Intermediate probability notes

Attila Máté
Brooklyn College of the City University of New York
March 12, 2020

Contents

1 Introduction 6

2 Sets 6
2.1 The empty set 7
2.2 Relations between sets 7
2.3 Set operations 7
2.4 Venn diagrams 9
2.5 The power set of a set 9

3 Combinatorics 9
3.1 Permutations: lists of length \( k \) of \( n \) items 9
   3.1.1 Permutations 10
   3.1.2 Permutations as mappings 10
3.2 Combinations 10
   3.2.1 The binomial theorem 11
3.3 Partitions 11
3.4 Counting functions from a finite set into another 12
3.5 The generalized distributive law 12
3.6 The multinomial theorem 13
3.7 Lists with repetition, order counts 13
3.8 Combinations with repetitions 13
3.9 The number of solutions of an equation 14
3.10 Reading 14
3.11 Homework 14

4 Probability: the Kolmogorov probability model 14
4.1 Outcomes and the sample space 14
4.2 Events and probabilities 15
4.3 Finite uniform probability spaces 15

*Written for the course Mathematics 3501 (Middle Level Probability) at Brooklyn College of CUNY.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Outcomes and elementary events</td>
<td>15</td>
</tr>
<tr>
<td>4.4</td>
<td>The algebra of events</td>
<td>15</td>
</tr>
<tr>
<td>4.5</td>
<td>Probability</td>
<td>16</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Why the algebra of events</td>
<td>17</td>
</tr>
<tr>
<td>4.6</td>
<td>Nondisjoint events</td>
<td>17</td>
</tr>
<tr>
<td>4.7</td>
<td>Reading</td>
<td>18</td>
</tr>
<tr>
<td>4.8</td>
<td>Homework</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Discrete sample spaces</td>
<td>18</td>
</tr>
<tr>
<td>5.1</td>
<td>Picking marbles with replacement</td>
<td>18</td>
</tr>
<tr>
<td>5.2</td>
<td>Picking marbles without replacement</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>Conditional probability</td>
<td>20</td>
</tr>
<tr>
<td>6.1</td>
<td>Reading</td>
<td>21</td>
</tr>
<tr>
<td>6.2</td>
<td>Homework</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Independence</td>
<td>21</td>
</tr>
<tr>
<td>7.1</td>
<td>Independence involving complements</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>Some classical problems</td>
<td>24</td>
</tr>
<tr>
<td>8.1</td>
<td>Russian roulette</td>
<td>24</td>
</tr>
<tr>
<td>8.2</td>
<td>Russian roulette revisited</td>
<td>24</td>
</tr>
<tr>
<td>8.3</td>
<td>Gambler's ruin</td>
<td>25</td>
</tr>
<tr>
<td>8.4</td>
<td>The probability of a union</td>
<td>26</td>
</tr>
<tr>
<td>8.5</td>
<td>Rencontre</td>
<td>27</td>
</tr>
<tr>
<td>9</td>
<td>Random variables</td>
<td>28</td>
</tr>
<tr>
<td>9.1</td>
<td>Functions of random variables</td>
<td>29</td>
</tr>
<tr>
<td>9.2</td>
<td>Reading</td>
<td>29</td>
</tr>
<tr>
<td>9.3</td>
<td>Homework</td>
<td>29</td>
</tr>
<tr>
<td>10</td>
<td>The Stieltjes integral</td>
<td>29</td>
</tr>
<tr>
<td>10.1</td>
<td>More on Stieltjes integrals</td>
<td>30</td>
</tr>
<tr>
<td>10.2</td>
<td>Integration by parts</td>
<td>31</td>
</tr>
<tr>
<td>10.3</td>
<td>Change of variables</td>
<td>32</td>
</tr>
<tr>
<td>10.4</td>
<td>Improper Stieltjes integrals</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>Expectation</td>
<td>33</td>
</tr>
<tr>
<td>11.1</td>
<td>Expectation via Stieltjes integrals</td>
<td>34</td>
</tr>
<tr>
<td>11.2</td>
<td>Reading</td>
<td>35</td>
</tr>
<tr>
<td>11.3</td>
<td>Homework</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>Discrete and continuous random variables</td>
<td>36</td>
</tr>
<tr>
<td>12.1</td>
<td>Discrete random variables</td>
<td>36</td>
</tr>
<tr>
<td>12.2</td>
<td>Continuous random variables</td>
<td>36</td>
</tr>
<tr>
<td>13</td>
<td>Properties of expectation</td>
<td>37</td>
</tr>
<tr>
<td>13.1</td>
<td>Linearity of expectation</td>
<td>37</td>
</tr>
<tr>
<td>13.1.1</td>
<td>Expectation of a constant</td>
<td>38</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>20.2.1</td>
<td>The distribution function of the general normal distribution</td>
<td>58</td>
</tr>
<tr>
<td>20.2.2</td>
<td>The density function of the general normal distribution</td>
<td>58</td>
</tr>
<tr>
<td>21</td>
<td>The gamma distribution</td>
<td>58</td>
</tr>
<tr>
<td>21.1</td>
<td>The gamma function</td>
<td>58</td>
</tr>
<tr>
<td>21.2</td>
<td>The gamma distribution</td>
<td>59</td>
</tr>
<tr>
<td>21.3</td>
<td>Sums of independent gamma distributions of the same scale</td>
<td>60</td>
</tr>
<tr>
<td>22</td>
<td>The Jacobian: change of variables in multiple integrals</td>
<td>60</td>
</tr>
<tr>
<td>22.1</td>
<td>Change of variables: revisiting the single-variable case</td>
<td>60</td>
</tr>
<tr>
<td>22.2</td>
<td>Change of variables in double integrals</td>
<td>60</td>
</tr>
<tr>
<td>23</td>
<td>The beta distribution</td>
<td>62</td>
</tr>
<tr>
<td>24</td>
<td>The chi squared and the lognormal distributions</td>
<td>62</td>
</tr>
<tr>
<td>24.1</td>
<td>The square of the standard normal distribution</td>
<td>63</td>
</tr>
<tr>
<td>24.2</td>
<td>The lognormal distribution</td>
<td>64</td>
</tr>
<tr>
<td>24.3</td>
<td>Reading</td>
<td>64</td>
</tr>
<tr>
<td>24.4</td>
<td>Homework</td>
<td>64</td>
</tr>
<tr>
<td>25</td>
<td>Joint distribution functions</td>
<td>65</td>
</tr>
<tr>
<td>25.1</td>
<td>Marginal distributions</td>
<td>65</td>
</tr>
<tr>
<td>25.2</td>
<td>Probability of a rectangle</td>
<td>66</td>
</tr>
<tr>
<td>25.3</td>
<td>Joint density functions</td>
<td>66</td>
</tr>
<tr>
<td>25.4</td>
<td>Sums of two random variables</td>
<td>67</td>
</tr>
<tr>
<td>26</td>
<td>Independent random variables. Sums of continuous variables</td>
<td>68</td>
</tr>
<tr>
<td>26.1</td>
<td>Sum of two independent continuous random variables: convolution</td>
<td>68</td>
</tr>
<tr>
<td>26.2</td>
<td>Sums of independent gamma variables</td>
<td>69</td>
</tr>
<tr>
<td>26.3</td>
<td>Sums of independent normal variables</td>
<td>69</td>
</tr>
<tr>
<td>27</td>
<td>Sums of independent discrete random variables</td>
<td>70</td>
</tr>
<tr>
<td>27.1</td>
<td>Sums of independent Poisson variables</td>
<td>71</td>
</tr>
<tr>
<td>27.2</td>
<td>Sums of independent binomial variables</td>
<td>71</td>
</tr>
<tr>
<td>27.3</td>
<td>Reading</td>
<td>72</td>
</tr>
<tr>
<td>27.4</td>
<td>Homework</td>
<td>73</td>
</tr>
<tr>
<td>28</td>
<td>Multivariate Stieltjes integral</td>
<td>73</td>
</tr>
<tr>
<td>28.1</td>
<td>The double Stieltjes integral</td>
<td>73</td>
</tr>
<tr>
<td>28.2</td>
<td>Iterated integrals: the difficulties</td>
<td>74</td>
</tr>
<tr>
<td>28.3</td>
<td>The Burkill integral</td>
<td>74</td>
</tr>
<tr>
<td>28.4</td>
<td>Improper integrals</td>
<td>75</td>
</tr>
<tr>
<td>29</td>
<td>Expectation of functions of two random variables</td>
<td>75</td>
</tr>
<tr>
<td>29.1</td>
<td>Expectation when (q(X, Y)) only depends on (X)</td>
<td>76</td>
</tr>
<tr>
<td>29.2</td>
<td>Expectation of a sum</td>
<td>77</td>
</tr>
<tr>
<td>29.3</td>
<td>Expectation of product of independent variables</td>
<td>77</td>
</tr>
<tr>
<td>30</td>
<td>Covariance and correlation</td>
<td>78</td>
</tr>
<tr>
<td>30.1</td>
<td>Random variable with zero variance</td>
<td>78</td>
</tr>
</tbody>
</table>
1 Introduction

This is a set of notes for a middle level probability course at Brooklyn College.

2 Sets

A set is a collection. The members of this collection are called its elements; the symbol \( x \in A \) indicates that \( x \) is an element of the set \( A \). We can describe a set \( A \) by listing its elements inside braces \( \{ \text{ and } \} \); for example,

\[
A = \{1, 3, 5, 7, 9\}
\]

is the set whose elements are the integers 1, 2, 3, 5, 7, and 9. Certain sets commonly occurring in mathematics have standard notation; for example \( \mathbb{Z} \) is the set of all integers (positive, negative, or zero), \( \mathbb{Q} \) is the set of all rationals, and \( \mathbb{R} \) is the set of all real numbers. Sets can also be described by the set-builder operation:

\[
S = \{ x : \phi(x) \}
\]

denotes the set of all things \( x \) for which the condition \( \phi(x) \) is satisfied. Here the variable \( x \) usually has a certain range, i.e., it can assume certain values specified in the context (for example, one might agree that \( x \) is a real number, i.e., that \( x \) runs over real numbers). Sometimes one can indicate the range in the set-builder operation. For example, the set \( A \) in equation (2.1) can also be described as

\[
A = \{ x \in \mathbb{Z} : 1 \leq x < 10 \& x \text{ is odd} \};
\]

here \( \& \) is the logical “and.” That is, for two statements \( \Phi \) and \( \Psi \), the statement \( \Phi \& \Psi \) is true only in case both \( \Phi \) and \( \Psi \) are true. The fact that the sets described in formulas (2.1) and (2.2) are the same is a special case of the following

Axiom 2.1 (Axiom of Extensionality). Two sets are equal if and only if they have the same elements.
2.1 The empty set

Once one uses the set-builder operation, it is almost inevitable that one encounters a set with no elements; such a set is called the empty set, denoted as $\emptyset$. By the Axiom of Extensionality (Axiom 2.1) the empty set is unique, that is, there is only one empty set. With the set-builder operation, one might occasionally write $\emptyset = \{x : x \neq x\}$. With the listing notation, sometimes one writes $\emptyset = \{}$, noting that nothing is listed between the braces.

2.2 Relations between sets

Definition 2.1. Given two sets $A$ and $B$, we say that $A$ is a subset of $B$ if every element of $A$ is an element also of $B$. In this case, we also say that $B$ is a superset of $A$ also this term is used less often than the term subset. The symbol $A \subset B$ expresses the statement that $A$ is a subset of $B$. We also say that $B$ includes $A$. The symbol $B \supset A$ can also be used. The use of the word "contain" should be used with extreme care, since it is often misused. $B$ contains $A$ should properly mean that $A$ is an element of $B$ (yes, a set can be an element of another set), but it is often misused to mean that $A$ is a subset of $B$. Such a misuse should absolutely avoided, The best way to say that $x \in A$ is that $x$ is an element of $A$, or that $x$ belongs to $A$.

To illustrate the difference between $\in$ and $\subset$, note that $\emptyset \notin \emptyset$, since the empty set has no element, while $\emptyset \subset \emptyset$ is vacuously true. In fact, for every set $A$, we have $\emptyset \subset A$ is satisfied: given that $\emptyset$ has no element, the no requirements are imposed on $A$ by saying that every element of $\emptyset$ is also an element of $A$.

The set $\{\emptyset\}$ is the set whose only element is $\emptyset$; it differs from $\emptyset$, since the former has one element, the latter has none. The sets $\{\emptyset\}$ and $\{\{\emptyset\}\}$ both have one elements, but they are not the same sets according to Axiom 2.1 since their elements are not the same, as we saw just before.

2.3 Set operations

Given sets $A$ and $B$, their union $A \cup B$ is the set that contains (correct use!) the elements of either $A$ or $B$, that is,

$$A \cup B \overset{\text{def}}{=} \{x : x \in A \lor x \in B\},$$

where $\lor$ is the symbol for logical “or.” That is, for two statements $\Phi$ and $\Psi$, the statement $\Phi \lor \Psi$ is true if $\Phi$ or $\Psi$ is true.

The intersection $A \cap B$ of the sets $A$ and $B$ is the set that contains only the elements that belong to both $A$ and $B$. That is,

$$A \cap B \overset{\text{def}}{=} \{x : x \in A \land x \in B\}.$$

The set difference of $A$ minus $B$, denoted as $A \setminus B$, is defined as the set of those elements of $A$ that are not elements of $B$:

$$A \setminus B \overset{\text{def}}{=} \{x : x \in A \land x \notin B\}.$$

---

2.1 One might say, “contains the elements of both,” but such use is ambiguous; this is why we clarify what we mean next.

2.2 In mathematics, logical “or” is always meant in the inclusive sense; that is $\Phi \lor \Psi$ is true if one of $\Phi$ and $\Psi$ is true, or if both are true.

2.3 Sometimes the notation $A - B$ is used, but it should be avoided, since the latter notation is also used with different meanings. In earlier times, typesetting $A \setminus B$ caused extra difficulty for printers, but with computerized typesetting that is no longer an issue.
One might read the right-hand side here as the set of those elements \( x \) for which \( x \in A \) but \( x \notin B \). While the word “but” expresses contrast, its meaning in this context is not any different from that of the word “and.”

The symmetric difference of the sets \( A \) and \( B \) is defined as the set
\[
A \triangle B \overset{\text{def}}{=} (A \setminus B) \cup (B \setminus A).
\]

If \( A \) is a set of sets, i.e., a set all whose elements are also sets then \( \bigcup A \) is the union of all elements of \( x \). Formally,
\[
\bigcup A \overset{\text{def}}{=} \{ x : \text{we have } x \in B \text{ for some } B \in A \},
\]
or, even more formally,
\[
\bigcup A \overset{\text{def}}{=} \{ x : (\exists y)(y \in A \& x \in y) \},
\]
where \((\exists y)\) is an existential quantifier, to be read as “there is a \( y \) such that . . .” Using restricted quantifiers, this can also be written as
\[
\bigcup A \overset{\text{def}}{=} \{ x : (\exists y \in A)(x \in y) \}.
\]

Often, mathematicians not trained in logic, and, for an irrational reason, prefer the notation where the elements of \( A \) are indexed. That is, let
\[
A = \{ B_\iota : \iota \in I \};
\]
that is, \( I \) is a set indexing the elements of \( A \), and the Greek letter \( \iota \) (iota) is used to indicate that \( I \) may not be a set of integers. Assuming that \( B_\iota \) is a set for all \( \iota \in I \), they prefer to use the symbol
\[
\bigcup_{\iota \in I} B_\iota,
\]
even though the simpler symbol \( \bigcup A \) means the same thing.

Similarly, if \( A \) is a set of sets, then \( \bigcap A \) can be defined as the intersection of all elements of \( A \):
\[
\bigcap A \overset{\text{def}}{=} \{ x : \text{we have } x \in B \text{ for all } B \in A \},
\]
or, even more formally,
\[
\bigcap A \overset{\text{def}}{=} \{ x : (\forall y)(y \in A \to x \in y) \};
\]
here \((\forall y)\) is a universal quantifier, to be read as “for all \( y \) we have . . .,” and for two statements \( \Phi \) and \( \Psi \), \( \Phi \to \Psi \) is the conditional, meaning “if \( \Phi \) is true then \( \Psi \) is also true”. The only time \( \Phi \to \Psi \) is false is in case \( \Phi \) is true and \( \Psi \) is false. Using restricted quantifiers, this can also be written as
\[
(2.3) \quad \bigcap A \overset{\text{def}}{=} \{ x : (\forall y \in A)(x \in y) \}.
\]

One needs to be a little careful with using the symbol \( \bigcap A \), since it is meaningless in case \( A \) is the empty set, since if \( x \) in equation (2.3) runs over all sets (as one would naturally expect), then the right-hand side describes the set of all sets, which is meaningless. Even if \( A \) is the empty set and \( C \) is another set, it is reasonable to interpret \( C \cap (\bigcap A) \) to be the set \( C \).

---

2 That is, it is meaningless in most versions of axiomatic theory. It make sense in Quine’s New Fundation and axiomatization of set theory of interest to mathematical logicians and to philosophers, but is not commonly used in mathematical practice. See [17].
2.4 Venn diagrams

Set operations can be illustrated in Venn diagrams. One draws two or three circles in a box, indicating two or three sets, and then shades the results of various set operations. In Figure 2.1, a) illustrates the set \( A \cap B \cap C \), which is defined as the set \( A \cap (B \cap C) = (A \cap B) \cap C \); the equality here can easily be proved, and justifies the omitting of the parentheses in the first expression. Similarly, a) illustrates the set \( A \cup B \cup C \), which is defined as the set \( A \cup (B \cup C) = (A \cup B) \cup C \); the equality here can easily be proved, and justifies the omitting of the parentheses in the first expression. Finally, a) illustrates the set \( (A \triangle B) \setminus C \). Venn diagrams may be helpful in illustrations, but they should not be used for proofs.

\[
\begin{align*}
\text{a)} & \quad A \cap B \cap C \\
\text{b)} & \quad A \cup B \cup C \\
\text{c)} & \quad (A \triangle B) \setminus C
\end{align*}
\]

Figure 2.1: Venn diagrams

2.5 The power set of a set

The set of all subsets of a set \( A \) is called its power set, and it is usually denoted by \( \mathcal{P}(A) \). That is,

\[
\mathcal{P}(A) \overset{\text{def}}{=} \{ x : x \subset A \}.
\]

3 Combinatorics

3.1 Permutations: lists of length \( k \) of \( n \) items

Let \( n \) and \( k \) be integers with \( 0 \leq k \leq n \) A \( k \)-permutation of \( n \) distinct items a list containing \( k \) of given \( n \) items, where we distinguish between two lists containing the same items given in different order. We can count these permutations as follows. The first item of the list can be picked in \( n \) different ways, depending on which of the items we pick. After this, \( n - 1 \) items remain, so the second item can be picked in \( n - 1 \) different ways. After with, the number of possible lists is \( n(n - 1) \). When picking the third item, we have to pick from among \( n - 2 \) items, so the number of possible lists containing three items is \( n(n - 1)(n - 2) \). Continuing this, the number of lists containing \( k \) items, denoted by \( P^n_k \), is

\[
P^n_k = n(n-1)(n-2)\ldots(n-k+1) = \prod_{j=0}^{k-1} (n-j),
\]

For \( k = 0 \) the product on the right is the empty product, interpreted as 1, saying that there is only one list of length 0, the empty list.
3.1.1 Permutations

For an integer \( n \geq 0 \), an \( n \)-permutation of \( n \) items is simply called a permutation of these items. Their number according to equation (3.1) is written as \( n! \), read as \( n \)-factorial:

\[
n! \overset{def}{=} \prod_{j=0}^{n-1} (n-j) = \prod_{j=1}^{n} j = 1 \cdot 2 \cdots n.
\]

Again, if \( n = 0 \), the product is the empty product, interpreted as 1. That is, \( 0! = 1 \).

