} \log |\mu|, \) and \( \sin(\bar{X}) \xrightarrow{P} \sin(\mu) \).
14.3 Convergence in Law

The other mode of convergence that we will study is called convergence in distribution, or convergence in law.

**Definition 14.2.** Let $Y_1, \ldots, \text{be a sequence of random variables with cumulative distribution functions } F_1, \ldots \text{. Let } Z \text{ be a continuous random variable with cumulative distribution function } G \text{. The sequence } Y_1, \ldots \text{ is said to **converge in law** to } Z \text{, or to converge in law to } G \text{, if } \lim_{n \to \infty} F_n(x) = G(x). \text{ We write } Y_n \xrightarrow{L} Z \text{ or } Y_n \xrightarrow{L} G \text{. The distribution } G \text{ is called the **limiting distribution** or **asymptotic distribution** of the sequence } Y_1, \ldots \text{.}

**Remark 14.1.** The cdfs $F_n(x)$ are said to **converge weakly** to $G(x)$.

**Remark 14.2.** The sequence $Y_1, \ldots, \text{ converges in law to } Z \text{ if and only if } \lim_{n \to \infty} P(Y_n \leq x) = P(Z \leq x) \text{ for all } x.$
Table I: Cumulative Normal Distribution

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.6</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>-3.5</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>-3.4</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>-3.3</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>-3.2</td>
<td>0.007</td>
<td>0.007</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.005</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>-3.1</td>
<td>0.010</td>
<td>0.009</td>
<td>0.009</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.008</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>-3.0</td>
<td>0.013</td>
<td>0.013</td>
<td>0.013</td>
<td>0.012</td>
<td>0.012</td>
<td>0.011</td>
<td>0.011</td>
<td>0.011</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td>-2.9</td>
<td>0.019</td>
<td>0.018</td>
<td>0.018</td>
<td>0.017</td>
<td>0.016</td>
<td>0.016</td>
<td>0.015</td>
<td>0.015</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>-2.8</td>
<td>0.026</td>
<td>0.025</td>
<td>0.024</td>
<td>0.023</td>
<td>0.022</td>
<td>0.021</td>
<td>0.021</td>
<td>0.020</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td>-2.7</td>
<td>0.035</td>
<td>0.034</td>
<td>0.033</td>
<td>0.032</td>
<td>0.031</td>
<td>0.030</td>
<td>0.029</td>
<td>0.028</td>
<td>0.027</td>
<td>0.026</td>
</tr>
<tr>
<td>-2.6</td>
<td>0.047</td>
<td>0.045</td>
<td>0.044</td>
<td>0.043</td>
<td>0.042</td>
<td>0.041</td>
<td>0.040</td>
<td>0.039</td>
<td>0.038</td>
<td>0.037</td>
</tr>
<tr>
<td>-2.5</td>
<td>0.062</td>
<td>0.060</td>
<td>0.059</td>
<td>0.057</td>
<td>0.055</td>
<td>0.054</td>
<td>0.052</td>
<td>0.051</td>
<td>0.049</td>
<td>0.048</td>
</tr>
<tr>
<td>-2.4</td>
<td>0.082</td>
<td>0.080</td>
<td>0.078</td>
<td>0.075</td>
<td>0.073</td>
<td>0.071</td>
<td>0.069</td>
<td>0.068</td>
<td>0.066</td>
<td>0.064</td>
</tr>
<tr>
<td>-2.3</td>
<td>0.107</td>
<td>0.104</td>
<td>0.102</td>
<td>0.099</td>
<td>0.096</td>
<td>0.094</td>
<td>0.091</td>
<td>0.089</td>
<td>0.087</td>
<td>0.084</td>
</tr>
<tr>
<td>-2.2</td>
<td>0.139</td>
<td>0.136</td>
<td>0.132</td>
<td>0.129</td>
<td>0.125</td>
<td>0.122</td>
<td>0.119</td>
<td>0.116</td>
<td>0.113</td>
<td>0.110</td>
</tr>
<tr>
<td>-2.1</td>
<td>0.179</td>
<td>0.174</td>
<td>0.170</td>
<td>0.166</td>
<td>0.162</td>
<td>0.158</td>
<td>0.154</td>
<td>0.150</td>
<td>0.146</td>
<td>0.143</td>
</tr>
<tr>
<td>-2.0</td>
<td>0.228</td>
<td>0.222</td>
<td>0.217</td>
<td>0.212</td>
<td>0.207</td>
<td>0.202</td>
<td>0.197</td>
<td>0.192</td>
<td>0.188</td>
<td>0.183</td>
</tr>
<tr>
<td>-1.9</td>
<td>0.287</td>
<td>0.281</td>
<td>0.274</td>
<td>0.268</td>
<td>0.262</td>
<td>0.256</td>
<td>0.250</td>
<td>0.244</td>
<td>0.239</td>
<td>0.233</td>
</tr>
<tr>
<td>-1.8</td>
<td>0.359</td>
<td>0.351</td>
<td>0.344</td>
<td>0.336</td>
<td>0.329</td>
<td>0.322</td>
<td>0.314</td>
<td>0.307</td>
<td>0.301</td>
<td>0.294</td>
</tr>
<tr>
<td>-1.7</td>
<td>0.446</td>
<td>0.436</td>
<td>0.427</td>
<td>0.418</td>
<td>0.409</td>
<td>0.401</td>
<td>0.392</td>
<td>0.384</td>
<td>0.375</td>
<td>0.367</td>
</tr>
<tr>
<td>-1.6</td>
<td>0.548</td>
<td>0.537</td>
<td>0.526</td>
<td>0.516</td>
<td>0.505</td>
<td>0.495</td>
<td>0.485</td>
<td>0.475</td>
<td>0.465</td>
<td>0.455</td>
</tr>
<tr>
<td>-1.5</td>
<td>0.668</td>
<td>0.655</td>
<td>0.643</td>
<td>0.630</td>
<td>0.618</td>
<td>0.606</td>
<td>0.594</td>
<td>0.582</td>
<td>0.571</td>
<td>0.559</td>
</tr>
<tr>
<td>-1.4</td>
<td>0.808</td>
<td>0.793</td>
<td>0.778</td>
<td>0.764</td>
<td>0.749</td>
<td>0.735</td>
<td>0.721</td>
<td>0.708</td>
<td>0.694</td>
<td>0.681</td>
</tr>
<tr>
<td>-1.3</td>
<td>0.968</td>
<td>0.951</td>
<td>0.934</td>
<td>0.918</td>
<td>0.901</td>
<td>0.885</td>
<td>0.869</td>
<td>0.853</td>
<td>0.838</td>
<td>0.823</td>
</tr>
<tr>
<td>-1.2</td>
<td>1.151</td>
<td>1.131</td>
<td>1.112</td>
<td>1.093</td>
<td>1.075</td>
<td>1.056</td>
<td>1.038</td>
<td>1.020</td>
<td>1.003</td>
<td>0.985</td>
</tr>
<tr>
<td>-1.1</td>
<td>1.357</td>
<td>1.335</td>
<td>1.314</td>
<td>1.292</td>
<td>1.271</td>
<td>1.251</td>
<td>1.230</td>
<td>1.210</td>
<td>1.190</td>
<td>1.170</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.587</td>
<td>1.562</td>
<td>1.539</td>
<td>1.515</td>
<td>1.492</td>
<td>1.469</td>
<td>1.446</td>
<td>1.423</td>
<td>1.401</td>
<td>1.379</td>
</tr>
<tr>
<td>-0.9</td>
<td>1.841</td>
<td>1.814</td>
<td>1.788</td>
<td>1.762</td>
<td>1.736</td>
<td>1.711</td>
<td>1.685</td>
<td>1.660</td>
<td>1.635</td>
<td>1.611</td>
</tr>
<tr>
<td>-0.8</td>
<td>2.119</td>
<td>2.090</td>
<td>2.061</td>
<td>2.033</td>
<td>2.005</td>
<td>1.977</td>
<td>1.949</td>
<td>1.922</td>
<td>1.894</td>
<td>1.867</td>
</tr>
<tr>
<td>-0.7</td>
<td>2.420</td>
<td>2.389</td>
<td>2.358</td>
<td>2.327</td>
<td>2.296</td>
<td>2.266</td>
<td>2.236</td>
<td>2.206</td>
<td>2.177</td>