3.1.2 Permutations as mappings

Instead of considering permutations as a rearrangement of a given list, it is often more amenable to a mathematical treatment to consider a permutation of a set \( A \) as a one-to-one mapping of the set \( A \) onto itself. These permutations clearly have the same number as permutations considered as rearrangements. Indeed, given an integer \( n > 0 \), if \( A = \{ k : 1 \leq k \leq n \} \) and \( \sigma : A \rightarrow A \), then \( \sigma(k) = l \) means that in the corresponding rearrangement the number \( k \) is placed at the \( l \)th place on the list.

3.2 Combinations

Given integers \( k \) and \( n \) with \( 0 \leq k \leq n \), a \( k \)-combination of \( n \) distinct items is a selection of \( k \) of these items, where the order the items are selected does not count. Their number is the same as the number of \( k \)-element subsets of an \( n \)-element sets (since, in a set, the order in which the elements are listed makes no difference). Given a \( k \)-combination, if you permute the selected \( k \) items, you get \( k! \) lists, which are \( k \)-permutations of the \( n \) given items. If you take all \( k \)-combinations and permute each of them, you get all \( k \)-permutations of the \( n \) given items. That is, the number of these permutations, \( P_n^k \), is \( k! \) times the number of \( k \)-combinations of \( n \)-items. Denoting by \( \binom{n}{k} \) the number of these combinations, read as \( n \) choose \( k \), according to (3.1), we have

\[
\binom{n}{k} = \frac{\prod_{j=0}^{k-1} (n-j)}{k!} = \frac{\prod_{j=0}^{k-1} (n-j)}{\prod_{j=0}^{k-1} (k-j)} = \frac{\prod_{j=0}^{n} n-j}{\prod_{j=0}^{k} j}.
\]

Note that \( \binom{n}{0} = 1 \), since in this case we pick the empty combination (or the empty subset), so there is only one pick. The formula correctly gives 1 as the empty product.

The quantity \( \binom{n}{k} \) is called a binomial coefficient, in view of its role in the Binomial Theorem given in equation (3.6) below. Formula (3.2) with \( n \) replacing \( \alpha \) can be used to define the binomial coefficient \( \binom{\alpha}{k} \) for any real (or complex) \( \alpha \) and any integer \( k \geq 0 \). Such binomial coefficients play a role in the binomial series, given below in equation (18.5), a generalization of the Binomial Theorem.

Noting that \( \binom{n}{k} \) is defined only for \( 0 \leq k \leq n \) we have

\[
(n-k)! \prod_{j=0}^{k-1} (n-j) = (n-k)! \prod_{j=n-k+1}^{n} j = (\prod_{j=1}^{n-k} j) \prod_{j=n-k+1}^{n} j = \prod_{j=1}^{n} j = n!,
\]

3.1 If \( A \) is a finite set and \( \sigma : A \rightarrow A \) is one-to-one, then \( \sigma \) is also clearly onto \( A \). However, when one considers permutations as mappings, one occasionally thinks of permutations of an infinite set.

3.2 An older, now deprecated, notation is \( C_n^k \), not to be used in this course.
where the parentheses surrounding the first product after the second equation sign is to indicate that the second product is not in the scope of the first product.\(^3\)

Hence, by (3.1) we have

\[
(3.3) \quad \binom{n}{k} = \frac{\prod_{j=0}^{k-1} (n-j)}{k!} = \frac{(n-k)! \prod_{j=0}^{k-1} (n-j)}{(n-k)! k!} = \frac{n!}{(n-k)! k!}.
\]

While the right-hand side is not useful in calculations, it helps us to show that

\[
(3.4) \quad \binom{n}{n-k} = \frac{n!}{k! (n-k)!} = \frac{n!}{(n-k)! k!} = \binom{n}{k}.
\]

### 3.2.1 The binomial theorem

Let \(n\) be a positive integer, and for \(k\) with \(0 \leq k \leq n\), let \(A_k\) denote all \(k\)-element subsets of the set \(\{i : 1 \leq i \leq n\}\). In particular, we have \(A_0 = \emptyset\). The number of elements of \(A_k\) is \(\binom{n}{k}\). Assume be are given real numbers \(x, y_i\) for \(1 \leq i \leq n\).\(^4\)

We have

\[
(3.5) \quad \prod_{i=1}^{n} (x + y_i) = \sum_{k=0}^{n} \sum_{S \subseteq A_k} x^{n-k} \prod_{i \in S} y_i.
\]

The reason this equality holds is that we can evaluate the product on the left by taking all the products that result by picking either \(x\) or \(y_i\) from each of the factors, and then adding all products resulting this way. When we pick a \(y_i\) from \(k\) of these factors \((0 \leq k \leq n)\), we have to pick \(n-k\) times, and we obtain the term of the sum corresponding to this \(k\) on the right-hand side.\(^5\)

If we assume that \(y_1 = y_2 = \ldots = y_n = y\), then the left-hand side becomes \((x + y)^n\). On the right-hand side, for a fixed \(k\), each of the terms after the second sum becomes \(x^{n-k} y^k\). Since there are \(\binom{n}{k}\) of these terms for a given \(k\), the sum right-hand side becomes \(\sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k\). That is, we obtain

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

This is the binomial theorem. As we pointed out in the footnote above, the commutativity of the product was used in way in the proof. For example, the binomial theorem is not true for noncommuting square matrices \(A\) and \(B\) of the same size.

### 3.3 Partitions

A partition of a set \(A\) is a set \(P \subset \mathcal{P}(A)\) (cf. (2.4)) such that \(A = \bigcup P\) and the elements of \(P\) are pairwise disjoint. Assuming \(A\) is a finite set, a partition sequence of \(A\) is a sequence \((A_i : 1 \leq i \leq k)\) such that the \(A_i\) are pairwise disjoint sets such that \(A = \bigcup_{i=1}^{k} A_i\). The difference between a partition and a partition sequence is that in the latter we are interested in the order the subsets of \(A\) are listed; further, in a partition sequence the empty set may occur several times.\(^6\)

\(^3\)The rules are not quite clear where the scope of a product ends; usually, it is ended by a + or − sign unprotected by parentheses. It is not safe to assume that the second product sign would end the scope of the first one (we would tend to think that it does not end the scope).

\(^4\)The quantities \(x\) and \(y_i\) may be elements of an arbitrary commutative ring, instead of being real numbers.

\(^5\)Note that in obtaining this equation, the commutativity of multiplication was used in an important way, to ensure that, in each product, the occurrences of \(x\) can be moved to the front.

\(^6\)This is because the empty set is disjoint to every set, including the empty set itself.
3.4 Counting functions from a finite set into another

For an integer \( n > 0 \), write \( M_n = \{1, 2, \ldots, n\} \). Give integers \( n > 0 \), \( N > 0 \) and integers \( k_i \geq 0 \) for \( 1 \leq i \leq N \) such that \( \sum_{i=1}^{N} k_i = n \), we want to count the number of functions \( f : M_n \rightarrow M_N \) such that

\[
\begin{align*}
|\{j : 1 \leq j \leq n \text{ and } f(j) = i\}| &= k_i \quad \text{for each } i \text{ with } 1 \leq i \leq N.
\end{align*}
\]

(3.7)

There is a one-to-one correspondence between such functions and partition sequences \( \langle A_i : 1 \leq j \leq j \rangle \) of \( M_n \) such that \(|A_i| = k_i\); in this correspondence \( A_i = \{j : f(j) = i\} \). The number of these partition sequences is denoted as

\[
\binom{n}{k_1, k_2, \ldots, k_N}
\]

(3.8)

In order to count the number of functions satisfying (3.7) let \( f_0 \) be on such (arbitrarily selected) function. The all such functions \( f \) can be obtained as a composition \( f = f_0 \circ g \), where \( g : M_n \rightarrow M_n \) is an appropriate permutation function of \( M_n \) (i.e., \( g \) is one-to-one and onto). The number of such permutations is \( n! \). In order to count the number of functions \( f \), we need to count the number of permutations \( g \) that \( f_0 \circ g = f_0 \circ g_1 \) for a given \( g_1 \). This is easy to do. Writing \( f_1 = f_0 \circ g_1 \), let \( \langle B_i : 1 \leq j \leq j \rangle \) be the partition sequence corresponding to \( f_0 \), that is, \( B_i = \{j : f_0(j) = i\} \). We have \( f_0 \circ g = f_0 \circ g_1 \) is and only if \( f_0 = (f_0 \circ g_1) \circ g^{-1} \), that is

\[
f_0(j) = f_0 \circ (g_1 \circ g^{-1})(j).
\]

for all \( j \) with \( 1 \leq j \leq n \). This means that if \( j \in B_i \) then we must have \( g_1 \circ g^{-1}(j) \in B_i \); that is, \( (g_1 \circ g) \upharpoonright B_i \) must be a permutation of \( B_i \). If we specify this permutation for each \( i \) with \( 1 \leq i \leq N \), the permutation \( g \) is uniquely determined. Since \( B_i \) has \( k_i! \) permutations, and these can be arbitrarily chosen for each \( i \), this shows that there are \( \prod_{i=1}^{N} k_i! \) choices for \( g \) such that \( f_0 \circ g \) gives write so the same partitioning function satisfying formula (3.7). Thus, the total number of functions satisfying this formula is the number \( n! \) of all permutations of \( M_n \) divided by this number, that is

\[
\binom{n}{k_1, k_2, \ldots, k_N} = \frac{n!}{\prod_{i=1}^{N} k_i!}.
\]

(3.9)

3.5 The generalized distributive law

Let \((a_{ij})\) be an \( n \times m \) matrix. Then

\[
\prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij} = \sum_{\sigma} \prod_{i=1}^{n} a_{i\sigma(i)},
\]

where \( \sigma \) runs over all mappings of the set \( M_n = \{1, 2, \ldots, n\} \) into the set \( M_n = \{1, 2, \ldots, m\} \). The left-hand side represent a product of sums. The right-hand side multiplies out this product by taking one term out of each of these sums, and adding up all the products that can be so formed. The equality of these two sides is obtained by the distributivity of multiplication over addition. For example, for \( m = n = 2 \), the above equation says that

\[
(a_{11} + a_{12})(a_{21} + a_{22}) = a_{1\sigma_1(1)}a_{2\sigma_1(2)} + a_{1\sigma_2(1)}a_{2\sigma_2(2)} + a_{1\sigma_3(1)}a_{2\sigma_3(2)} + a_{1\sigma_4(1)}a_{2\sigma_4(2)},
\]

where \( \sigma_1(1) = 1, \sigma_1(2) = 1, \sigma_2(1) = 1, \sigma_2(2) = 2, \sigma_3(1) = 2, \sigma_3(2) = 1, \sigma_4(1) = 2, \sigma_4(2) = 2 \).

\[3.7\text{Note that } (f_0 \circ g_1) \circ g^{-1} = f_0 \circ (g_1 \circ g^{-1}).\]
3.6 The multinomial theorem

Let \((a_{ij})\) be an \(n \times m\) matrix with \(m\) identical rows, and write \(x_j = a_{ij}\) for each \(i\) and \(j\) with \(1 \leq i \leq n\) and \(1 \leq j \leq m\). Using the distributive law in (3.10), we have

\[
\left( \sum_{j=1}^{m} x_j \right)^n = \prod_{i=1}^{n} \sum_{j=1}^{m} a_{ij} = \sum_{\sigma} \prod_{i=1}^{n} a_{i\sigma(i)} = \sum_{\sigma} \prod_{i=1}^{n} x_{\sigma(i)}
\]

Each term on the right hand side is a product \(\prod_{j=1}^{m} x_j^{k_j}\), where \(\sum_{j=1}^{m} k_j = n\) and \(k_j \geq 0\); such a term is corresponds to a mapping \(\sigma : M_n \to M_m\) for which

\[
|\{i : 1 \leq i \leq n\} \text{ and } \sigma(i) = j| = k_j \quad \text{for each } j \text{ with } 1 \leq j \leq m.
\]

The number of such \(\sigma\) is

\[
\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{\prod_{j=1}^{m} k_j!}.
\]

according to equation (3.9). This gives the equation

\[
\left( \sum_{j=1}^{m} x_j \right)^n = \sum_{k_j : 1 \leq j \leq m, k_j \geq 0, \sum_{j=1}^{m} k_j = n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{j=1}^{m} x_j^{k_j}
\]

(3.11)

\[
= \sum_{k_j : 1 \leq j \leq m, k_j \geq 0, \sum_{j=1}^{m} k_j = n} \frac{n!}{\prod_{j=1}^{m} k_j!} \prod_{j=1}^{m} x_j^{k_j}.
\]

3.7 Lists with repetition, order counts

Let \(n, k \geq 0\) be integers, and assume we are given \(n\) items. We want to count the number of the lists of length \(k\) of these items when items are allowed to repeat. The order in which the items are listed counts. This is a simple task, since there are \(n\) ways to pick the first item on the list, then \(n\) ways to pick the second, third, etc. items. After \(k\) picks, we can create \(n^k\) lists. This is the total number of lists that can be created.

3.8 Combinations with repetitions

Let \(n, k \geq 0\) be integers, and assume we are given \(n\) items. We want to count the number of the lists of length \(k\) of these items when items are allowed to repeat, but the order in which the items are listed does not count. Such lists are called \(k\)-combinations of \(n\) items with repetitions.

The number of these items can be counted as follows. Assume the items to be listed are the numbers \(1, 2, 3, \ldots, n\). Define a mapping \(f\) from the set of such lists as follows. Let \(\lambda = (i_1, i_2, i_3, \ldots, i_k)\) be such a list, arranged in nondecreasing order; that is, assume \(i_1 \leq i_2 \leq i_3 \leq \ldots \leq i_k\). Then put

\[
f(\lambda) = \{i_l + l - 1 : 1 \leq l \leq k\} = \{i_1 + 0, i_2 + 1, i_3 + 2, \ldots, i_k + (k - 1)\}.
\]

Noting that the numbers listed in the set on the right-hand side are all distinct; in fact, \(i_1 + 0 < i_2 + 1 < i_3 + 2 < \ldots < i_k + (k - 1)\). It is easy to see that \(f\) is a one-to-one mapping from the set of all such lists onto the set of all \(k\)-element subsets of the set

\[
\{1, 2, 3, \ldots, n + k - 1\} = \{m \in \mathbb{Z} : 1 \leq m \leq n + k - 1\}.
\]

Since the number of these subsets is \(\binom{n+k-1}{k}\) according to Subsection 3.2 this is also the number of lists with repetitions we are considering.
3.9 The number of solutions of an equation

Here we solve the following question. Let \( n \) and \( r \) be positive integers with \( 1 \leq r \leq n \). How many positive integer solutions \( x_1, x_2, \ldots, x_n \), are there of the equation

\[
\sum_{i=1}^{r} x_i = n.
\]

The number can be given as follows. Consider the list 111...1 of length \( n \), and place \( r - 1 \) commas in this list in such a way at most one comma can be places between two adjacent 1s. Let \( x_i \) be the number of 1s between the \( i \)-th and \( i+1 \)-th comma. This gives all possible solutions of equation (3.12). Since \( r - 1 \) places need to be selected to place the commas at from among \( n - 1 \) available spaces between two adjacent numbers, the number of ways the commas can be placed is \( \binom{n-1}{r-1} \). This is the number of positive solutions of equation (3.12).

To count the number of nonnegative solutions of equation (3.12) note that this equation is equivalent to

\[
\sum_{i=1}^{r} (x_i + 1) = n + r.
\]

The number of nonnegative solutions of this equation is the same as the number of positive solutions of the equation

\[
\sum_{i=1}^{r} y_i = n + r.
\]

This number is \( \binom{n+r-1}{r-1} \), giving the number of nonnegative solutions of equation (3.12).

3.10 Reading


3.11 Homework

[19] Chapter 1, pp. 15–17, Problems 1, 8, 13, 20, 33.


Also look at all of [19] Chapter 1, pp. 20–21, Self-Test Problems and Exercises. Note that solutions to there are given at in the book [19], starting on p. 470.

4 Probability: the Kolmogorov probability model

We will describe the mathematical model of probability due to [Andrey Kolmogorov].

4.1 Outcomes and the sample space

A result of a random experiment is an outcome. For example, when rolling a die, there are six possible outcomes, corresponding to the number rolled. The set of possible outcomes is called the sample space. The set of all outcomes is called the sample space, usually denoted by \( \Omega \) is our discussions.\(^4\)

\(^4\)Perhaps this is too simplistic. In certain cases, modeling a random event may not be quite easy, and often it is not clear what the sample space should be.
4.2 Events and probabilities

Events are certain sets of outcomes. That is the set of events $\mathcal{B}$ is a certain subset of the power set $\mathcal{P}(\Omega)$ of $\Omega$. Events will be assigned probabilities: a probability function will be a function $P : \mathcal{B} \to [0,1]$ mapping $\mathcal{B}$ into the interval $[0,1]$ of real numbers. The set of events and the probability function $P$ will have to satisfy certain properties, to be discussed below.

4.3 Finite uniform probability spaces

One is often faced with a situation that in an experiment there are only finitely many possible outcomes, and each outcome is equally likely. For example, when rolling a fair die, the possible outcomes are 1, 2, 3, 4, 5, 6. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. In this case, the set of events $\mathcal{B}$ is the set of all subsets $\mathcal{P}(\Omega)$ of $\Omega$. For example, the event $\{1, 2, 5\}$ means that the outcome of a roll is 1, 2, or 5. For a finite set $A$ denoting by $|A|$ its number of element, the probability of an event $E \subset \Omega$ will be taken to be

$$P(E) = \frac{|E|}{|\Omega|}.$$  

4.3.1 Outcomes and elementary events

As we mentioned, outcomes are elements of the sample space $\Omega$. That is, if $x \in \Omega$, then $x$ is an outcome. An elementary event is the occurrence of an outcome; that is, the elementary event corresponding to $x$ is $\{x\}$; i.e., the one element set containing only $x$, called the singleton of $x$, or, more simply, singleton $x$. From a set theoretical point of view, $x$ and $\{x\}$ must be distinguished, so as to be in compliance with the axiom of extensionality (Axiom 2.1). In probability theory, one might be more lax, and sometimes overlook such distinctions (this is a fact of life, not something to be encouraged).

4.4 The algebra of events

We mentioned that the set of events $\mathcal{B}$ is a subset of the power set $\mathcal{P}(\Omega)$ of $\Omega$. We will require that $\mathcal{B}$ satisfy certain requirements:

**Definition 4.1 (Algebra of events).** An algebra of events is a set $\mathcal{B} \subset \mathcal{P}(\Omega)$ satisfying the following requirements:

1. We have $\Omega \in \mathcal{B}$.
2. If $A, B \in \mathcal{B}$, then $A \setminus B \in \mathcal{B}$.
3. Assume $A_n \in \mathcal{B}$ for every positive integer $n$. Then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{B}$.