<td>2.148</td>
</tr>
<tr>
<td>-0.6</td>
<td>2.743</td>
<td>2.709</td>
<td>2.676</td>
<td>2.643</td>
<td>2.611</td>
<td>2.578</td>
<td>2.546</td>
<td>2.514</td>
<td>2.483</td>
<td>2.451</td>
</tr>
<tr>
<td>-0.5</td>
<td>3.085</td>
<td>3.050</td>
<td>3.015</td>
<td>2.981</td>
<td>2.946</td>
<td>2.912</td>
<td>2.877</td>
<td>2.843</td>
<td>2.810</td>
<td>2.776</td>
</tr>
<tr>
<td>-0.4</td>
<td>3.446</td>
<td>3.409</td>
<td>3.372</td>
<td>3.336</td>
<td>3.300</td>
<td>3.264</td>
<td>3.228</td>
<td>3.192</td>
<td>3.156</td>
<td>3.121</td>
</tr>
</tbody>
</table>
Table I: Cumulative Normal Distribution (continued)

<table>
<thead>
<tr>
<th>$z$</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>.5000</td>
<td>.5040</td>
<td>.5080</td>
<td>.5120</td>
<td>.5160</td>
<td>.5199</td>
<td>.5239</td>
<td>.5279</td>
<td>.5319</td>
<td>.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>.5398</td>
<td>.5438</td>
<td>.5478</td>
<td>.5517</td>
<td>.5557</td>
<td>.5596</td>
<td>.5636</td>
<td>.5675</td>
<td>.5714</td>
<td>.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>.5793</td>
<td>.5832</td>
<td>.5871</td>
<td>.5910</td>
<td>.5948</td>
<td>.5987</td>
<td>.6026</td>
<td>.6064</td>
<td>.6103</td>
<td>.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>.6179</td>
<td>.6217</td>
<td>.6255</td>
<td>.6293</td>
<td>.6331</td>
<td>.6368</td>
<td>.6406</td>
<td>.6443</td>
<td>.6480</td>
<td>.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>.6554</td>
<td>.6591</td>
<td>.6628</td>
<td>.6664</td>
<td>.6700</td>
<td>.6736</td>
<td>.6772</td>
<td>.6808</td>
<td>.6844</td>
<td>.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>.6915</td>
<td>.6950</td>
<td>.6985</td>
<td>.7020</td>
<td>.7054</td>
<td>.7088</td>
<td>.7123</td>
<td>.7157</td>
<td>.7190</td>
<td>.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>.7257</td>
<td>.7291</td>
<td>.7324</td>
<td>.7357</td>
<td>.7389</td>
<td>.7422</td>
<td>.7454</td>
<td>.7486</td>
<td>.7517</td>
<td>.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>.7580</td>
<td>.7611</td>
<td>.7642</td>
<td>.7673</td>
<td>.7704</td>
<td>.7734</td>
<td>.7764</td>
<td>.7794</td>
<td>.7823</td>
<td>.7852</td>
</tr>
<tr>
<td>0.8</td>
<td>.7881</td>
<td>.7910</td>
<td>.7939</td>
<td>.7967</td>
<td>.7995</td>
<td>.8023</td>
<td>.8051</td>
<td>.8078</td>
<td>.8106</td>
<td>.8133</td>
</tr>
<tr>
<td>0.9</td>
<td>.8159</td>
<td>.8186</td>
<td>.8212</td>
<td>.8238</td>
<td>.8264</td>
<td>.8289</td>
<td>.8315</td>
<td>.8340</td>
<td>.8365</td>
<td>.8389</td>
</tr>
<tr>
<td>1.0</td>
<td>.8413</td>
<td>.8438</td>
<td>.8461</td>
<td>.8485</td>
<td>.8508</td>
<td>.8531</td>
<td>.8554</td>
<td>.8577</td>
<td>.8599</td>
<td>.8621</td>
</tr>
<tr>
<td>1.1</td>
<td>.8643</td>
<td>.8665</td>
<td>.8686</td>
<td>.8708</td>
<td>.8729</td>
<td>.8749</td>
<td>.8770</td>
<td>.8790</td>
<td>.8810</td>
<td>.8830</td>
</tr>
<tr>
<td>1.2</td>
<td>.8849</td>
<td>.8869</td>
<td>.8888</td>
<td>.8907</td>
<td>.8925</td>
<td>.8944</td>
<td>.8962</td>
<td>.8980</td>
<td>.8997</td>
<td>.9015</td>
</tr>
<tr>
<td>1.3</td>
<td>.9032</td>
<td>.9049</td>
<td>.9066</td>
<td>.9082</td>
<td>.9099</td>
<td>.9115</td>
<td>.9131</td>
<td>.9147</td>
<td>.9162</td>
<td>.9177</td>
</tr>
<tr>
<td>1.4</td>
<td>.9192</td>
<td>.9207</td>
<td>.9222</td>
<td>.9236</td>
<td>.9251</td>
<td>.9265</td>
<td>.9279</td>
<td>.9292</td>
<td>.9306</td>
<td>.9319</td>
</tr>
<tr>
<td>1.5</td>
<td>.9332</td>
<td>.9345</td>
<td>.9357</td>
<td>.9370</td>
<td>.9382</td>
<td>.9394</td>
<td>.9406</td>
<td>.9418</td>
<td>.9429</td>
<td>.9441</td>
</tr>
<tr>
<td>1.6</td>
<td>.9452</td>
<td>.9463</td>
<td>.9474</td>
<td>.9484</td>
<td>.9495</td>
<td>.9505</td>
<td>.9515</td>
<td>.9525</td>
<td>.9535</td>
<td>.9545</td>
</tr>
<tr>
<td>1.7</td>
<td>.9554</td>
<td>.9564</td>
<td>.9573</td>
<td>.9582</td>
<td>.9591</td>
<td>.9599</td>
<td>.9608</td>
<td>.9616</td>
<td>.9625</td>
<td>.9633</td>
</tr>
<tr>
<td>1.8</td>
<td>.9641</td>
<td>.9649</td>
<td>.9656</td>
<td>.9664</td>
<td>.9671</td>
<td>.9678</td>
<td>.9686</td>
<td>.9693</td>
<td>.9699</td>
<td>.9706</td>
</tr>
<tr>
<td>1.9</td>
<td>.9713</td>
<td>.9719</td>
<td>.9726</td>
<td>.9732</td>
<td>.9738</td>
<td>.9744</td>
<td>.9750</td>
<td>.9756</td>
<td>.9761</td>
<td>.9767</td>
</tr>
<tr>
<td>2.0</td>
<td>.9772</td>
<td>.9778</td>
<td>.9783</td>
<td>.9788</td>
<td>.9793</td>
<td>.9798</td>
<td>.9803</td>
<td>.9808</td>
<td>.9812</td>
<td>.9817</td>
</tr>
<tr>
<td>2.1</td>
<td>.9821</td>
<td>.9826</td>
<td>.9830</td>
<td>.9834</td>
<td>.9838</td>
<td>.9842</td>
<td>.9846</td>
<td>.9850</td>
<td>.9854</td>
<td>.9857</td>
</tr>
<tr>
<td>2.2</td>
<td>.9861</td>
<td>.9864</td>
<td>.9868</td>
<td>.9871</td>
<td>.9875</td>
<td>.9878</td>
<td>.9881</td>
<td>.9884</td>
<td>.9887</td>
<td>.9890</td>
</tr>
<tr>
<td>2.3</td>
<td>.9893</td>
<td>.9896</td>
<td>.9898</td>
<td>.9901</td>
<td>.9904</td>
<td>.9906</td>
<td>.9909</td>
<td>.9911</td>
<td>.9913</td>
<td>.9916</td>
</tr>
<tr>
<td>2.4</td>
<td>.9918</td>
<td>.9920</td>
<td>.9922</td>
<td>.9925</td>
<td>.9927</td>
<td>.9929</td>
<td>.9931</td>
<td>.9932</td>
<td>.9934</td>
<td>.9936</td>
</tr>
<tr>
<td>2.5</td>
<td>.9938</td>
<td>.9940</td>
<td>.9941</td>
<td>.9943</td>
<td>.9945</td>
<td>.9946</td>
<td>.9948</td>
<td>.9949</td>
<td>.9951</td>
<td>.9952</td>
</tr>
<tr>
<td>2.6</td>
<td>.9953</td>
<td>.9955</td>
<td>.9956</td>
<td>.9957</td>
<td>.9959</td>
<td>.9960</td>
<td>.9961</td>
<td>.9962</td>
<td>.9963</td>
<td>.9964</td>
</tr>
<tr>
<td>2.7</td>
<td>.9965</td>
<td>.9966</td>
<td>.9967</td>
<td>.9968</td>
<td>.9969</td>
<td>.9970</td>
<td>.9971</td>
<td>.9972</td>
<td>.9973</td>
<td>.9974</td>
</tr>
<tr>
<td>2.8</td>
<td>.9974</td>
<td>.9975</td>
<td>.9976</td>
<td>.9977</td>
<td>.9977</td>
<td>.9978</td>
<td>.9979</td>
<td>.9979</td>
<td>.9980</td>
<td>.9981</td>
</tr>
<tr>
<td>2.9</td>
<td>.9981</td>
<td>.9982</td>
<td>.9982</td>
<td>.9983</td>
<td>.9984</td>