When stating a definition, one tries to be as economical as possible. In this case, one may need to show that a certain set $\mathcal{B}$ satisfies this definition; the result of the economy is that one has less work to do when doing this. In fact, there is a lot more to this definition then is immediately clear. For example, it follows by Clauses [1] and [2] that $\emptyset = \Omega \setminus \Omega \in \mathcal{B}$. Further, if $A, B \in \mathcal{B}$, then $A \cup B \in \Omega$. This follows from Clause [3] with $A = A_1$, $B = A_2$, and $A_n = \emptyset$ for $n \geq 2$. We also have

$$A \cap B = \Omega \setminus ((\Omega \setminus A) \cup (\Omega \setminus B)) \in \mathcal{B}$$

\footnote{Note that we are discussing finite uniform probability spaces here. In a more general situation, if $x \in \Omega$, it is not a requirement that singleton $x$ be an event; that is, we may or may not have $\{x\} \in \mathcal{B}$.}
according to the above clauses.\footnote{4.3}{We have a similar result for infinite intersections: if \( A_n \in \mathcal{B} \) for every positive integer \( n \), then we also have
\[
\bigcap_{n=0}^{\infty} A_n = \Omega \setminus \bigcup_{n=0}^{\infty} (\Omega \setminus A_n) \in \Omega.
\]
A set satisfying the properties given in this definition is called a \( \sigma \)-algebra of sets. If Clause \( (3) \) is only required for a finite number of events, the set is called an algebra of sets. Since we will always require this clause for infinitely many sets, we will use the simpler term algebra, even though we always mean a \( \sigma \)-algebra.

4.5 Probability

Two sets \( A \) and \( B \) are called disjoint if \( A \cap B = \emptyset \), i.e., if \( A \) and \( B \) have no elements in common. A list of sets are called pairwise disjoint if any two among them are disjoint; i.e., if \( A_1, A_2, \ldots \) are sets (there may be a finite or an infinite number among them) if \( A_i \cap A_j = \emptyset \) unless \( i = j \); this also means that the same set should not be listed twice, unless it is the empty set (which can be listed any number of times).

\textbf{Definition 4.2} (Probability function). Given an algebra of events \( \mathcal{B} \), a probability function \( P : \mathcal{B} \to [0, 1] \) is a function satisfying the following requirements:

(1) We have \( P(\Omega) = 1 \).

(2) If \( A_n \in \mathcal{B} \) for every positive integer \( n \), and the sets \( A_n \) are pairwise disjoint, then
\[
P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).
\]

The property described in Clause \( (2) \) is called \( \sigma \)-additivity; a similar property stated for a finite number of events is called finite additivity, or simply additivity. Intuitively, it is not immediately clear why \( \sigma \)-additivity is needed as opposed to just additivity; this will be clear in the discussion of Russian roulette below. Subsection \( 8.1 \).

4.4 There is a lot more in this definition than immediately meets the eye. Clause \( (2) \) immediately implies that \( P(\emptyset) = 0 \). Indeed, choosing \( A_n = \emptyset \) for every \( n \), the equality in this clause would not hold if \( P(\emptyset) > 0 \). If \( A, B \in \mathcal{B} \) and \( A \) and \( B \) are disjoint, then \( P(A \cup B) = P(A) + P(B) \). This follows from Clause \( (2) \) with the choice \( A_1 = A, A_2 = B, \) and \( A_n = \emptyset \) for \( n \geq 2 \). Hence, if \( A \in \mathcal{B} \), then \( P(A) + P(\Omega \setminus A) = P(\Omega) = 1 \); hence, we have
\[
P(\Omega \setminus A) = 1 - P(A).
\]

\footnote{4.3}{The equality between the sets in the last display is one of the \textbf{De Morgan identities} (click on the link for an explanation and nice Venn diagram illustrations). The sister identity of this is
\[
A \cup B = \Omega \setminus \left((\Omega \setminus A) \cap (\Omega \setminus B)\right) \in \Omega.
\]
There are also analogous identities in logic.

\footnote{4.4}{On account of \( \sigma \)-additivity, one needs to remember that an infinite sum, or sum of a series, is taken to be the limit of its partial sums. While usual addition is commutative, the same is not necessarily true for infinite series – however, it is always true for a series with nonnegative terms, or, more generally, for an absolutely convergent series. However, a conditionally convergent series can always be rearranged to converge to something else or to diverge, according to a theorem of \textbf{Dirichlet}.}

16
The set \( \Omega \setminus A \) is called the complement\footnote{First, note that the words “complement” and “compliment” have very different meaning (you may look up the word in Wiktionary to see the difference. Second, if \( A \) and \( B \) are sets and \( A \subset B \), then \( B \setminus A \) is occasionally called the relative complement of \( A \) with respect to \( B \). If \( B \) is the universal set in a discourse, that is, the set of all things under consideration, then \( B \setminus A \) is simply called the complement of \( A \). In most mathematical discussions, there is no universal set, but in discussing probabilities and events, one might think of \( \Omega \) as the universal set.} of \( A \), and one often uses a special notation for it. Following the book \[3\], we will use the notation

\[ A^* \overset{\text{def}}{=} \Omega \setminus A. \]

With this notation, we have \( P(A^*) = 1 - P(A) \).

**Definition 4.3** (Probability space). A probability space is a triple \((\Omega, \mathcal{B}, P)\), where, as described above, \( \Omega \) is a set of outcomes, \( \mathcal{B} \) is an algebra of events on \( \Omega \), and \( P : \Omega \to [0,1] \) is a probability function.

By an abuse of language, one often says, let \( \Omega \) be a probability space when one in fact means the triple \((\Omega, \mathcal{B}, P)\).

### 4.5.1 Why the algebra of events

One may ask, why one does not take the set of events all subsets of \( \Omega \). In fact, this is what one usually does when \( \Omega \) is finite or a countably infinite set. When \( \Omega \) is an uncountable set, this is usually impossible. In an important paper \[20\] Stanislaw Ulam\footnote{Ulam made many other contributions, discussed in the Wikipedia article, to the Manhattan Project (the program to create the first atomic bomb), and he had the basic idea for the Teller–Ulam design, the design on which the hydrogen bomb is based.}. The question whether it is possible to have a probability function on all subsets of an uncountable set leads to deep questions of the foundation of mathematics. Ulam showed that the existence of such an uncountable set is unprovable with the usual axioms of set theory.

### 4.6 Nondisjoint events

We have

**Lemma 4.1.** Let \((\Omega, \mathcal{B}, P)\) be a probability space, and assume that \( A, B \in \mathcal{B} \). Then

\[(4.2) \quad P(A \cup B) = P(A) + P(B) - P(A \cap B).\]

**Proof.** The sets \( A \setminus (A \cap B) \), \( B \setminus (A \cap B) \), and \( A \cap B \) are pairwise disjoint, and we have

\[ A \cup B = (A \setminus (A \cap B)) \cup (B \setminus (A \cap B)) \cup (A \cap B). \]

Hence

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

That is,

\[(4.3) \quad P(A \cup B) = \left( P(A \setminus (A \cap B)) + P(A \cap B) \right) \]

\[ + \left( P(B \setminus (A \cap B)) + P(A \cap B) \right) - P(A \cap B) \]

Noting that we have

\[ P(A) = P\left(A \setminus (A \cap B)\right) + P(A \cap B) \]
and
\[ P(B) = P(B \setminus (A \cap B)) + P(A \cap B), \]
and substituting these into equation (4.3), equation (4.2) follows.

4.7 Reading
[19, § Chapter 2, pp. 22–57]. A part of this material is related to Section 5. Another part of the material is discussed in Subsection 8.4 and again, with the help of indicator variables, in Subsection 13.2.

4.8 Homework
[19, Chapter 2, pp. 50–54], Problems 9, 12, 17, 25, 31, 40, 53, 55. [19, § Chapter 2, pp. 55–56], Theoretical Exercises, 4, 5, 8b, 18. [19, Chapter 2, pp. 56–57], Self-Test Problems and Exercises. A part of this homework is related to Section 5.

5 Discrete sample spaces
A discrete probability space is a finite or countably infinite probability space in which each subset is an event; that is, it is a triple \((\Omega, B, P)\) where \(\Omega\) is finite or countably infinite, and \(B = P(\Omega)\). According to Clause [2] of Definition [4.2] in order to define the probability function \(P\), it is enough to define \(P(\{x\})\) for each \(x \in \Omega\). Here we will usually consider finite uniform probability spaces, where probabilities can be calculated by formula (4.1).

5.1 Picking marbles with replacement
In what follows, “marbles” mean colored glass balls, usually red or green, placed in a container, usually called an urn in probability theory. Given the urn with marbles, one randomly picks one of these marbles after thoroughly mixing the marbles – of course, the marbles are indistinguishable to the touch. One can pick with replacement, in which the marble picked will be put back in the urn, and without replacement, in which the marble is not put back. First we consider picking with replacement.

Given \(a\) red marbles and \(b\) green marbles in an urn; here \(a\) and \(b\) must be positive integers. We pick \(n\) marbles with replacement; what is the probability of picking exactly \(k\) red marbles; here \(0 \leq k \leq n\). In order to solve this problem, the probability space \(\Omega\) will be the set of all sequences of picks. There are \((a+b)\) items, and \(n\) picks with repetition allowed; the number of these picks are \((a+b)^n\) according to Subsection 3.7 That is, \(|\Omega| = (a+b)^n\).

As for counting the number of favorable outcomes, we want first to count those outcomes in which we first pick \(k\) red marbles, then we pick \(n-k\) green marbles. The number of ways we can pick \(k\) red marbles is the number of lists of length \(k\) that we can form with the \(a\) red marbles with repetitions allowed. This is \(a^k\). The number of ways we can then pick \(n-k\) green marbles is, similarly, \(b^{n-k}\). The number of ways first picking \(k\) red marbles and then \(n-k\) green marbles is \(b^{n-k}\). The probability of this happening is

\[
\frac{a^k b^{n-k}}{(a+b)^n} = \left(\frac{a}{a+b}\right)^k \left(\frac{b}{a+b}\right)^{n-k}.
\]

5.1.1 The words “discrete” and “discreet” have totally different meanings. Look them up in [Wiktionary] to see the difference.
In order to clarify the situation, we first want to consider the simple example when \( k = 3 \) and \( n = 5 \). Picking a red marble will be written as \( R \) and picking a green marble, \( G \). The sequence \( RRRGG \) indicates first picking 3 red marbles, and then 2 green marbles. If we drop the restriction that the red marbles will be picked first, then the possibilities that we pick 3 red marbles and 2 green ones can be described by the sequences \( RRRGG, RRGRG, RRGGR, RGRGG, RGRGR, RGGRR, GRRRG, GRRGR, GRGRR, GRRRR \). Each of these possibilities represent mutually exclusive events. For example, \( RRRGR \) and \( GRRGR \) cannot happen at the same time, since the first sequence indicates that the second pick is a red marble, while the second sequence indicates that the second pick is green. Finally each of these sequences correspond to the same count \( a^3b^2 \) (or, in the general case \( a^kb^{n-k} \) sequences of marbles.

Next, we want to count the number of sequences formed by the letters \( R \) and \( G \) listed above. This is easy, since each sequence corresponds to a combination of the letters \( \{1, 2, 3, 4, 5\} \), since one way to represent these sequences is to list the places where a red letter is located. For example, the sequence \( RGRRG \) corresponds to the set \( \{1, 3, 4\} \); that is, the number of these sequences is the number of 3-element subsets of a 5 element set, that is, \( \binom{5}{3} \). In the general case, this number is the number of \( k \)-element subsets of an \( n \)-element set; that is, \( \binom{n}{k} \).

We obtain the probability of picking \( k \) red marbles and \( n-k \) green marbles by adding up the probabilities of each particular sequence of picks, since these sequences represent mutually exclusive events. Since each of these probabilities is the same, as given in formula (5.1), and the number of these sequences is \( \binom{n}{k} \), our work is done since one way to represent these sequences is to list the places where a red letter is located. For example, the sequence \( RGRRG \) corresponds to the set \( \{1, 3, 4\} \); that is, the number of these sequences is \( \binom{5}{3} \). In the general case, this number is the number of \( k \)-element subsets of an \( n \)-element set; that is, \( \binom{n}{k} \).

Since each of these probabilities is \( \binom{n}{k} \), and the number of these sequences is \( \binom{n}{k} \), we obtain that this probability is

\[
\binom{n}{k} \frac{a^k b^{n-k}}{(a+b)^n} = \binom{n}{k} \left( \frac{a}{a+b} \right)^k \left( \frac{b}{a+b} \right)^{n-k}.
\]

The second form will be of particular interest to us, since it is a special case of the binomial distribution, to be discussed later.

### 5.2 Picking marbles without replacement

Given \( a \) red marbles and \( b \) green marbles in an urn, where \( a \) and \( b \) are positive integers, we want to pick \( n \) marbles without replacement, and the question is, what is the probability that \( k \) or these marbles are red. For the question to be meaningful, we must have \( 0 \leq k \leq n \leq a+b \). The first issue is the selection of the probability space \( \Omega \). It is much simpler to consider the set of chosen marbles, and ignore the order in which they were picked; doing this will give the same answer. The reason is that if we consider the sequence of the picks, the number of sets would be multiplied by \( n! \), the number the picked marbles can be put in different orders.

\[
\binom{n}{k} \binom{a+b}{n-k} = \binom{a+b}{n-k} \binom{a}{k} \binom{b}{n-k} \binom{a+b}{n}.
\]

The reason is that if we consider the sequence of the picks, the number of sets would be multiplied by \( n! \), the number the picked marbles can be put in different orders.

19
6 Conditional probability

Given a probability space \( \Omega \) and two events \( A \) and \( B \) with \( P(B) \neq 0 \), we define the conditional probability \( P(A \mid B) \), to be read as the probability of \( A \) given that \( B \) occurs is

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

As an example, consider the case of rolling a die, and assume we know that the number rolled is \( \leq 4 \), what is the probability that the number rolled is 1, 3, 4, or 5. Then, calculating this probability directly, we can take \( \Omega \) direct \( = \{1, 2, 3, 4\} \) so

\[
P_{\text{direct}}(E_{\text{direct}}) = \frac{|E_{\text{direct}}|}{|\Omega_{\text{direct}}|} = \frac{3}{4}.
\]

Calculating the same probability as a conditional probability, let the probability space be \( \Omega = \{1, 2, 3, 4, 5, 6\} \). Let \( A = \{1, 3, 4, 5\} \) and \( B = \{1,2,3,4\} \). Then \( A \cap B = \{1, 3, 4\} \) and so \( P(A \cap B) = 3/6 \), and \( B = \{1, 2, 3, 4\} \), and so \( P(B) = 4/6 \), the calculation is simpler if we do not reduce these fractions. Hence

\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{3/6}{4/6} = \frac{3}{4};
\]

we obtain the same result. We have

**Theorem 6.1** (Total probability theorem). Let \( \Omega \) be a probability space. Assume \( H_i \subset \Omega \) pairwise disjoint events, where \( 1 \leq i \leq n \) for some integer \( n \) or \( 1 \leq i < \infty \), such that \( \bigcup_i H_i = \Omega \), and let \( A \subset \Omega \) be another event. Assume, further, that \( P(H_i) \neq 0 \) for any \( i \). Then

\[
P(A) = \sum_i P(H_i) P(A \mid H_i).
\]

In case \( P(H_i) = 0 \), the conditional probability \( P(A \mid H_i) \) is not defined; this is why the assumption \( P(H_i) \neq 0 \) is needed. If we define the product \( P(H_i) P(A \mid H_i) \) to be 0 is case \( P(H_i) = 0 \), then this assumption can be dropped. The other assumptions can be somewhat weakened to say that \( P(A_i \cap A_i) = 0 \) instead of \( A_i \cap A_j = \emptyset \), and \( \sum_i P(H_i) = 1 \) instead of \( \bigcup_i H_i = \Omega \).

**Proof.** We have

\[
A = A \cap \Omega = A \cap \left( \bigcup_i H_i \right) = \bigcup_i (A \cap H_i).
\]

Hence,

\[
P(A) = \sum_i P(A \cap H_i) = \sum_i P(H_i) P(A \mid H_i);
\]

here, the first equation follows by Clause \([2]\) in Definition \([4.2]\) and the second equation holds by the definition of conditional probability given in equation \([6.1]\). \( \square \)

**Theorem 6.2** (Bayes’s theorem). Let \( \Omega \) be a probability space. Assume \( H_i \subset \Omega \) pairwise disjoint, where \( 1 \leq i \leq n \) for some integer \( n \) or \( 1 \leq i < \infty \), such that \( \bigcup_i H_i = \Omega \), and let \( A \subset \Omega \) be another

\[\text{6.1 Of course, it is ridiculous to include 5 if we know that the number rolled is } \leq 4, \text{ but the question is perfectly meaningful.}\]

\[\text{6.2 Of course, we do not put 5 into the set } E_{\text{direct}}, \text{ since 5 is not even an element of the probability space } \Omega_{\text{direct}}.\]
event. Assume, further, that \( P(H_i) \neq 0 \) for any \( i \). Then assuming \( P(A) \neq 0 \), for any \( k \) in the same range as \( i \), we have

\[
P(H_k \mid A) = \frac{P(H_k) P(A \mid H_k)}{\sum_i P(H_i) P(A \mid H_i)}.
\]

Proof. By using equation (6.1) twice, we have

\[
P(H_k \mid A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(H_k) P(A \mid H_k)}{P(A)}.
\]

Using equation (6.2) for \( P(A) \) in the denominator, formula (6.3) follows. \( \square \)

**Problem 6.1.**

a). In a factory, parts are manufactured by three machines, \( M_1, M_2, \) and \( M_3 \) in proportions 20 : 30 : 50. The percentages 5%, 2%, and 4% of these parts are defective, respectively. Find the probability that a randomly chosen part is defective.

b). Find the probability that a defective part was manufactured on the second machine.

**Solution.**

a). Write \( A \) for the event that a part is defective, and \( M_i \) with \( i = 1, 2, 3 \) for the event that it was manufactured on machine \( i \). We have

\[
P(A) = \sum_{i=1}^{3} P(A \mid M_i) P(M_i) = .05 \cdot .2 + .02 \cdot .3 + .04 \cdot .5 = .036.
\]

b). Using Bayes’s Theorem 6.2 we have

\[
P(M_2 \mid A) = \frac{P(A \mid M_2) P(M_2)}{\sum_{i=1}^{3} P(A \mid M_i) P(M_i)} = \frac{.02 \cdot .3}{.05 \cdot .2 + .02 \cdot .3 + .04 \cdot .5} = \frac{.006}{.036} = \frac{1}{6} \approx .166667;
\]

note that the denominator in the third member is the same as the answer to part a).

6.1 Reading

[19] Chapter 3, pp. 58–118. § 3.5 has many examples. Only Example 5c will be covered in the class, but all examples provide interesting (but not easy) reading. Some of this reading relates to Section 7.

6.2 Homework

[19] Chapter 3, pp. 103–112, Problems 1, 6, 10, 25, 31, 43, 52, 66, 81, 95. [19] Chapter 3, pp. 113–116, Theoretical exercises 1, 5, 6, 12, 15, 28 (counterexample: Roll two dice. \( E_1 \): 4 is rolled on the first die. \( E_2 \): 4 is rolled on the second die. \( F \): a total of 7 is rolled. Check out the details).

Also look at all of [19] Chapter 3, pp. 116–116, Self-Test Problems and Exercises. Some of this homework relates to Section 7.