<td>.9984</td>
<td>.9985</td>
<td>.9985</td>
<td>.9986</td>
<td>.9986</td>
</tr>
<tr>
<td>3.0</td>
<td>.9987</td>
<td>.9987</td>
<td>.9987</td>
<td>.9988</td>
<td>.9988</td>
<td>.9988</td>
<td>.9989</td>
<td>.9989</td>
<td>.9989</td>
<td>.9990</td>
</tr>
<tr>
<td>3.1</td>
<td>.9990</td>
<td>.9990</td>
<td>.9991</td>
<td>.9991</td>
<td>.9991</td>
<td>.9992</td>
<td>.9992</td>
<td>.9992</td>
<td>.9992</td>
<td>.9993</td>
</tr>
<tr>
<td>3.2</td>
<td>.9993</td>
<td>.9993</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9994</td>
<td>.9995</td>
<td>.9995</td>
<td>.9995</td>
</tr>
<tr>
<td>3.3</td>
<td>.9995</td>
<td>.9995</td>
<td>.9995</td>
<td>.9995</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9996</td>
<td>.9997</td>
</tr>
<tr>
<td>3.4</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
<td>.9997</td>
</tr>
<tr>
<td>3.5</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
<td>.9998</td>
</tr>
<tr>
<td>3.6</td>
<td>.9998</td>
<td>.9998</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
<td>.9999</td>
</tr>
</tbody>
</table>
Table II: Upper Percentage Points for the Student’s t Distribution

<table>
<thead>
<tr>
<th>( \nu )</th>
<th>.40</th>
<th>.25</th>
<th>.10</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
<th>.001</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.325</td>
<td>1.000</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.309</td>
<td>636.619</td>
</tr>
<tr>
<td>2</td>
<td>0.289</td>
<td>0.816</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>22.327</td>
<td>31.599</td>
</tr>
<tr>
<td>3</td>
<td>0.277</td>
<td>0.765</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.812</td>
<td>10.125</td>
<td>12.924</td>
</tr>
<tr>
<td>4</td>
<td>0.271</td>
<td>0.741</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>0.267</td>
<td>0.727</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>5.939</td>
<td>6.869</td>
</tr>
<tr>
<td>6</td>
<td>0.265</td>
<td>0.718</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>5.208</td>
<td>5.959</td>
</tr>
<tr>
<td>7</td>
<td>0.263</td>
<td>0.711</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td>8</td>
<td>0.262</td>
<td>0.706</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>0.261</td>
<td>0.703</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>4.297</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>0.260</td>
<td>0.700</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>4.144</td>
<td>4.587</td>
</tr>
<tr>
<td>11</td>
<td>0.260</td>
<td>0.697</td>
<td>1.363</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
<td>4.025</td>
<td>4.437</td>
</tr>
<tr>
<td>12</td>
<td>0.259</td>
<td>0.695</td>
<td>1.356</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
<td>3.930</td>
<td>4.318</td>
</tr>
<tr>
<td>13</td>
<td>0.259</td>
<td>0.694</td>
<td>1.350</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
<td>3.852</td>
<td>4.221</td>
</tr>
<tr>
<td>14</td>
<td>0.258</td>
<td>0.692</td>