7 Independence

In common sense, a number of events are independent if they have no influence on one another. The mathematical definition is somewhat more complicated. We want to give the definition for any \( n \) events. The definition will be recursive in that the the independence of \( n \) events will rely on the independence of \( n-1 \) events.
Lemma 7.1. Let $\Omega$ be a probability space, let $n \geq 1$ be an integer, and let $A_i$ for $1 \leq i \leq n$ be events. If $n = 1$, we will call the event independent. If $n > 1$, we will call them independent if any $n - 1$ of them are independent, and, in addition, we have

\[
P \left( \bigcap_{i=1}^{n} A_i \right) = \prod_{i=1}^{n} P(A_i). \tag{7.1}\]

For $n = 1$, independence means no restriction on the event $A_1$, that is, $A_1$ by itself is always independent; we allowed the case $n = 1$ to allow the recursive definition and to support some proofs by induction. For $n = 1$, this says that $A_1$ and $A_2$ are independent if $P(A_1 \cap A_2) = P(A_1)P(A_2)$. If $P(A_2) \neq 0$, then this is equivalent to the first equation in

\[
P(A_1) = \frac{P(A_1 \cap A_2)}{P(A_2)} = P(A_1 | A_2),\]

of (6.1). That is, in this case the the probabilty of $A_1$ conditional on $A_2$ occurring is the same as the unconditional probability of $A_1$. For $n = 3$, independence means that we have where the second equation holds in view $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$, $P(A_1 \cap A_2) = P(A_1)P(A_2)$, $P(A_1 \cap A_3) = P(A_1)P(A_3)$, $P(A_2 \cap A_3) = P(A_2)P(A_3)$. It is clear from the definition, that given $n$ events, wether or not they are independent does not depend on the order the events are listed,

7.1 Independence involving complements

Lemma 7.1. Let $\Omega$ be a probability space, let $n \geq 1$ be an integer. Assume the events $A_i$ for $1 \leq i \leq n$ are independent. Then the events $A_i^*$ and $A_i$ for $2 \leq i \leq n$ are independent.

Proof. For $n = 1$ this is true, since independence imposes no requirements in case $n = 1$. We use induction on $n$, so assume $n \geq 2$ and the assertion is true for any $n'$ events with $1 \leq n' < n$. Recall that $A_i^*$ denotes the complement of $A_i$. The only thing we need to show that equation (7.1) remains valid if $A_1$ is replaced by $A_1^*$. Writing $B = \bigcap_{i=2}^{n} A_i$, we have $(A_1 \cap B) \cup (A_1^* \cap B) = B$, and the sets $A_1 \cap B$ and $A_1^* \cap B$ on the left-hand side are disjoint; so $P(A_1 \cap B) + P(A_1^* \cap B) = P(B)$. Hence, using equation (7.1) with $A_1 \cap B = \bigcap_{i=1}^{n} A_i$ and the analogous equation with $B = \bigcap_{i=2}^{n} A_i$, we have

\[
P(A_1^* \cap B) = P(B) - P(A_1 \cap B) = \prod_{i=2}^{n} P(A_i) - \prod_{i=1}^{n} P(A_i) = \prod_{i=2}^{n} P(A_i) - P(A_1)\prod_{i=2}^{n} P(A_i) = P(A_1^*)\prod_{i=2}^{n} P(A_i).\]

This shows that equation (7.1) holds with $A_1^*$ replacing $A_1$, completing the proof. \qed

Corollary 7.1. Given a list of independent events, the list of events obtained by replacing any number of them by their complements, the list of events so obtained is independent.

Proof. When replacing the desired events on the list one by one, the lemma just proved shows that at each step we obtain a list of independent events. \qed

Problem 7.1. Given $n > 0$ hunters, they shoot at a deer at the same time; each has a probability $\alpha$ of hitting the deer; these events are independent. What is the probability that the deer will be hit at least once.
Solution. For $i$ with $1 \leq i \leq n$ let $A_i$ be the event that the $i$th hunter hits the deer. Then $\bigcap_{i=1}^{n} A_i^c$ is the event that none of them hits. Since the events $A_i^c$ are independent according to Corollary 7.1, we have
\[
P\left(\bigcap_{i=1}^{n} A_i^c\right) = \prod_{i=1}^{n} P(A_i^c) = (1 - \alpha)^n.
\]
The deer will be hit if this event does not happen; that is, the deer will be hit with probability $1 - (1 - \alpha)^n$.

Problem 7.2. Let $m$ and $n$ be positive integers. Assume an experiment $A$ of probability $p$ is repeated independently. What is the probability of the event $E$ that $n$ consecutive successes occur before $m$ consecutive failures?

Solution. Write $q = 1 - p$. Let $H$ be the event that the first experiment is a success. Then we have
\[
P(E|H) = p^{n-1} + (1 - p^{n-1}) P(E|H^*).
\]
This equation can be explained as follows: after the first success, if the next $n - 1$ trials are all successes, then $E$ occurs; the probability of this is $p^{n-1}$. If this does not occur, then disregarding everything before the first failure, the probability of $E$ occurring is $P(E|H^*)$ (because we are starting with a failure).

Similarly, we have
\[
P(E|H^*) = (1 - q^{m-1}) P(E|H).
\]
This equation can be explained as follows. For $E$ to occur after the first failure, the next $m - 1$ trials must not be failures; this occurs with probability $1 - q^{m-1}$. That is, one of the first $m$ trials must be a success. Disregarding the trials before the first success, for $E$ to occur, we must have $E$ after the initial success.

These two equations can easily be solved for $P(E|H)$ and $P(E|H^*)$. We have
\[
P(E|H) = \frac{p^{n-1}}{p^{n-1} + q^{m-1} - p^{n-1}q^{m-1}}
\]
and
\[
P(E|H^*) = \frac{p^{n-1}(1 - q^{m-1})}{p^{n-1} + q^{m-1} - p^{n-1}q^{m-1}}.
\]
Given that we have
\[
P(E) = p P(E|H) + q P(E|H^*),
\]
it follows that
\[
P(E) = \frac{p^n + q^{n-1}(1 - q^{m-1})}{p^n - p^{n-1}q^{m-1}} = \frac{p^{n-1}(1 - q^m)}{p^{n-1} + q^{m-1} - p^{n-1}q^{m-1}}.
\]
for the last equation, note that $q = 1 - p$.

1 See [19] p. 97–98. We tried to simplify the explanation there, which we find a little too formal. Some readers may, however, prefer the more formal explanation.

2 If one is uncomfortable with this explanation, one can proceed as follows: If a failure occurs in the $n - 1$ trials following the first success, divide this event into $n - 1$ mutually exclusive events according to whether the first failure occurs on the 2nd, 3rd, ..., or $n$th trial. The probabilities of these events add up to $1 - p^{n-1}$. Looking ahead from this failure, $E$ occurs with probability $P(E|H^*)$ in the rest of the sequence in each of these cases; given that we start out with a failure (this failure can be the first one of $m$ successive failures, in which case $E$ does not occur, so this failure must be taken into account.)
8 Some classical problems

8.1 Russian roulette

Russian roulette is a game, hopefully only played in a theoretical model and not in practice, in which a single round placed a revolver, a handgun with a rotating cylinder capable of holding six rounds. Two players, A and B, play alternate turns until the losing player gets killed as follows. Before the player puts the gun to his head, the chamber rotated to stop at a random location, then the player puts the gun to his head and pulls the trigger. The player survives only if the gun does not fire, i.e., if the active chamber does not contain a round. If he survives, it is the turn of the other player to do the same. Player A goes first. The mathematical question is, what is the probability that player A gets killed. This question gives a clear illustration why \( \sigma \)-additivity, stated in Clause (2) of Definition 4.2, is important, and finite additivity would not suffice (see the discussion after the definition quoted).

In order to more easily analyze the situation, we reformulate the game with a nonlethal description as follows. The players alternately roll a die, and the player that first rolls a 1 loses, but for easier analysis, assume that the game still goes on indefinitely. Let \( K_i \) for \( i = 1, 2, 3, \ldots \), be the event that the player on the \( i \)th turn rolls a 1. The events \( K_i \) are assumed to be independent, and we have \( P(K_i) = \frac{1}{6} \) for each \( i \).

Let \( H_i \) be the event that a 1 is rolled for the first time in the \( i \)th turn. We have

\[
P(H_i) = P(K_1) \cdot \prod_{j=1}^{i-1} P(K_j) = \frac{1}{6} \cdot \prod_{j=1}^{i-1} \frac{5}{6} = \frac{1}{6} \left( \frac{5}{6} \right)^{i-1};
\]

the first equation holds in view of the independence of the events in question (cf. Corollary 7.1). The events \( H_i \) are mutually exclusive. Player A loses if \( H_i \) occurs for some odd \( i \). That is, the probability that player A loses is

\[
P\left( \bigcup_{k=0}^{\infty} H_{2k+1} \right) = \sum_{k=0}^{\infty} \frac{1}{6} \left( \frac{5}{6} \right)^{2k} = \frac{1}{6} \sum_{k=0}^{\infty} \left( \frac{5^2}{6^2} \right)^k = \frac{1}{6} \cdot \frac{1}{1 - 5^2/6^2} = \frac{6}{36 - 25} = \frac{6}{11};
\]

the third equation uses the sum formula for the geometric series

\[
\sum_{n=0}^{\infty} x^n = \frac{1}{1 - x}.
\]

The reason for reinterpreting the original version with Russian roulette is the difficulty in imagining the event \( K_i \) meaning that the player in the \( i \)th turn gets killed in case he was killed earlier – it would have been somewhat problematic to deal with the independent events \( K_i \). One can still do a similar analysis with the events \( K_i \), involving conditional probabilities, but that is a bit more cumbersome.

8.2 Russian roulette revisited

Here we give a different analysis Russian roulette; for this, we will use the original description. Here, let \( A \) be the event the first player gets killed, and \( B \), the second player gets killed. Let \( T \) be the event that the first player gets killed in the first round. Noting that \( T \subset A \), thus \( T \cap A = T \), we have

\[
P(A) = P(T) + P(T^* \cap A) = P(T) + P(T^*) P(A | T^*) = P(T) + P(T^*) P(B);
\]

This is a reasonable assumption if the players don’t cheat.
the third equality holds since $P(A \mid T^*) = P(B)$, if the first player does not get killed in the first round then at the beginning of the second round he is in exactly the same situation as the second player at the beginning of the game: before his turn, the other player will play. Note that $P(T) = 1/6$, and so $P(T^*) = 5/6$. Given that $A \cap B = \emptyset$, and $A \cup B \cup (A^* \cap B^*) = \Omega$, we have $P(A) + P(B) + P(A^* \cap B^*) = 1$. Noting that $P(A^* \cap B^*) = 0$ we have $P(B) = 1 - P(A)$. With these observation, equation (8.3) becomes

$$P(A) = \frac{1}{6} + \frac{5}{6}(1 - P(A)).$$

Solving this equation, we obtain $P(A) = 5/11$. This agrees with the result we obtained above. It is interesting to note that this consideration implicitly evaluates the geometric series in (8.1) 8.2

8.3 Gambler’s ruin

The problem a version of which we will be discussing has a long history; see [12, Corollary on p. 129, pdf p. 135] for the history of the problem. In the version of the game we are going to discuss, there are two players, $A$ and $B$, $A$ having $a$ dollars and $B$ having $b$ dollars, where $a$ and $b$ are positive integers. In each turn, they toss a fair coin, and if it comes out head, $B$ gives $A$ one dollar, and if tail, $A$ gives $B$ one dollar. When one of the players has 0 dollars, he loses the game and the game ends. The player that loses is said to be ruined. The question is, what is the probability of $A$ losing the game.

In answering this question, let $P_n$ denote the probability that $A$ loses the game when he has $n$ dollars; at this point $B$ would have $a + b - n$ dollars. Let $A$ be the event that $A$ loses at this point, having $n$ dollars. Let $H$ and $T$ be the event that the next toss is head or tail. We have

$$P_n = P(A) = P((A \cap H) + P(A \cap T) = P(H)P(A \mid H) + P(T)P(A \mid T) = \frac{1}{2}P_{n+1} + \frac{1}{2}P_{n-1},$$

since if the next toss is head, $A$ will have $n + 1$ dollars, and if its a tail, he will have $n - 1$ dollars.

$$P_n = \frac{1}{2}P_{n+1} + \frac{1}{2}P_{n-1}.$$

This is a homogeneous linear recurrence equation, and it is well known how to solve such equations. The method of solving them is discussed in the section on Recurrence equations in [12, Corollary on p. 129, pdf p. 135]. According to this, such equations are associated with a polynomial equation, called the characteristic equation of the recurrence equation, and the zeros of the former determine the solutions of the latter. The characteristic equation in our case is $\zeta = (1/2)\zeta^2 + (1/2)$, i.e. $\zeta^2 - 2\zeta + 1 = 0$. This can be written as $(\zeta - 1)^2 = 0$, and its only zero is $\zeta = 1$, but this is a double zero, which somewhat complicates the situation. In any case, according to the result just quoted, the general solution is this recurrence equation is

$$P_n = C_1 + C_2n,$$

where $C_1$ and $C_2$ are arbitrary constants. In the present case, these constants can be determined from the initial conditions $P_{a+b} = 0$ and $P_0 = 1$. This is because if $A$ has $a + b$ dollars, then $B$ has lost, since he has 0 dollars, so the probability of $A$ losing is 0. On the other hand, if $A$ has 0 dollars, he has lost the game, so the probability of him losing is 1. Substituting these initial conditions, we

\footnote{We are not saying that $A^* \cap B^* = \emptyset$, since it is not impossible in principle that both players stay alive for ever, but this event has zero probability. We will leave the formal verification of this to the reader. The problem is that to do this formal analysis, we would have to repeat arguments similar to the preceding subsection.}
obtain \(0 = C_1 + (a+b)C_2\) and \(1 = C_1 + C_2 \cdot 0\); that is, \(C_1 = 1\) and \(C_2 = -1/(a+b)\). Substituting these into the equation describing the solution, we obtain \(P_a = 1 - n/(a+b)\). With \(n = a\), this gives the probability that \(A\) will be ruined at the start of the game:

\[
P_a = 1 - \frac{a}{a+b} = \frac{b}{a+b}.
\]

### 8.4 The probability of a union

Lemma 4.1 can be generalized as follows:

**Theorem 8.1.** Given an integer \(n \geq 1\) and events \(A_1, A_2, \ldots, A_n\), we have

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{S: S \subseteq \{1,2,\ldots,n\}, S \neq \emptyset} (-1)^{|S|+1} P\left(\bigcap_{i \in S} A_i\right).
\]

Here \(|S|\) indicates the number of elements of the set \(S\). In words, this formula says that to calculate the probability of a union of events, for all \(k\) with \(1 \leq k \leq n\), one has to take the probabilities of the intersection of any \(k\) of these events, and add this probability if \(k\) is odd and subtract it if \(k\) is even.

The validity of this result can easily be seen intuitively as follows. Assume \(x \in \Omega\) belongs to \(k > 0\) among the sets \(A_1, A_2, \ldots, A_n\). When taking the intersection of \(l\) of these sets for \(l \leq 1 \leq n\), then such intersections can be taken \(\binom{k}{l}\) ways. because we need to select \(l\) sets form among those containing \(x\). That is, in the sum of probabilities above, the number of ways \(x\) is counted is

\[
\sum_{i=1}^{k} \binom{k}{i} (-1)^{i+1} = (-1) \left(\sum_{i=1}^{k} \binom{k}{i} (-1)^i - 1\right) = -\left((1 + (-1))^k - 1\right) = 1,
\]

where the second equation follows by the binomial theorem. This shows that \(x\) is counted exactly once, as it should be. While this is not a rigorous proof, in a more advanced framework it can easily be reformulated in terms of integrals and characteristic functions as a rigorous proof. Here we will give a proof using induction that goes roughly along the lines of an induction proof of the binomial theorem. In Subsection 3.2.1 we gave a proof of the binomial theorem that does not use induction. The ideas of that proof, rather than the ideas outlined above, will be used to give a proof of Theorem 8.1 in Subsection 13.2 based indicator random variables (these are essentially the same as characteristic functions mentioned above, but we will work with the framework of random variables).

**Proof.** For \(n = 1\), the theorem simply says that \(P(A_1) = P(A_1)\), and for \(n = 2\), the result was proved as Lemma 4.1. So assume \(n \geq 3\), and assume the result is true if \(n\) is replaced by \(n-1\). We have

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(A_n \cup \bigcup_{i=1}^{n-1} A_i\right) = P(A_n) + P\left(\bigcup_{i=1}^{n-1} A_i\right) - P\left(A_n \cap \bigcup_{i=1}^{n-1} A_i\right) = P(A_n) + P\left(\bigcup_{i=1}^{n-1} A_i\right) - P\left(\bigcup_{i=1}^{n-1} (A_i \cap A_n)\right),
\]

\[^{8,3}\text{Given a a set } A, \text{ its characteristic function (defined on a set } B \text{ with } A \subseteq B \text{ is a function } \chi_A \text{ such that } \chi_A(x) = 1 \text{ if } x \in A \text{ and } \chi_A(x) = 0 \text{ if } x \in B \setminus A.}\]
where the second equation holds according to Lemma 4.1. Using the induction hypothesis for the second and third terms, we obtain

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = P(A_n) + \sum_{S: S \subseteq \{1,2,\ldots,n-1\} \text{ and } S \neq \emptyset} (-1)^{|S|+1} P\left(\bigcap_{i \in S} A_i\right) - \sum_{S: S \subseteq \{1,2,\ldots,n-1\} \text{ and } S \neq \emptyset} (-1)^{|S|+1} P\left(\bigcap_{i \in S} (A_i \cap A_n)\right)
\]

We have \(\bigcap_{i \in S} (A_i \cap A_n) = A_n \cap \bigcap_{i \in S} A_i\), and so

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = P(A_n) + \sum_{S: S \subseteq \{1,2,\ldots,n-1\} \text{ and } S \neq \emptyset} (-1)^{|S|+1} P\left(\bigcap_{i \in S} A_i\right) - \sum_{S: S \subseteq \{1,2,\ldots,n-1\} \text{ and } S \neq \emptyset} (-1)^{|S|+1} P\left(\bigcap_{i \in S} (A_i \cap A_n)\right).
\]

This can also be written as

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{S: S \subseteq \{1,2,\ldots,n\} \text{ and } n \notin S} (-1)^{|S|+1} P\left(\bigcap_{i \in S} A_i\right) + \sum_{S: S \subseteq \{1,2,\ldots,n\} \text{ and } n \in S} (-1)^{|S|+1} P\left(\bigcap_{i \in S} A_i\right),
\]

where, in the last term, the \(-\) sign was changed to \(+\), since previously, \(n\) was not counted as an element of \(S\) in the corresponding term; furthermore, the term \(P(A_n)\) on the right-hand side was incorporated into the second sum in case \(S = \{n\}\). Combining the two sums on the right-hand side into a single sum, we obtain equation \((8.4)\), completing the proof. \(\square\)

### 8.5 Rencontre

The word “rencontre” means encounter in French. In this game, \(n > 0\) paper slips numbered 1, 2, \ldots, \(n\) placed in a hat. An encounter is the event that for some \(i\) with \(1 \leq i \leq n\), the slip labeled \(i\) is drawn. Let this event be denoted by \(A_i\). In this case, \(\Omega\) can be taken to be the set of all permutations \(S_n\) of the set \(\{k: 1 \leq k \leq n\}\), \(\text{8.4}\) each of these being equally likely. \(\text{8.5}\) Interpreting these permutations as one-to-one mappings of this set onto itself (cf. Subsubsection 3.1.2), \(\sigma \in A_i\) means \(\sigma(i) = i\). That is,

\[A_i = \{\sigma \in S_n : \sigma(i) = i\}.
\]

The probability of a rencontre, \(P\left(\bigcup_{i=1}^{n} A_i\right)\), can be calculated using formula \((8.4)\). To use this formula, note that for \(S \subseteq \{k: 1 \leq k \leq n\}\), we have

\[\bigcap_{i \in S} A_i = \{\sigma \in S_n : \sigma(i) = i \text{ for } i \in S\}.
\]

\(\text{8.4}\) In fact, the symbol \(S_n\) for this set of permutations is standard in elementary group theory.