<td>1.345</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
<td>3.787</td>
<td>4.140</td>
</tr>
<tr>
<td>15</td>
<td>0.258</td>
<td>0.691</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
<td>3.733</td>
<td>4.073</td>
</tr>
<tr>
<td>16</td>
<td>0.258</td>
<td>0.690</td>
<td>1.337</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
<td>3.686</td>
<td>4.015</td>
</tr>
<tr>
<td>17</td>
<td>0.257</td>
<td>0.689</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.900</td>
<td>3.646</td>
<td>3.965</td>
</tr>
<tr>
<td>18</td>
<td>0.257</td>
<td>0.688</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.610</td>
<td>3.922</td>
</tr>
<tr>
<td>19</td>
<td>0.257</td>
<td>0.688</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
<td>3.579</td>
<td>3.883</td>
</tr>
<tr>
<td>20</td>
<td>0.257</td>
<td>0.687</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.552</td>
<td>3.850</td>
</tr>
<tr>
<td>21</td>
<td>0.257</td>
<td>0.686</td>
<td>1.323</td>
<td>1.721</td>
<td>2.080</td>
<td>2.518</td>
<td>2.831</td>
<td>3.527</td>
<td>3.819</td>
</tr>
<tr>
<td>22</td>
<td>0.256</td>
<td>0.686</td>
<td>1.321</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
<td>3.505</td>
<td>3.792</td>
</tr>
<tr>
<td>23</td>
<td>0.256</td>
<td>0.685</td>
<td>1.319</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
<td>3.485</td>
<td>3.768</td>
</tr>
<tr>
<td>24</td>
<td>0.256</td>
<td>0.685</td>
<td>1.318</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
<td>3.467</td>
<td>3.745</td>
</tr>
<tr>
<td>25</td>
<td>0.256</td>
<td>0.684</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
<td>3.450</td>
<td>3.725</td>
</tr>
<tr>
<td>26</td>
<td>0.256</td>
<td>0.684</td>
<td>1.315</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
<td>3.435</td>
<td>3.707</td>
</tr>
<tr>
<td>27</td>
<td>0.256</td>
<td>0.684</td>
<td>1.314</td>
<td>1.703</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
<td>3.421</td>
<td>3.690</td>
</tr>
<tr>
<td>28</td>
<td>0.256</td>
<td>0.683</td>
<td>1.313</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
<td>3.408</td>
<td>3.674</td>
</tr>
<tr>
<td>29</td>
<td>0.256</td>
<td>0.683</td>
<td>1.311</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
<td>3.396</td>
<td>3.659</td>
</tr>
<tr>
<td>30</td>
<td>0.256</td>
<td>0.683</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>35</td>
<td>0.255</td>
<td>0.682</td>
<td>1.306</td>
<td>1.690</td>
<td>2.030</td>
<td>2.438</td>
<td>2.724</td>
<td>3.340</td>
<td>3.591</td>
</tr>
<tr>
<td>40</td>
<td>0.255</td>
<td>0.681</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>60</td>
<td>0.254</td>
<td>0.679</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>120</td>
<td>0.254</td>
<td>0.677</td>
<td>1.289</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
<td>3.160</td>
<td>3.373</td>
</tr>
<tr>
<td>( \infty )</td>
<td>0.253</td>
<td>0.674</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.090</td>
<td>3.291</td>
</tr>
</tbody>
</table>