\(\text{8.5}\) That is, for each \(\sigma \in \Omega\), \(P(\{\sigma\}) = 1/n!\). One always needs to remember that \(P(\sigma)\) is meaningless: elements of \(\Omega\) are not events; events are subsets of \(\Omega\).
It is easy to count the number of these permutations: the numbers not belonging to $S$ can be freely permuted, and so this number is $(n-|S|)!/n!$. Given $r$ with $1 \leq r \leq n$, in formula (8.4) there are $\binom{n}{r}$ terms for which $|S| = r$; that is, the sum of probabilities associated with these terms is

$$
\binom{n}{r} \frac{(n-r)!}{n!} \cdot \frac{(n-r)!}{n!} = \frac{1}{r!}.
$$

These sums must be multiplied by $(-1)^{r+1}$ and added for $r$ with $1 \leq r \leq n$ to obtain the left-hand side of formula (8.4). That is,

$$
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{r=1}^{n} \frac{(-1)^{r+1}}{r!} = 1 + \sum_{r=n+1}^{\infty} \frac{(-1)^{r}}{r!} - \sum_{r=0}^{\infty} \frac{(-1)^{r}}{r!}.
$$

On the right-hand side, the second sum equals $1/e$ and the first sum tends to 0 when $n \to \infty$. Thus, for large $n$, the probability is approximately $1 - 1/e$.

9 Random variables

Intuitively, a random variable is a quantity $X$ that assumes random values. In a more mathematical setting, given a probability space $\Omega$, a random variable is a function $X : \Omega \to \mathbb{R}$, where $\mathbb{R}$ is a set of real numbers. Since we will be calculating probabilities involving random variables, we need to impose additional restrictions on $X$ so that certain probabilities are meaningful:

**Definition 9.1.** Let $(\Omega, \mathcal{B}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called a random variable if for all real numbers $\alpha$ we have \{\omega \in \Omega : X(\omega) \leq \alpha\} \in \mathcal{B}.

This means that if $X$ is a random variable, then for all $\alpha \in \mathbb{R}$ the probability on the right-hand side of the equation

$$(9.1) \quad F_X(\alpha) \overset{\text{def}}{=} P\{\omega \in \Omega : X(\omega) \leq \alpha\}
$$

is meaningful. The function $F_X$ defined by this equation is called the **distribution function** of the random variable $X$.

In many considerations involving random variables, there are different possible choices for the space $\Omega$, and often $\Omega$ is not even mentioned in the discussion explicitly. For this reason, it is customary to denote the event \{\omega \in \Omega : X(\omega) \leq \alpha\} as $[X \leq \alpha]$ or more simply as $X \leq \alpha$. Given a random variable $X$ and a real number $\alpha$, the probabilities of the events $X < \alpha$, $X \geq \alpha$, $X > \alpha$, $X = \alpha$, and $X \neq \alpha$ are also defined. To see this, note that

$$(9.2) \quad \{\omega \in \Omega : X(\omega) < \alpha\} = \bigcup_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq \alpha - \frac{1}{n}\};$$

Recall that

$$e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!}
$$

for all $x$.

That is, we are discussing real-valued random variables here. In a more general setting, a random variable may assume complex values, matrix values, etc.

Also called the **cumulative distribution function**.
hence it follows that the probability of the event $X < \alpha$ is defined in view of Clause [3] of Definition 4.1. The event $X \geq \alpha$ is the complement of $X < \alpha$, and the event $X > \alpha$ is the complement of $X \leq \alpha$, so these events also have probabilities. Further, we have

$$\{\omega \in \Omega : X(\omega) = \alpha\} = \{\omega \in \Omega : X(\omega) \leq \alpha\} \cap \{\omega \in \Omega : X(\omega) \geq \alpha\}.$$  

Further, for any interval $I$, the probability of the event $X \in I$ is also defined. For example, if $I = (\alpha, \beta]$, then

$$\{\omega \in \Omega : X(\omega) \in I\} = \{\omega \in \Omega : X(\omega) \leq \beta\} \cap \{\omega \in \Omega : X(\omega) > \alpha\}.$$ 

An important consequence of formula (9.2) is that

$$(9.3) \quad P(X < \alpha) = \lim_{n \to \infty} F_X(\alpha - \frac{1}{n}) = \lim_{x \to \alpha} F_X(x).$$

### 9.1 Functions of random variables

If $X$ is a random variable on $\Omega$ and if $g : \mathbb{R} \to \mathbb{R}$ is a function, then the composition $g \circ X$ is a function defined on $\Omega$ with real numbers as values. Whether it is a random variable depends on whether the condition given in Definition 9.1 is satisfied. This certainly happens if $g$ is an nice function. The random variable $g \circ X$ is usually denoted as $g(X)$, since, intuitively, one thinks of $X$ as a random quantity rather than a function.

### 9.2 Reading

[19, Chapter 4, pp. 119–188]. This reading is related to the material of several of the coming section in these notes. While the material here is not covered in exactly in the same order as in [19], the present section mainly concerns [19, §4.1–4.2].

### 9.3 Homework

[19] Chapter 4, pp. 175], Problems 1, 5, 6.

### 10 The Stieltjes integral

In order to give a unified treatment of random variables, we need the concept of the Stieltjes integral. The next three definitions describe the Riemann–Stieltjes integral.

**Definition 10.1** (Partition). A *partition* of the interval $[a, b]$ is a finite sequence $\langle x_0, x_1, \ldots, x_n \rangle$ of points such that

$$(10.1) \quad P : a = x_0 < x_1 < x_2 < \ldots < x_n = b.$$ 

The *width* or *norm* of a partition is

$$\|P\| = \max\{x_i - x_{i-1} : 1 \leq i \leq n\}.$$ 

9.3 The widest class of function $g$ for which this can be true is the class of *Borel measurable* functions; unfortunately, we cannot explain here what this class is. It certainly contains all continuous functions, but, actually, it is a much larger class.

29
Definition 10.2 (Riemann–Stieltjes sum). Given a partition

\[ P : a = x_0 < x_1 < x_2 < \ldots < x_n = b. \]

of the interval \([a, b]\), a tag for the interval \([x_{i-1}, x_i]\) with \(1 \leq i \leq n\) is a number \(\xi_i \in [x_{i-1}, x_i]\) for each \(i\). A partition with a tag for each interval \([x_{i-1}, x_i]\) is called a tagged partition. Given a tagged partition as described, and given the functions \(f\) and \(g\) on \([a, b]\), the corresponding Riemann–Stieltjes sum is

\[ S = \sum_{i=1}^{n} f(\xi_i)(g(x_i) - g(x_{i-1})). \]

The Riemann–Stieltjes integral

\[ \int_{a}^{b} f(x) \, dg(x) \]

is defined as the limit of the Riemann–Stieltjes sums \(S\) associated with the partition \(P\) as \(\|P\| \to 0\), independently of the choice of the tags. While not important for our purposes, we will give a rigorous definition (if we take \(g(x) = x\) in this definition, this gives the usual definition of the Riemann integral):

Definition 10.3 (Riemann–Stieltjes integral). If there is a real number \(A\) such that for every \(\epsilon > 0\) there is a \(\delta > 0\) such that for any Riemann–Stieltjes sum \(S\) for \(f\) and \(g\) associated with a partition of width \(< \delta\) of \([a, b]\) we have \(|A - S| < \epsilon\), then we call \(A\) the Riemann–Stieltjes integral of \(f\) with respect to \(g\) on \([a, b]\), and we write \(A = \int_{a}^{b} f \, dg\). In this case we call \(f\) Riemann–Stieltjes integrable with respect to \(g\) on \([a, b]\).

10.1 More on Stieltjes integrals

We will include here some simple results to put Stieltjes integrals in the proper context. The first one converts Stieltjes integrals into Riemann integrals in certain cases (but not in the case of interest to us above, when the function playing the role of \(g\) is not continuous).

Theorem 10.1. Assume \(g\) is differentiable on \([a, b]\). Assume further that the Riemann integral \(\int_{a}^{b} f(x)g'(x) \, dx\) and the Riemann–Stieltjes integral \(\int_{a}^{b} f(x) \, dg(x)\) exist. Then

\[ \int_{a}^{b} f(x) \, dg(x) = \int_{a}^{b} f(x)g'(x) \, dx. \]

Proof. Let

\[ P : a = x_0 < x_1 < x_2 < \ldots < x_n = b. \]

a partition of the interval \([a, b]\). By the mean-value theorem of differentiation, for each \(i\) with \(1 \leq i \leq n\) there is a \(\xi_i \in [x_{i-1}, x_i]\) such that \(g'(\xi_i)(x_i - x_{i-1}) = g(x_i) - g(x_{i-1})\). Hence we have

\[ \sum_{i=1}^{n} f(\xi_i)(g(x_i) - g(x_{i-1})) = \sum_{i=1}^{n} f(\xi_i)g'(\xi_i)(x_i - x_{i-1}). \]

Making \(\|P\| \to 0\), the left-hand side tends to \(\int_{a}^{b} f(x) \, dg(x)\) and the right-hand side tends to \(\int_{a}^{b} f(x)g'(x) \, dx\), completing the proof. \(\square\)

\(^{10}\) In order to apply the mean-value theorem, we need to assume that \(g\) is real valued, since the mean-value theorem is not valid for complex-valued functions. The result can nevertheless be proved in case \(g\) is complex valued by establishing it separately for the real and the imaginary parts of \(g\).
10.2 Integration by parts

**Theorem 10.2 (Integration by Parts).** Assume the integral \( \int_a^b f(x) \, dg(x) \) is defined. Then the integral \( \int_a^b g(x) \, df(x) \) is also defined and we have

\[
\int_a^b f(x) \, dg(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(x) \, df(x).
\]

**Proof.** For the proof, we redefine the concept of partition by allowing \( P = \{x_i : 1 \leq i \leq n\} \) to be a nondecreasing sequence. This is a harmless change, since the terms \( f(\xi_i)(g(x_i) - g(x_{i-1})) \) for which \( x_{i-1} = x_i \) do not contribute to the Riemann–Stieltjes sum. Let \( P \) be such an arbitrary partition; that is

\[
P : a = x_0 \leq x_1 \leq x_2 \leq \ldots \leq x_n = b,
\]

and let \( \xi_i \in [x_{i-1}, x_i] \) be arbitrary tags. We have the identity

\[
\sum_{i=1}^n f(\xi_i)(g(x_i) - g(x_{i-1})) = f(x_n)g(\xi_n) - f(\xi_1)g(x_0) + \sum_{i=1}^{n-1} g(x_i)(f(\xi_i) - f(\xi_{i+1})).
\]

This is easy to verify; namely, the same terms are added on both sides, in different order. Indeed, for \( i \) with \( 1 \leq i \leq n \), both sides adds the term \(+f(\xi_i)g(x_i)\) except that on the right-hand time for \( i = n \) this term is written out separately. Further, both sides adds the terms \(-f(\xi_i)g(x_{i-1})\) for \( i \) with \( 1 \leq i \leq n \), even though on the right-hand side this term is written as \(-f(\xi_{i+1})g(x_i)\) for \( i \) with \( 1 \leq i \leq n - 1 \), and the term corresponding to \( i = 0 \), i.e., the term \(-f(\xi_1)g(x_0)\), is written out separately. This rearrangement of a sum is called partial summation or Abel rearrangement, named after the Norwegian mathematician Niels Henrik Abel.

Making the assumption \( a = x_0 = \xi_1 = x_1 \) and \( x_{n-1} = \xi_n = x_n = b \), the above identity becomes

\[
\sum_{i=1}^n f(\xi_i)(g(x_i) - g(x_{i-1})) = f(b)g(b) - f(a)g(a) - \sum_{i=1}^{n-1} g(x_i)(f(\xi_{i+1}) - f(\xi_i)).
\]

Considering

\[
P' : a = \xi_1 \leq \xi_2 \leq \xi_3 \leq \ldots \leq \xi_n = b
\]

with the tags \( x_i \in [\xi_i, \xi_{i+1}] \) for \( i \) with \( 1 \leq i \leq n-1 \), the right-hand side contains a Riemann–Stieltjes sum for the integral \( \int_a^b g(x) \, df(x) \), and the left-hand side contains a Riemann–Stieltjes sum for the integral \( \int_a^b f(x) \, dg(x) \); the fact that we allow \( x_{i-1} = x_i \) makes no difference here, since the terms with \( x_{i-1} = x_i \) make no contribution to the sum. Since \( \xi_{i-1} \leq x_i \leq \xi_i \leq x_{i+1} \leq \xi_{i+1} \) for all \( i \) with \( 1 \leq i \leq n - 1, x_0 = x_1, x_{n-1} = x_n \), we have \( \|P\| \leq 2\|P'\| \). Hence, making \( \|P'\| \to 0 \), we also have \( \|P\| \to 0 \); hence the left-hand side tends to \( \int_a^b f(x) \, dg(x) \), since this integral was assumed to exist. So, the right-hand side also has a limit; thus, the integral \( \int_a^b g(x) \, df(x) \) also exists, it being the limit of the sum on the right-hand side. This completes the proof of the theorem.

Using Theorems 10.1 and 10.2 we obtain the integration by parts formula for Riemann integrals:

\[
\int_a^b f(x)g'(x) \, dx = f(b)g(b) - f(a)g(a) - \int_a^b f'(x)g(x) \, dx
\]

under appropriate conditions. We leave the formulation of these conditions to the reader.

---

\[10.2\] The equality \( \xi_i = \xi_{i+1} \) is possible, whether or not we allow the possibility that \( x_{i-1} = x_i \). This causes no trouble, just as allowing \( x_{i-1} = x_i \) causes no trouble.
10.3 Change of variables

We also have a change of variables (i.e., substitution) formula for Riemann–Stieltjes integrals; it is even simpler than the one for regular Riemann integrals. For this, we need to put

\[ \int_a^b f(x) \, dg(x) \overset{\text{def}}{=} - \int_a^b f(x) \, dg(x) \quad (a < b), \]

as is usual in case of Riemann integrals. At this point, we might as well put \( \int_a^b f(x) \, dg(x) = 0 \) in case \( a = b \).

**Theorem 10.3.** Assume the integral \( \int_a^b f(x) \, dg(x) \) exists, and let \( h : [A, B] \to [a, b] \) be a nondecreasing or nonincreasing function onto \([a, b]\). Then the integral \( \int_A^B f(h(t)) \, dg(h(t)) \), exists and we have

\[ (10.4) \quad \int_A^B f(h(t)) \, dg(h(t)) = \int_{h(A)}^{h(B)} f(x) \, dg(x). \]

Note that \( h(A) = a \) and \( h(B) = b \) in case \( h \) is nondecreasing, and \( h(A) = b \) and \( h(B) = a \) in case \( h \) is nonincreasing. As for the proof, it is fairly direct and straightforward except that it involves simple results about uniform continuity, and so we omit the proof. Readers familiar with uniform continuity can easily construct a proof.

**Problem 10.1.** Let \( f \) be a function on \([-1,1]\) that is continuous at 0, and let \( g \) be the function that is

\[ g(x) = \begin{cases} 0 & \text{if } -1 \leq x < 0, \\ 1 & \text{if } 0 \leq x \leq 1. \end{cases} \]

Show that

\[ \int_{-1}^1 f(x) \, dg(x) = f(0). \]

**Solution.** Let

\[ P : -1 = x_0 < x_1 < x_2 < \ldots < x_n = 1 \]

be a partition and let \( \xi_i \in [x_{i-1}, x_i] \) be a tag for each \( i \) with \( 1 \leq i \leq n \). Let \( k = k(P) \) with \( 1 \leq k \leq n \) be such that \( x_i < 0 \) for \( i < k \) and \( x_i \geq 0 \) for \( i \geq k \); clearly, \( k \) depends on the partition \( P \). Then

\[ g(x_i) - g(x_{i-1}) = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k. \end{cases} \quad (1 \leq i \leq n). \]

Hence

\[ S(P) = \sum_{i=1}^n f(\xi_i) (g(x_i) - g(x_{i-1})) = f(\xi_k) = f(\xi_{k(P)}); \]

\( S(P) \) depends on also on the tags, not just on \( P \), but this dependence is not indicated. Making \( \|P\| \to 0 \), we have \( \xi_{k(P)} \to 0 \). Since \( f \) is continuous at 0, we have

\[ \int_{-1}^1 f(x) \, dg(x) = \lim_{\|P\| \to 0} S(P) = \lim_{\|P\| \to 0} f(\xi_{k(P)}) = f(0). \]

\[ ^{10.3} \text{A function } h \text{ satisfying the requirements of Theorem } 10.3 \text{ is necessarily continuous, and so also uniformly continuous.} \]
10.4 Improper Stieltjes integrals

In probability theory, Stieltjes integrals are often used on the interval \((-\infty, \infty)\). The above definition only allows the use of finite intervals. For infinite intervals, the integral is taken as the limit of integrals over finite intervals. Such integrals are called improper Stieltjes integrals. We have

\[
\int_a^\infty f(x) \, dg(x) \overset{\text{def}}{=} \lim_{b \to +\infty} \int_a^b f(x) \, dg(x),
\]

\[
\int_{-\infty}^b f(x) \, dg(x) \overset{\text{def}}{=} \lim_{a \to -\infty} \int_a^b f(x) \, dg(x),
\]

and

\[
\int_{-\infty}^\infty f(x) \, dg(x) \overset{\text{def}}{=} \lim_{a \to -\infty} \lim_{b \to +\infty} \int_a^b f(x) \, dg(x).
\]

In the last limit, \(a \to -\infty\) and \(b \to +\infty\) independently.\(^{10.4}\)

11 Expectation

Intuitively, the expectation of a random variable will be the average value of a random variable, weighted by the probability of assuming this or nearby values. A definition that works for a large class of distributions is as follows.

**Definition 11.1.** Let \(H_i \subset \Omega\) pairwise disjoint nonempty events, where \(1 \leq i \leq n\) for some integer \(n\) or \(1 \leq i < \infty\), such that \(\bigcup_i H_i = \Omega\). Let \(X\) be a random variable such that \(X\) is constant on each \(H_i\). Let \(\omega_i \in H_i\) for each \(i\). Then the expectation \(E(X)\) of \(X\) is defined as

\[
E(X) = \sum_i X(\omega_i) P(H_i);
\]

in case the sum on the right-hand side is an infinite sum, we need to assume the sum is absolutely convergent; if it is not, \(E(X)\) will not be defined.

The assumption of absolute convergence is necessary in view of footnote \(^{11.4}\) on p. \(16\). There it is explained the the sum on the right-hand side of equation (11.1) is not defined without the assumption of absolute convergence, since there is no clear way to specify the order of the summands. Another issue that arises with the above definition is its soundness; that is, one needs to show that it actually defines something. This is an issue at present, since given a random variable \(X\) for which such events \(H_i\) can be found, the choice of these events is not unique. What needs to be proven is that if one chooses a different collection \(K_j\) of events with the same property, the corresponding sum in equation (11.1) gives the same value. This is not very difficult to do, but we will omit a proof here.

Given a general random variable \(X\), the expectation of \(X\) can be defined by approximating

\[^{10.4}\text{That is, the limit is approached when } a \text{ is a very large negative number and } b \text{ is a very large positive number, but there is no connection between } a \text{ and } b; \text{ for example, it would be wrong to assume that } a = -b. \text{ We will not give a precise definition. Suffice it to say that instead of defining the last integral as a double limit, we can define it as}

\[
\int_{-\infty}^0 f(x) \, dg(x) + \int_0^\infty f(x) \, dg(x).
\]
X with random variables for which the expectation is determined by the above definition – we omit the specific details.

This only gives the expectation of certain large class of random variables, to be called discrete (see a discussion below, in Section 12) it is possible to define expectation for a much larger class by approximating those random variables by members of the class for which it was defined above. The general definition suggested by this line of thought is a kind of integral:

(11.2) \[ E(X) = \int_{\Omega} X(\omega) \, dP(\omega). \]

Unfortunately, the precise definition of the integral on the right is beyond the level of the course. Therefore, we have to take a less elegant approach, and define expectation in terms of an improper Stieltjes integral (see Subsection 10.4).

11.1 Expectation via Stieltjes integrals

Let \( X \) be a random variable with distribution function \( F_X \) (see formula (9.1)) and let \( g \) be a nice function; we will consider the expectation of the random variable \( f(X) \); see Subsection 9.1. For the sake of simplicity, we will assume that \( g \) is continuous, though this assumption can be relaxed. We first assume that \( X \) is bounded; that is, there are \( a \) and \( b \) such that \( a < X < b \). Taking tags \( \xi_i \) as in Definition 10.2 the Riemann-Stieltjes sum

\[ S = \sum_{i=1}^{n} g(\xi_i) (F_X(x_i) - F_X(x_{i-1})) \]

where \( F_X \) is the distribution function (cf. (9.1)) of the random variable \( X \), approximate the Riemann-Stieltjes integral

(11.3) \[ \int_{a}^{b} g(x) \, dF_X(x) \]

Now, if we define the sets \( H_i \) as

\[ H_i = \{ \omega \in \Omega : x_{i-1} < X \leq x_i \}, \]

then \( P(H_i) = F_X(x_i) - F_X(x_{i-1}) \) according to (9.1), and we define the random variable \( X_S \) with \( X_S = \xi_i \) for \( 1 \leq X \leq n \). then the Riemann sum \( S \) can also be written as

\[ S = \sum_{i=1}^{n} g(X_S) P(H_i). \]

See [8, Chapter IX: Probability, pp. 184-215] for an introduction to probability theory based on modern integration theory.

11.2 Here \( a \) may be a huge negative number, and \( b \) may be a huge positive number, but they must be finite.
According to Definition 11.1, we have \( E(g(X_S)) = S \) for the Riemann-Stieltjes sums \( S \). If the integral in (11.3), this integral will be considered the expectation of \( E(X) \). If \( X \) is not bounded\(^{11.3}\), we need to take an improper Riemann–Stieltjes integral

\[
E(g(X)) = \int_{-\infty}^{\infty} g(x) \, dF_X(x),
\]

assuming that this integral is absolutely convergent, i.e., that the integral

\[
\int_{-\infty}^{\infty} |g(x)| \, dF_X(x),
\]

also exists. The assumption of absolute convergence is required since absolute convergence is important in Definition 11.1.

From the viewpoint of modern integration theory, the random variable \( X : \Omega \to \mathbb{R} \) is a measurable transformation, the differential \( dF_X(x) \) is just the measure \( d(PX^{-1}) \) and the expectation of the random variable \( g(X) \) expressed as

\[
E(g(X)) = \int_{\Omega} g(X(\omega)) \, dP(\omega)
\]

with the aid of formula (11.2) is rewritten in formula (11.4) as an integral on \( \mathbb{R} \) with the measure \( d(PX^{-1}) \) according to [8, §39 (Measurable Transformations), Theorem C, p. 163].

Note that in the above considerations the random variables \( X_S \) approximate the random variable \( X \), and in fact, the equation

\[
E(X) = \int_{-\infty}^{\infty} x \, dF_X(x)
\]

(this is equation (11.4) with \( g(x) = x \)) will be taken as the definition of \( E(X) \). As for equation (11.3), this should be taken as a consequence of the definition (quite a useful consequence in applications below), though it is not easy to derive the latter equation from the definition given in (11.3). In fact, this equation is not especially suited for theoretical discussions. For example, to derive the equation \( E(X+Y) = E(X)+E(Y) \) from this definition requires methods of multivariate calculus (see Section 29). The definition in equation (11.2) is much more suited for such discussion - unfortunately, that definition requires more mathematical background than is supposed in these notes.

In equation (11.5) it is no longer necessary to explicitly assume absolute convergence, since this is implicitly ensured by the definition of the improper Stieltjes integral in Subsection 10.4 that the in the limit \( a \to -\infty \) and \( b \to \infty \) \( a \) and \( b \) change independently (most importantly, it is not assumed that \( a = -b \). If the integral in this equation is not convergent, then \( E(X) \) will not be defined.

11.2 Reading

Continue reading [19, Chapter 4, pp. 119–188]. The current section is related to [19, §4.3–4.4]

11.3 Homework

[19, Chapter 4, pp. 177], 25, 27, 28, 33.

\(^{11.3}\)We may also allow \( g \) not to be bounded, but we would still want to require that \( g \) is bounded on finite integrals. More precisely, it is necessary to require that the Riemann–Stieltjes integral exists. The function \( F_X \) is clearly nondecreasing, and it is known that if \( g \) is continuous and \( F_X \) is nondecreasing, then the integral \( \int_{a}^{b} g \, dF_X \) exists, but it may exist also under less stringent assumptions.
12 Discrete and continuous random variables

To overcome the difficulty presented by not being able to use the definition of expectation given in equation (11.2), we will need to a traditional distinction of discrete and continuous random variables. While not all random variables fall in either of these two classes, this will be sufficient for us to discuss a wide range of material. The distinction between random and continuous random variables is unnecessary in a more advanced presentation.

12.1 Discrete random variables

A **discrete** random variable $X$ is one that only assumes countably many values. This is exactly the kind of random variable covered by Definition 11.1. Instead of using the distribution function, it is customary to use the probability function for such a variable. Given a discrete random variable, this is the function

\[(12.1) \quad p_X(x) = P(X = x).\]

This function can be defined for all real $x$, but it will be nonzero only for the countably many values $x \in \text{ra}(X)$, i.e., only if $X$ actually assumes the value $x$ (and even then, it may be zero). By formula (9.3), it is easy to see that

\[(12.2) \quad p_X(x) = F_X(x) - \lim_{t \nearrow x} F_X(t).\]

The expectation of a discrete $X$ can be calculated as

\[(12.3) \quad E(X) = \sum_{x \in \text{ra}(X)} xp_X(x).\]

This corresponds to formula (11.1) when writing $x_i$ for the elements of $\text{ra}(X)$ and $H_i = \{\omega \in \Omega : X(\omega) = x_i\}$, assuming this sum is absolutely convergent. If $g : \mathbb{R} \to \mathbb{R}$ is a function, then we can calculate the expectation of $g(X)$ as

\[E(g(X)) = \sum_{x \in \text{ra}(X)} g(x)p_X(x),\]

again assuming absolute convergence. This equation is also covered by Definition 11.1.

12.2 Continuous random variables

A random variable $X$ is called continuous if its distribution function $F_X$ is continuous. We need to impose more restrictions of the distribution function for us to be able to calculate the expectation of $X$ without Riemann–Stieltjes integrals. In addition to assuming that $F_X$ is continuous, we will assume that it is differentiable, perhaps with the exceptions of a few points (to be discussed later). The derivative of $F_X'$ is called the **density function** of the random variable $X$:

\[(12.4) \quad f_X(x) \overset{\text{def}}{=} F_X'(x) = \frac{d}{dx} P\{\omega \in \Omega : X(\omega) \leq x\}.\]

---

12.1 The word “discrete” has been discussed before. See footnote 5.1 on p. 18.

12.2 $\text{ra}(X)$ is the range of the function $X$, i.e., the set of values assumed by the random variable $X$.

12.3 The choice of $H_i$ in Definition 11.1 is of course not unique, but the one we gave here is a suitable choice.
The issue with the definition of the density function is that for the density function to be useful, we need to have

\begin{equation}
F_X(x) = \int_{-\infty}^x f_X(t) \, dt
\end{equation}

for all \(x \in \mathbb{R}\). If \(f_X(x) = F'_X(x)\) for all \(x \in \mathbb{R}\), then this equation is satisfied, but it is known to hold also in other situations. For example, if \(f_X\) is Riemann integrable on all finite intervals, the condition is satisfied even if \(f_X\) is not continuous everywhere. But there are other situations when a density function \(f_X\) satisfying equation (12.5) exists even when \(F_X\) is not differentiable everywhere.

According to Theorem 10.1, the expectation of \(X\) given by formula (11.5) can be calculated as

\begin{equation}
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx
\end{equation}

assuming that \(f_X\) is continuous. Even if \(f_X\) is continuous, this integral may not exist; but then the Riemann-Stieltjes integral in (11.5) does not exist, either, and \(E(X)\) is not defined. Similarly, given a nice function \(g : \mathbb{R} \to \mathbb{R}\), according to formula (11.4), the expectation of \(g(X)\) can be calculated as

\begin{equation}
E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx.
\end{equation}

Here, in addition to the continuity of \(f_X\), we need to assume that this integral is absolutely convergent – the reason is similar to the situation with formulas (11.4) and (11.5), where we needed to assume absolute convergence for the former, while for the latter this was ensured by the definition of the improper integral (note that \(f_X \geq 0\) since \(F_X\) is nondecreasing). The assumption on the continuity can be somewhat relaxed. For example, we may even allow \(f_X = F'_X\) to be not defined at finitely many points.

13 Properties of expectation

13.1 Linearity of expectation

If \(X\) and \(Y\) are random variables and \(c\) is a real number such that the expectations of \(X\) and \(Y\) exist, then we have

\begin{equation}
E(X + cY) = E(X) + c E(Y).
\end{equation}

This follows immediately from equation (11.2); unfortunately, the integral in this equation cannot be explained in these notes. While equation (11.5) could be taken as a definition of expectation, proving the above equation by using this equation is too cumbersome, and requires multivariate calculus. The approach we are going to take is to note that it is sufficient to approach the result for discrete random variables, since all random variables can be approximated by discrete random variables.
variables; in fact, in the proof of (11.4) and (11.5) we used the discrete random variable $X_S$ to approximate $X$.

For discrete variables, formula (11.2) is not the one to use. Instead, we can use equation (11.1) in Definition 11.1 with sets $H_i$ that work for both of the discrete random variables $X$ and $Y$. Such sets can be defined as the nonempty ones among all the sets

$$\{\omega : X(\omega) = x \text{ and } Y(\omega) = y\}$$

for $x, y \in \mathbb{R}$. Since $X$ and $Y$ assume only countable many values, the number of these sets is countable. Labeling them as $H_i$ and taking $\omega_i \in H_i$ for some or all positive integers $i$ (depending on whether there are finitely many or infinitely many among these sets), according to equation (11.1) we have

$$E(X + cY) = \sum_i (X(\omega_i) + cY(\omega_i)) P(H_i)$$

$$= \sum_i X(\omega_i) P(H_i) + c \sum_i Y(\omega_i) P(H_i) = E(X) + cE(Y),$$

establishing equation (13.1) for discrete random variables.

### 13.1.1 Expectation of a constant

A constant $a$ can be considered as a random variable $Y$ such that $Y(\omega) = a$ all the time. We naturally have $E(X) = a$ in this case. While this remark may seem frivolous, using equation (13.1) it shows that

(13.2) \[ E(X + a) = E(X) + a. \]

### 13.2 Probability of union, revisited

Given an event $A$, its indicator random variable $I_A$ is the random variable defined as

(13.3) \[ I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases} \]

It is immediate that $E(I_A) = P(A)$. We are going to give a new proof of Theorem 8.1. Indeed, writing $I_i$ for the indicator variable of the event $A_i$ for $i$ with $1 \leq i \leq n$ in that theorem, then we have

$$\prod_{i=1}^{n} (1 + (-1)I_i) = \sum_{k=0}^{n} \sum_{S : S \subseteq \{1,2,\ldots,n\}} \prod_{i \in S} (-1)I_i;$$

indeed, this is just equation (3.5) with $x = 1$ and $y_i = (-1)I_i$; here there was no reason here to break up the sum according to the cardinality of $S$, but we wanted to show the parallel with equation (3.5).

This can be written more appropriately for our purposes as

$$1 - \prod_{i=1}^{n} (1 - I_i) = 1 + \sum_{S : S \subseteq \{1,2,\ldots,n\}} (-1)^{|S|+1} \prod_{i \in S} I_i = \sum_{S : S \subseteq \{1,2,\ldots,n\}} \sum_{i \in S} (-1)^{|S|+1} \prod_{i \in S} I_i;$$

\[13.1\] The easy calculation is shown in equation (15.1), where indicator variables are discussed in more detail.
here, the minus sign before the product is accounted for by the 1 in the exponent of \((-1)^{|S|+1}\) after the first equation. Further, the term corresponding to \(S = \emptyset\) cancels against the 1 added to the sum – so this 1 does not appear on the right-hand side. Taking expectations and using equation (13.1), this equation becomes

\[
E\left(1 - \prod_{i=1}^{n}(1 - I_i(\omega))\right) = \sum_{S: S \subset \{1, 2, \ldots, n\}, S \neq \emptyset} (-1)^{|S|+1} E\left(\prod_{i \in S} I_i\right).
\]

Note that

\[
\bigcup_{i=1}^{n} A_i = \Omega - \bigcap_{i=1}^{n} A_i^c,
\]

we have

\[
1 - \prod_{i=1}^{n}(1 - I_i(\omega)) = 1 \quad \text{if and only if} \quad \omega \in \bigcup_{i=1}^{n} A_i;
\]

furthermore,

\[
\prod_{i \in S} I_i(\omega) = 1 \quad \text{if and only if} \quad \omega \in \bigcap_{i \in S} A_i.
\]

Using these, equation (13.4) becomes identical to equation (8.4), giving an alternate proof of Theorem 8.4.

### 13.3 Independent random variables

Given \(n \geq 1\), the random variables \(X_1, X_2, \ldots, X_n\) are called independent, if for all intervals \(I_i \subset \mathbb{R}\) \((1 \leq i \leq n)\), the events \(\{\omega \in \Omega : X_i(\omega) \in I_i\}\), \(1 \leq i \leq n\), are independent. Here the intervals can be closed, open, finite or infinite, and the one-point interval \([a, a]\) = \{a\} for \(a \in \mathbb{R}\) is allowed. It is not difficult to verify that if these assumptions are only made for intervals of type \(I_i = (-\infty, a_i]\), then they can be extended to all types of intervals.\(^{13.2}\)

We have

**Theorem 13.1.** Let \(n \geq 1\) be an integer, and let \(X_1, X_2, \ldots, X_n\) be independent random variables. Then

\[
E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).
\]

We will prove this only for \(n = 2\), and only in case the variables \(X_i\) are discrete. The extension to larger values of \(n\) is fairly routine, and the extension to arbitrary random variables can be done similarly as in the proof of equation (13.1).

**Proof.** Assume \(X\) and \(Y\) are independent discrete variables. We have

\[
E(XY) = \sum_{x,y \in \mathbb{R}} xy P(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\});
\]

in view of equation (11.1). This is because the sets

\[\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}\]

\(^{13.2}\)The easiest way to do this is by induction on the number of components for which we are allowed to replace the special kind of intervals \(I_i = (-\infty, a_i]\) with intervals of an arbitrary kind.
are pairwise disjoint (for different among these sets \(X\) or \(Y\) will assume different values, so those sets cannot have any element in common), there are only countably many nonempty ones among them since \(X\) and \(Y\) only assume countably many values, and those can be taken as \(H_i\) in equation (11.1). As \(X\) and \(Y\) are independent, we have

\[
P(\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}) = P(\{\omega \in \Omega : X(\omega) = x\}) \cdot P(\{\omega \in \Omega : Y(\omega) = y\}).
\]

Hence we have

\[
E(XY) = \sum_{x,y \in \mathbb{R}} xy P(\{\omega \in \Omega : X(\omega) = x\}) \cdot P(\{\omega \in \Omega : Y(\omega) = y\})
\]

\[
= \sum_{x \in \mathbb{R}} x P(\{\omega \in \Omega : X(\omega) = x\}) \cdot \sum_{y \in \mathbb{R}} y P(\{\omega \in \Omega : Y(\omega) = y\}) = E(X) E(Y),
\]

where the last equation also holds by equation (11.1). 

The following lemma is almost immediate.

**Lemma 13.1.** Let \(n \geq 1\), and assume the random variables \(X_i\) for \(1 \leq i \leq n\) are independent. Let \(a_i, 1 \leq i \leq n\) be real numbers. Then the random variables \(X_i + a_i, 1 \leq i \leq n\) are also independent.

**Proof.** Let \(I_i\) be an interval. We have \(X_i + a_i \in I_i\) if and only if \(X_i \in \{x - a_i : x \in I_i\}\). Noting that the set \(\{x - a_i : x \in I_i\}\) is also an interval, the result follows from the definition of independence. 

### 14 Variance

Given a random variable \(X\), its variance is defined as

\[
V(X) = E\left( (X - E(X))^2 \right).
\]

As we noted before, the expectation of a random variable is not always defined (see the discussion following equation (11.5).) *A fortiori* its variance is not always defined either. The variance of a random variable measures the deviation of a random variable from its expectation.

Some simple observations are in order. Let \(X\) be a random variable that has a variance, and let \(c \in \mathbb{R}\). Then

\[
V(X + c) = V(X).
\]

Indeed, writing \(Y = X + c\), we have \(E(Y) = E(X) + c\); hence \(Y - E(Y) = X - E(X)\). Hence \(V(Y) = V(X)\) according to equation (14.1), establishing (14.2). Given \(X\) and \(c\) as before, we also have

\[
V(cX) = c^2 V(X).
\]

In showing this, in view of (14.2) we may assume that \(E(X) = 0\) (since we may replace \(X\) with \(Y = X - E(X)\)), and \(E(Y) = 0\). Then we also have \(E(cX) = cE(X)\), and so

\[
V(cX) = E((cX)^2) = E(c^2 X^2) = c^2 E(X^2) = c^2 V(X),
\]

as we wanted to show.

---

\[\text{Latin, meaning “even more so,” or “with a stronger reason.”} \]
14.1 Moments; the calculation of variance

Given a positive integer \( n \), the \( n \)th moment of a random variable \( X \) is defined as \( \text{E}(X^n) \). The second moment has an important use in calculating the variance. Indeed, writing \( \mu = \text{E}(X) \), we have

\[
\text{V}(X) = \text{E}((X - \mu)^2) = \text{E}(X^2 - 2\mu X + \mu^2) = \text{E}(X^2) - 2\mu \text{E}(X) + \mu^2.
\]

For the last equation, note that \( \mu^2 \) is just a number, which can be taken as a random variable assuming a constant value, so its expectation is just \( \mu^2 \). Given that \( \mu = \text{E}(X) \), this shows with

\[
(14.4) \quad \text{V}(X) = \text{E}(X^2) - (\text{E}(X))^2,
\]

14.2 Pairwise independent variables

Given an integer \( n > 0 \), the random variables \( X_1, X_2, \ldots, X_n \) are called pairwise independent if any two of them are independent.

It is important to note that pairwise independent random variables are not necessarily independent, as the following example due to Sergei Bernstein shows. Let \( X_1, X_2, \) and \( X_3 \) be three random variables be defined as follows. Tossing a fair coin twice, let \( X_1 = 1 \) if the first toss is a head, 0 otherwise, let \( X_2 = 1 \) if the second toss is a head, 0 otherwise, let \( X_3 = 1 \) if exactly one of the two tosses is a head, 0 otherwise. We leave it to the reader to show that these variables are pairwise independent, but not independent.

The following theorem has important applications:

**Theorem 14.1.** Let \( n > 0 \) be an integer, and assume the random variables variables \( X_i \) for \( i \) with \( 1 \leq i \leq n \) pairwise independent each of which has variance. Then

\[
(14.5) \quad \text{V} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{V}(X_i).
\]

**Proof.** We may assume that \( \text{E}(X_i) = 0 \) for each \( i \), otherwise we may replace \( X_i \) with \( X_i - \text{E}(X_i) \) in view of equation (14.2); further, this replacement does not affect the assumption of pairwise independence in view of Lemma [13.1]. With this assumption we have

\[
\text{V} \left( \sum_{i=1}^{n} X_i \right) = \text{E} \left( \left( \sum_{i=1}^{n} X_i \right)^2 \right) = \text{E} \left( \sum_{i=1}^{n} \sum_{j=1}^{n} X_i X_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{E}(X_i X_j).
\]

For \( i = j \) the terms on the right-hand side become \( \text{E}(X_i^2) \). When \( i \neq j \), noting that \( X_i \) and \( X_j \) are independent, by Theorem [13.1] for these terms we have \( \text{E}(X_i X_j) = \text{E}(X_i) \text{E}(X_j) = 0 \), since we assumed that \( \text{E}(X_i) = 0 \) for each \( i \). Thus we have

\[
\text{V} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{E}(X_i^2) = \sum_{i=1}^{n} \text{V}(X_i),
\]

as we wanted to show. \( \square \)

14.3 Reading

Continue reading [19, Chapter 4, pp. 119–188]. The current section is related to [19, §4.5]
15 Some discrete distributions

15.1 One-point distribution
This is the distribution of the random variable $X$ for which $X(\omega) = a$ with some $a$ for all $\omega \in \Omega$. For his variable, we have

$$p_X(a) = 1.$$ for its probability function.

15.2 Two-point distribution
If the random variable $X$ assumes two values, $a$ and $b$. Its probability function is

$$p_X(a) = p, \quad p_X(b) = q,$$for some $p$ and $q$ with $0 \leq p \leq 1$ and $q = 1 - p$.

15.3 Indicator variables
Indicator variables have been described before, in equation (13.3). Given an event $A$, its indicator variable is defined as the variable $I_A = 1$ is $A$ occurs, and $I_A = 0$ otherwise. Let $p = P(A)$; then

$$p_X(1) = p, \quad p_X(0) = q,$$where $q = 1 - p$. We have

(15.1) \quad E(I_A) = 1 \cdot p + 0 \cdot q = p.$$

Noting that $I_A^2 = I_A$ we also have $E(I_A^2) = p$, and using equation (14.4), we obtain

(15.2) \quad V(I_A) = p - p^2 = p(1 - p) = pq,$$

15.4 Uniform distribution
In this distribution, the variable $X$ that assumes the values $1, 2, \ldots, n$ ($n > 0$) with equal probability is said to have uniform distribution. We have

$$p_X(i) = \frac{1}{n} \quad (1 \leq i \leq n).$$

15.5 Geometric distribution
Given two numbers real $p$ and $q$ with $0 < p < 1$ and $q = 1 - p$, a random variable with geometric distribution assumes the values $k$ for all integers $k$ with $k \geq 0$ such that

$$p_X(k) = q^k p \quad (k \geq 0).$$
Note that the total probability is 1, as it should be for all probability functions:
\[ \sum_{k=0}^{\infty} qX(k) = \sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = p \frac{1}{1-q} = 1, \]
where the third equation holds in view of the sum formula of the geometric series, and the last equation holds because \( p = 1 - q \).

### 15.6 For the first time (fft) distribution

This is almost the same as the geometric distribution, but here \( X \) assumes the values \( k \), where \( k \) is an integer with \( k \geq 1 \), and given numbers real \( p \) and \( q \) with \( 0 < p < 1 \) and \( q = 1 - p \), we have
\[ p_X(k) = q^{k-1}p \quad (k \geq 0). \]
This random variable can be realized by the following experiment. Assume in repeated independent trials of an experiment, each time the probability of success is \( p \), and numbering these trials by 1, 2, 3, \ldots, let \( X = k \) if the experiment succeeds first on the \( k \)th trial. The expectation of \( X \) can be calculated as follows.

\[
\begin{align*}
    E(X) &= \sum_{k=1}^{\infty} k p_X(k) = \sum_{k=1}^{\infty} k q^{k-1} = p \sum_{k=1}^{\infty} \frac{d}{dq} q^k = p \frac{d}{dq} \sum_{k=1}^{\infty} q^k \\
    &= p \frac{d}{dq} \left( -1 + \sum_{k=0}^{\infty} q^k \right) = p \frac{d}{dq} \left( -1 + \frac{1}{1-q} \right) \\
    &= p \frac{1}{(1-q)^2} = \frac{1}{1-q} = \frac{1}{p};
\end{align*}
\]
for the third equation, we made use of the theorem that power series can be termwise differentiated inside the interval of convergence.

For calculating the variance, we will make use of equation (14.4); instead of directly calculating the second moment \( E(X^2) \), it will, however, be more natural to first calculate \( E(X(X-1)) \), and note that \( E(X^2) = E(X) + E(X(X-1)) \). We have
\[
\begin{align*}
    E(X(X-1)) &= \sum_{k=1}^{\infty} k(k-1) p_X(k) = \sum_{k=2}^{\infty} k(k-1) p_X(k) = \sum_{k=2}^{\infty} k(k-1) p q^{k-1} \\
    &= q p \sum_{k=2}^{\infty} k(k-1) q^{k-2} = q p \sum_{k=2}^{\infty} \frac{d^2}{dq^2} q^k = q p \frac{d^2}{dq^2} \sum_{k=2}^{\infty} q^k \\
    &= q p \frac{2}{(1-q)^3} = q(1-q) \frac{2}{(1-q)^3} = \frac{2q}{(1-q)^2};
\end{align*}
\]
here the second equation holds since \( k-1 = 0 \) for \( k = 1 \). Thus,
\[ E(X^2) = E(X) + E(X(X-1)) = \frac{1}{1-q} + \frac{2q}{(1-q)^2} = \frac{1-q + 2q}{(1-q)^2} = \frac{1+q}{(1-q)^2}. \]

Hence, using equation (14.4), we have
\[ (15.4) \quad V(X^2) = E(X^2) - (E(X))^2 = \frac{1+q}{(1-q)^2} - \frac{1}{(1-q)^2} = \frac{q}{(1-q)^2} = \frac{q}{p^2}. \]
16 Binomial distribution

Let \( A \) be an event of probability \( p \) indicating the success of a certain trial. We want to repeat the trial \( n \) times (\( n > 0 \)) independently, and let \( A_i \) indicate that the \( i \)th trial is successful. Write \( q = 1 - p \). Given \( k \) with \( 0 \leq k \leq n \), the events

\[
\left( \bigcap_{i=1}^{k} A_i \right) \cap \left( \bigcap_{i=k+1}^{n} A_i^{*} \right)
\]

indicate that the first \( k \) trials are successful, and the rest of them are not. Since these events are independent (cf. Corollary 7.1), and the probability of \( A_i \) is \( p \), so that of \( A_i^{*} \) is \( 1 - p = q \), the probability of this event is \( p^k q^{n-k} \).

Let \( X \) be the random variable assuming values \( k \) with \( 0 \leq k \leq n \) such that \( X = k \) indicates that out of \( n \) independent repetition of the trials just described, exactly \( k \) is successful, and let \( 1 \leq i_1 < i_2 < \cdots < i_k \leq n \) describe the places where the trial is successful. For each fixed choice \( \{i_l : 1 \leq l \leq k \} \) of the \( k \)-element set the probability of such a result of the trials is the same as above, where successes occurred in the beginning. For different \( k \)-element sets these events are mutually exclusive, since if \( i \) is an element of one of the sets but not the others, this indicates a success at the \( i \)th place in one sequence of trials, and failure in the other sequence. Since there are \( \binom{n}{k} \) such \( k \)-element sets, the total probability of \( k \) successful outcomes, we have

\[
(16.1) \quad p_X(k) = P(X = k) = \binom{n}{k} p^k q^{n-k} \quad (0 \leq k \leq n).
\]

A variable \( X \) having this distribution is said to have a binomial distribution with parameters \( n \) and \( p \); in symbols, \( X \sim \text{Bin}(n, p) \).

16.1 Binomial variable as a sum of indicator variables

For \( i \) with \( 1 \leq i \leq n \), let \( I_i \) be the random variable defined as

\[
I_i(\omega) = \begin{cases} 
1 & \text{if } \omega \in A_i, \\
0 & \text{if } \omega \notin A_i;
\end{cases}
\]

that is, \( I_i = 1 \) if the \( i \)th trial is successful and \( I_i = 0 \) otherwise. Then we have for the binomial variable \( X \):

\[
X = \sum_{i=1}^{n} I_i.
\]

16.1 Some caution is needed here. The probability space corresponding to a single trial is not the same as the probability space corresponding to \( n \) repetitions of the trial. For example, rolling a die can be described by the probability space \( \Omega = \{i : 1 \leq i \leq 6\} \), but rolling a die twice needs to be described by the probability space \( \Omega' = \{(i, j) : 1 \leq i, j \leq 6\} \). In fact, \( \Omega' \) is the Cartesian product \( \Omega \times \Omega \). Similarly, the probability space corresponding to the \( n \) independent repetitions of the above trial can be described by a Cartesian product. We will omit the technical details.

16.2 For \( k = 0 \), the intersection for \( i = 1 \) to \( k \) indicates the empty intersection, which naturally would be meaningless, since it indicates the set of all sets, which is not permitted. However, note that the notation is allowed in a context when the intersection is taken with another set (see the comment at the end of Subsection 2.3), and then the resulting set is just the other set in the intersection, i.e. the intersection from \( k + 1 \) to \( n \). A similar discussion applied to the case \( k = n \).

16.3 The argument that follows is similar to the one used in Subsection 5.1 to describe picking marbles with replacement.
Hence, according to equation (15.1), we have

\[(16.2) \quad E(X) = \sum_{i=1}^{n} E(I_i) = np.\]

Similarly, given that the random variables \(I_i\), independent, by equation (15.2) and by Theorem 14.1 we have

\[(16.3) \quad V(X) = \sum_{i=1}^{n} V(I_i) = npq.\]

### 16.2 Direct determination of the expectation and the variance of the binomial distribution

While the direct determination of the expectation and the variance of a binomial variable is more complicated than using indicator variables, the method is worth studying since it shows important techniques of calculation. First note that for \(k\) and \(n\) with \(1 \leq k \leq n\) we have

\[(16.4) \quad \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)!}{k(k-1)!(n-k)!} = \frac{(n-1)!}{(k-1)!(n-k)!} = n \binom{n-1}{k-1}.\]

Using this for the binomial variable \(X\), we have

\[E(X) = \sum_{k=0}^{n} kpX(k) = \sum_{k=1}^{n} kpX(k) = \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k q^{n-k};\]

in the second equality, the zero term corresponding to \(k = 0\) was omitted from the summation, and in the fourth equality, we used equation (16.4). Hence, we have

\[E(X) = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} = np \sum_{l=0}^{n-1} \binom{n-1}{l} p^l q^{(n-1)-l} = np(p + q)^{n-1} = np;\]

here, in the second equation, we replaced the summation variable \(k\) with \(l = k - 1\). The third equation was obtained by the binomial theorem, and, for the fourth, note that \(p + q = 1\).

The calculation of the variance is using similar ideas. As in Subsection 15.6 to this end, we calculate \(E(X(X-1))\). we have

\[E(X(X-1)) = \sum_{k=0}^{n} k(k-1)pX(k) = \sum_{k=2}^{n} k(k-1)pX(k)\]

\[= \sum_{k=2}^{n} k(k-1) \binom{n}{k} p^k q^{n-k} = \sum_{k=2}^{n} n(n-1) \binom{n-2}{k-2} p^k q^{n-k};\]

here, in the second equality, we omitted the first two terms that are 0, and in fourth equality, we used equation (16.4) twice. Further, we have

\[E(X(X-1)) = n(n-1)p^2 \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} q^{(n-2)-(k-2)}\]

\[= n(n-1)p^2 \sum_{l=0}^{n-2} \binom{n-2}{l} p^l q^{(n-2)-l} = n(n-1)p^2(p + q)^{n-2} = n(n-1)p^2;\]
here, similarly as before in the second equation, we replaced the summation variable $k$ with $l = k - 2$.

The third equation was obtained by the binomial theorem, and, for the fourth, note that $p + q = 1$.

Hence, we have

$$E(X^2) = E(X) + E(X(X - 1)) = np + n(n - 1)p^2$$

$$= np - np^2 + n^2 p^2 = np(1 - p) + n^2 p^2 = np + n^2 p^2.$$ 

and

$$V(X) = E(X^2) - (E(X))^2 = npq + n^2 p^2.$$ 

16.3 Reading

Continue reading [19, Chapter 4, pp. 119–188]. The current section is related to [19, §4.6]

16.4 Homework

[19, Chapter 4, pp. 178], 40, 41, 43, 46, 53.

17 The Poisson distribution

The Poisson distribution describes the distribution of the following random variable $X$. Assume a certain event may happen several times in the time interval $[0, 1]$; the occurrences are independent, and in any time interval of length $\delta$ the probability of $A$ occurring is approximately $\lambda \delta$, where $\lambda > 0$ is a given constant, the approximation is better for small $\delta$, it becoming exact as $\delta \rightarrow 0$. That, for an intervals $I \subset [0, 1]$ we have

$$\lim_{|I| \rightarrow 0} \frac{P(A \text{ occurs in } I)}{|I|} = \lambda,$$

where $|I|$ denotes the length of the interval $I$. Then $X$ denotes the number of occurrences of $A$ in the interval $[0, 1]$. $X$ can assume the values $0, 1, 2, \ldots$.

To find the probability $P(X = k)$ for each $k \geq 0$, for a large integer $n$, divide the interval $[0, 1)$

$$\text{into } n \text{ equal parts: } I_i = \left[ \frac{i - 1}{n}, \frac{i}{n} \right] \quad (1 \leq i \leq n).$$

The probability of $A$ occurring in $I_i$ is approximately $\lambda/n$: these events being independent. So the probability of $A$ occurring in $k$ of these intervals follows the Binomial distribution $\text{Bin}(n, \lambda/k)$:

$$P(X = k) \approx \binom{n}{k} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k} = \frac{\prod_{i=0}^{k-1}(n-i) \lambda^k}{k!} \frac{n^k}{n^k} \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k}$$

$$= \frac{\lambda^k}{k!} \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k} \prod_{i=0}^{k-1} \frac{n^k}{n^k} \left( 1 - \frac{\lambda}{n} \right)^n \left( 1 - \frac{\lambda}{n} \right)^{-k} \prod_{i=0}^{k-1} \frac{n - i}{n}$$

Making $n \rightarrow \infty$ we obtain the exact formula for the probability function of $X$. Noting that we have

$$\lim_{n \rightarrow \infty} \left( 1 + \frac{x}{n} \right)^n = e^x$$

$\text{The occurrence of } A \text{ at any particular time (i.e., in an interval of length zero) is zero, so we omit the point 1 from the interval } [0, 1] \text{ so as to simplify the calculation.}$
for any real $x$ (see [14]), we find that

$$P(X = k) = \lim_{n \to \infty} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \prod_{i=0}^{n-k} \frac{n-i}{n} = \frac{\lambda^k}{k!} e^{-\lambda},$$

given that the limit of the last two factors is 1.

In the calculation we neglected the possibility that the event $A$ may occur more than once in an interval $I_i$. However, the probability for $A$ to occur $l$ times in an interval $I_i$ is about $\lambda^i / i!$, so using the formula

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad (|x| < 1)$$

for the geometric series, we can calculate the probability for $A$ to occur more than in the interval $I_i$ for a single value of $i$ is about

$$\sum_{l=1}^{\infty} \frac{\lambda^l}{n^l l!} = \frac{\lambda^2}{n^2} \sum_{l=2}^{\infty} \frac{\lambda^{l-2}}{n^{l-2} (l-2)!} = \frac{\lambda^2}{n^2} \sum_{l=2}^{\infty} \frac{\lambda^{l-2}}{n^{l-2} (l-2)!} = \frac{\lambda^2}{n^2} \sum_{m=0}^{\infty} \frac{\lambda^m}{n^m} = \frac{\lambda^2}{n^2} \frac{1}{1 - \lambda/n} = \frac{\lambda^2}{n(n-\lambda)},$$

where the third equation was obtained by replacing the summation variable $l$ by $m + 2$, in which case $m = l - 2$ and the summation runs from $m = 0$ to $m = \infty$. So the probability that $A$ occurs at least once in one of the intervals $I_i$ for some $1 \leq i \leq n$ is at most $\lambda^i / (n-\lambda)$; actually, it is somewhat less, since the events $A$ occurring more than once in $I_i$ for different values of $i$ are not mutually exclusive.

Note that the probabilities $P(X = k)$ for $k = 0, 1, 2, \ldots$ add up to 1, as they should:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^\lambda = 1,$$

where the penultimate equality holds in view of the Taylor series of $e^\lambda$. The expectation and the variance of the Poisson distribution can be calculated as the limit of the approximating binomial distribution. Here we give direct calculation.

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

here in the second equation we changed the lower limit of summation from $k = 0$ to $k = 1$, since the term for $k = 0$ is 0 anyway (since it is multiplied by $k$). For the third equation, we moved $k$ to the denominator as $1/k$, and for the fourth equation we used the fact that $k! = (k-1)! k$ for $k \geq 1$ ($0! = 1$ by definition) in the form $k!/k = (k-1)!$. We also wrote $\lambda^k$ as $\lambda \lambda^{k-1}$.

Next, we will move $\lambda$ and $e^{-\lambda}$ outside from the scope of the summation sign, and replace the summation variable $k$ with $l + 1$, in which case $l = k - 1$; so we will sum from $l = 0$ to $l = \infty$. We obtain

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

\[\begin{align*}
17.2 & \text{We are intentionally using } n \text{ as the summation variable to stress the point that the summation variable has no meaning outside the sum, so this use of } n \text{ does not conflict of the use of } n \text{ denoting the number of interval. We would have a problem if we needed to use that } n \text{ inside the summation. Further note that } 0^0 \text{ is meaningless, but in situations where the sum involves } x^n, \text{ for } n = 0 \text{ and } x = 0 \text{ we take } x^n \text{ to be } 1 \text{ by a wide-spread mathematical convention.} \\
17.3 & \text{The one before the last one.}
\end{align*}\]
where for the second equation we used the Taylor series of $e^\lambda$.

The variance $V(X)$ of $X$ can be calculated from the formula $V(X) = E(X^2) - (E(X))^2$, where, to calculate the first term on the right-hand side, we will use the formula $E(X^2) = E(X(X-1)) + E(X)$. We have

$$E((X(X-1)) = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda}$$

$$= \sum_{k=2}^{\infty} \frac{\lambda^k}{k!/(k(k-1))} e^{-\lambda} = \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} e^{-\lambda};$$

Similarly to the calculation of $E(X)$, we omitted the terms for $k = 0$ and $k = 1$, which are zero in view of their being multiplied by $k(k-1)$. After the third equation, we moved $k(k-1)$ to the to the denominator, and then noted that $k!/(k(k-1)) = (k-2)!$ for $k \geq 2$. We also replaced $\lambda^k$ with $\lambda^{k-2}$. We will next replace the summation variable $k$ with $l + 2$, in which case $l = k - 2$; so we will sum from $l = 0$ to $l = \infty$. We obtain

$$E((X(X-1)) = \lambda^2 e^{-\lambda} \sum_{l=0}^{\infty} \frac{\lambda^l}{l!} = \lambda^2 e^{-\lambda} e^\lambda = \lambda^2,$$

where for the second equation we used the Taylor series of $e^\lambda$. Thus,

$$E(X^2) = E(X(X-1)) + E(X) = \lambda^2 + \lambda.$$

Hence

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

The description of the Poisson distribution in terms of events in the time interval $[0,1]$ can naturally be restated in terms of distances, areas, volumes, or weight. For example, assuming infinitely small chocolate chips, given the average density of chips in a unit amount of dough, one can describe the probability that there are $k$ chocolate chips in a cookie using a unit amount of dough.

### 17.1 Reading

Continue reading [19, Chapter 4, pp. 119–188]. The current section is related to [19, §4.7]

### 17.2 Homework

[19, Chapter 4, pp. 178], 54, 58, 63, 64, 66.

### 18 Negative binomial distribution

Let $p$ be a real number with $0 < p < 1$ and consider independent trials with probability $p$ of success; as usual, we will write $q = 1 - p$ when convenient to do so. Let $r \geq 1$ be an integer; the random variable $X$ will assume the value $n$ if in $n$ independent trials $r$ successes are accumulated. Then $X$ can assume integer values $n \geq r$. Such a random variable will be said to have a negative binomial

\[^{18,1}\text{Different probability texts adopt slightly different definitions of the negative binomial distribution. We use the definition given in [19, Section 4.8.2 on p. 160].}\]
distribution with parameters \((r, p)\). The probability \(P(X = n)\) can be calculated as follows. In the first \(n - 1\) trials, \(r - 1\) successes need to be accumulated; the probability of this is

\[
\binom{n - 1}{r - 1} p^{r - 1} (1 - p)^{(n - 1) - (r - 1)} = \binom{n - 1}{r - 1} p^{r - 1} (1 - p)^{n - r}.
\]

Then the \(n\)th trial must be a success; this has probability \(p\). The product of these probabilities gives the probability we seek:

\[
(18.1) \quad P(X = n) = \binom{n - 1}{r - 1} p^r (1 - p)^{n - r}.
\]

We will show that these probabilities add up to 1, as required.

To this end, first note that the definition of the binomial coefficient \(\binom{\alpha}{k}\) given in formula (3.2) can be extended for any real (or even complex) \(\alpha\) and any integer \(k \geq 0\) by the same formula:

\[
(18.2) \quad \binom{\alpha}{k} = \frac{\prod_{j=0}^{k-1} (\alpha - j)}{k!} = \frac{\prod_{j=0}^{k-1} (\alpha - j)}{\prod_{j=0}^{k-1} (k - j)} = \frac{\prod_{j=0}^{k-1} \alpha - j}{k!}.
\]

Writing \(n = k + r\), where \(n \geq r\), and so \(k \geq 0\), we have

\[
P(X = n) = \binom{n - 1}{r - 1} p^r (1 - p)^{n - r} = p^r \binom{k + r - 1}{r - 1} (1 - p)^k.
\]

Using equations (3.4) and then using (3.2) twice, we have

\[
\binom{k + r - 1}{r - 1} = \binom{k + r - 1}{k} = \frac{1}{k!} \prod_{j=0}^{k-1} (k + r - 1 - j) = \frac{1}{k!} \prod_{j=0}^{k-1} (r + j) = (-1)^k \frac{1}{k!} \prod_{j=0}^{k-1} (-r - j) = (-1)^k \binom{-r}{k}.
\]

Substituting this into the previous equation, we obtain

\[
(18.3) \quad P(X = k + r) = p^r (-1)^k \binom{-r}{k} (1 - p)^k = p^r \binom{-r}{k} (p - 1)^k.
\]

Thus, adding up the probabilities \(P(X = n)\) for \(n \geq r\), we obtain

\[
\sum_{n=r}^{\infty} P(X = n) = \sum_{k=0}^{\infty} P(X = k + r) = \sum_{k=0}^{\infty} p^r \binom{-r}{k} (p - 1)^k
\]

\[
(18.4) \quad = p^r \sum_{k=0}^{\infty} \binom{r}{k} (p - 1)^k = p^r p^{-r} = 1;
\]

here, the fourth equation holds according to the Maclaurin expansion of \((1 + x)^{\alpha}\) with \(x = p - 1\) and \(\alpha = -r\). According to this,

\[
(18.5) \quad (1 + x)^{\alpha} = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k \quad (|x| < 1); \tag{18.5}
\]

this series is called the binomial series, convergent when \(|x| < 1\). \(^{18.2}\) This shows that the probabilities indeed add up to 1.

\(^{18.2}\) It is also convergent in other cases and the equation is valid for certain special values of the pair \((x, \alpha)\).
18.1 Moments of the negative binomial distribution.

Let $X$ be a negative binomial distribution with parameters $(r, p)$. Let $m \geq 0$ be an integer. Using formula (18.3), we have

$$E(X^m) = \sum_{n=r}^{\infty} n^m P(X = n) = \sum_{k=0}^{\infty} (k + r)^m P(X = k + r)$$

(18.6)

$$= \sum_{k=0}^{\infty} (k + r)^m p^r \binom{-r}{k} (p - 1)^k = p^r (p - 1)^{-r} f_m(p - 1),$$

where

$$f_m(x) = \sum_{k=0}^{\infty} (k + r)^m \binom{-r}{k} x^{k+r} \quad (|x| < 1).$$

We have

$$f_0(x) = x^r (1 + x)^{-r} \quad (|x| < 1)$$

according to formula (18.5). Further, it is easy to see that

$$f_{m+1}(x) = x \frac{d}{dx} f_m(x) \quad (|x| < 1).$$

This gives

$$f_1(x) = \frac{rx^r}{(1 + x)^{r+1}} \quad (|x| < 1),$$

$$f_2(x) = \frac{rx^r(r - x)}{(1 + x)^{r+2}} \quad (|x| < 1),$$

and

$$f_3(x) = \frac{rx^r((x-r)^2 - (r+1)x)}{(x+1)^{r+3}} \quad (|x| < 1).$$

18.2 Expectation and variance of the negative binomial distribution

Substituting the equations for $f_1(x)$ and $f_2(x)$ into formula (18.6), we obtain

$$E(X) = p^r (p - 1)^{-r} \frac{r(p - 1)^r}{p^{r+1}} = \frac{r}{p}$$

and

$$E(X^2) = p^r (p - 1)^{-r} \frac{r(p - 1)^r(r - p + 1)}{p^{r+2}} = \frac{r(r + 1 - p)}{p^2}.$$

Hence

$$V(X) = E(X^2) - (E(X))^2 = \frac{r(1 - p)}{p^2} = \frac{rq}{p^2}.$$
18.3 Alternative determination of the moments of the negative binomial distribution.

Let $X$ be a negative binomial distribution with parameters $(r, p)$. The identity

$$\frac{\alpha}{r} \left( \frac{\alpha - 1}{r - 1} \right) = \left( \frac{\alpha}{r} \right)$$

$(r \geq 1)$

can easily be verified by formula (18.2) for any real $\alpha$. Using this identity with $n$ replacing $\alpha$, for $k \geq 0$ formula (18.1) gives

$$E(X^k) = \sum_{n=r}^{\infty} n^k \left( \frac{n-1}{r-1} \right) p^{r-1} (1-p)^{n-r} = \sum_{n=r}^{\infty} r n^{k-1} \left( \frac{n}{r} \right) p^{r-1} (1-p)^{n-r}$$

(18.7)

$$= \frac{r}{p} \sum_{n=r}^{\infty} n^{k-1} \left( \frac{n}{r} \right) p^{r-1} (1-p)^{(n+1)-(r+1)} = \frac{r}{p} \sum_{n=r}^{\infty} n^{k-1} P(Y = n + 1)$$

$$= \frac{r}{p} E((Y-1)^{k-1}),$$

where $Y$ is a negative binomial distribution with parameters $(r+1, p)$. For $k = 1$ we have $n^k = 1$, and the sum before the last equation is 1 in view of equation (18.4), agreeing with the interpretation that $E(Y-1)^0 = E(1) = 1$. Formula (18.7) gives an alternative way to determine the moments of a negative binomial distribution.

18.4 The negative binomial variable as a sum of independent first variables

Let $X$ be a negative binomial variable with parameters $(r, p) \ (r \geq 1)$. Then, given independent trials with probability $p$ of success, $X = n$ indicates that $n$ trials are needed to accumulate $r$ successes. For $i = 1, 2, \ldots, r$, let $X_i$ be independent first (for the first time) random variables for these trials; that is, $X_i = k$ success is first achieved on the $k$th trial; see Subsection 15.6. We have

$$X = \sum_{i=1}^{r} X_i,$$

if $X_i$ is related to the trials starting after the $i-1$st success and ending with the $i$th success. The expected value and the variance of the variables $X_i$ were calculated in formulas (15.3) and (15.4) as $1/p$ and $q/p^2$, respectively. Given that the variables $X_i$ are independent, according to these results we have

$$E(X) = \sum_{i=1}^{r} E(X_i) = \frac{r}{p},$$

and

$$V(X) = \sum_{i=1}^{r} V(X_i) = \frac{rq}{p^2}.$$
18.6 Homework

[19] Chapter 4, pp. 178. [19] Chapter 4, pp. 182. Theoretical exercises 4, 5, 7, 10 (hint: first show that we have \( \binom{n}{k+1} \binom{n}{k} = \binom{n+1}{k+1} \)), 16, 36.

Hint for Problem 36. At each pick, one new ball is added, so at the \( n \)th pick there are \( n+1 \) balls to pick from. If no blue balls have been picked, i.e., if \( X \geq n \), then there are \( n \) red balls and one blue ball to pick from. Hence, the probability that a red ball is picked on the \( n \) pick, i.e., that \( X > n \), is

\[
P(X > n) = P(X > n \mid X \geq n) = \frac{n}{n+1}.
\]

You may also want to look at the Self-Test Problems and Exercises in [19] pp. 186–188.

19 Some continuous distributions

19.1 Uniform distribution

Given an interval \([a, b]\), a random variable \( X \) is said to be uniformly distributed on this interval if \( X \) assumes values only in this interval, and the probability of \( X \) assuming a value in a subinterval of this interval is proportional to the length of the interval. The density function of \( X \) can be described as follows:

\[
f_X(x) = \begin{cases} 
\frac{1}{b-a} & \text{if } a \leq x \leq b, \\
0 & \text{otherwise.}
\end{cases}
\]

As for the distribution function, for \( x \in [a, b] \) we have

\[
F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt = \int_{a}^{x} \frac{1}{b-a} \, dt = \frac{1}{b-a} \bigg\rvert_{t=a}^{x} = \frac{x-a}{b-a}.
\]

That is,

\[
F_X(x) = \begin{cases} 
0 & \text{if } -\infty < x < a, \\
(x-a)/(b-a) & \text{if } a \leq x \leq b, \\
1 & \text{if } b < x.
\end{cases}
\]

Note that \( f_X(x) = F'_X(x) \) for all \( x \) different from \( a \) and \( b \). For \( x = a \) and \( x = b \) \( F'_X(x) \) is not defined; at these points, \( f_X(x) \) could be assigned any value, and it will not make any difference; the choice we made above was convenient, but arbitrary.

As for the expectation of \( X \), we have

\[
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{a}^{b} x \frac{1}{b-a} \, dx = \frac{1}{2(b-a)} x^2 \bigg\rvert_{x=a}^{b} = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.
\]

As for the second moment, we have

\[
E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_{a}^{b} x^2 \frac{1}{b-a} \, dx = \frac{1}{3(b-a)} x^3 \bigg\rvert_{x=a}^{b} = \frac{b^3 - a^3}{3(b-a)}
\]

\[
= \frac{(b-a)(a^2 + ab + b^2)}{3(b-a)} = \frac{1}{3} (a^2 + ab + b^2).
\]

52
Hence, for the variance, we obtain
\[ V(X) = E(X^2) - (E(X))^2 = \frac{1}{3} (a^2 + ab + b^2) - \frac{1}{4} (a^2 + 2ab + b^2) \]
\[ = \frac{1}{12} (4a^2 + 4ab + 4b^2 - 3a^2 - 6ab - 3b^2) = \frac{1}{12} (a^2 - 2ab + b^2) = \frac{1}{12} (b - a)^2. \]

19.2 **Cauchy distribution**

Given a real number \( a > 0 \), the random variable \( X \) with density function
\[ f_X(x) = \frac{a}{\pi} \frac{1}{a^2 + x^2} \quad (x \in \mathbb{R}). \]
is said to have Cauchy distribution. As for the factor \( a/\pi \), it is there to make sure that
\[ \int_{-\infty}^{\infty} f_X(x) \, dx = 1. \]
Indeed, we have
\[ \int_{-\infty}^{\infty} \frac{1}{a^2 + x^2} \, dx = \frac{1}{a} \lim_{B \to -\infty} \int_{A}^{B} \frac{1}{1 + (x/a)^2} \, dx = \frac{1}{a} \lim_{B \to -\infty} \int_{A/a}^{B/a} \frac{1}{1 + t^2} \, dt \]
\[ = \frac{1}{a} \lim_{B \to -\infty} \left( \arctan \frac{B}{a} - \arctan \frac{A}{a} \right) = \frac{\pi}{a}. \]
The Cauchy distribution has important applications in physics, but our main interest in it is that it is an example for a distribution that has no expectation. Indeed, the integral expressing the expectation is
\[ E(X) = \int_{-\infty}^{\infty} x \frac{1}{a^2 + x^2} \, dx; \]
this integral, however, is divergent. This is easily shown by using the comparison test for the integral on the interval \([0, \infty)\). Furthermore, the integral can easily be calculated on finite intervals by making the substitution \( t = a^2 + x^2 \), when \( dt = 2x \, dx \), and a direct calculation shows that the integral on the interval \([a, b]\) diverges when we make \( a \to -\infty \) and \( b \to \infty \).

19.3 **Exponential distribution**

Assume a certain equipment is placed in service at time \( x = 0 \), and for any small \( \Delta x \) the probability of failure in the interval \([x, x + \Delta x]\) is approximately \( \lambda \Delta x \), assuming it has not failed before. The random variable \( X \) will have value \( x \) if the equipment fails at time \( x \). Clearly, \( x \geq 0 \), since the equipment was not in service before time 0. The random variable \( X \) so described is said to have an exponential distribution with parameter \( 1/\lambda \); in symbols, \( X \sim \text{Exp}(1/\lambda) \).

---

<sup>19.1</sup> When \( a = -b \), the integral on the interval \([a, b]\) is zero, but when taking the limit, \( a \) and \( b \) must vary independently.

<sup>19.2</sup> We used a closed interval here, but since the probability of failure at any single point of time is 0, it makes no difference whether the interval is open or closed. We need to say \( \Delta x \) is small, since the dependence on the length of the interval is not linear, so this relation can only hold in the limit when \( \Delta x \to 0 \).
19.3.1 The distribution function of the exponential distribution

Given $x > 0$, to find the probability of $X > x$, let $\Delta x > 0$ be small. We have

$$\frac{P(X > x + \Delta x)}{P(X > x)} = \frac{P(X > x + \Delta x \& X > x)}{P(X > x)} = \frac{P(X > x + \Delta x \mid X > x)}{P(X > x)} = 1 - \frac{P(X \leq x + \Delta x \mid X > x)}{P(X > x)} \approx 1 - \lambda \Delta x;$$

here the second equation holds in view of the definition of conditional probability in equation (6.1).

The fourth equation uses the fact that for any two events we have $P(A \mid B) = P(A \cap B \mid B)$, also an immediate consequence of equation (6.1). The approximate equation at the end holds in view of the description of the random variable $X$. Writing $G(x) = P(X > x)$, this equation can be written as

$$\frac{G(x + \Delta x) - G(x)}{\Delta x} = \frac{G(x + \Delta x)}{G(x)} - 1 \approx -\lambda \Delta x,$$

or else, as

$$\frac{G(x + \Delta x) - G(x)}{\Delta x} \approx -\lambda G(x).$$

Making $\Delta x \to 0$, this gives the equation

$$G'(x) = -\lambda G(x).$$

This is a differential equation; we can solve it as follows. Writing $y = G(x)$, this equation says $dy/dx = -\lambda y$, i.e.,

$$\frac{dy}{y} = -\lambda dx.$$

Integrating this, we obtain

$$\int \frac{dy}{y} = \int (-\lambda)dx,$$

that is,

$$\log |y| = -\lambda x + C.$$

for some constant $C$. The absolute value is not needed since $y$ denotes a probability, so $0 \leq y \leq 1$. This means that $y = e^{-\lambda x + C}$. That is,

$$P(X > x) = G(x) = e^{-\lambda x + C}.$$

Noting that we have $G(0) = P(X > 0) = 1$, this means that $e^C = 1$, and so $C = 0$. Thus, we have $G(x) = e^{-\lambda x}$. Therefore,

$$F_X(x) = P(X \leq x) = 1 - G(x) = 1 - e^{-\lambda x}.$$

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x \leq 0, \\ 1 - e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

19.3 We used the logic symbol $\&$ for “and” rather than the intersection symbol $\cup$, since the events are described as relations rather than as sets.

19.4 Here $\log u$ denotes the natural logarithm of $u$. In mathematics, one usually uses log rather than $\ln$, which was commonly used in calculus courses.
We have \( f_X(x) = F'_X(x) \), that is
\[
f_X(x) = \begin{cases} 
0 & \text{if } -\infty < x < 0, \\
\lambda e^{-\lambda x} & \text{if } x \geq 0.
\end{cases}
\]

At \( x = 0 \) the derivative \( F'_X(x) \) is not defined, and the value of \( f_X(0) \) is chosen arbitrarily; this will have no effect on any of the calculations.

### 19.3.2 Alternative derivation of the distribution function

Given \( x > 0 \), to find the probability of \( X > x \), let \( n \) be a large positive integer, and divide the interval \([0, x)\) into \( n \) parts
\[
I_i = \left( \frac{i}{n} - \frac{1}{n}, \frac{i}{n} \right) \quad (1 \leq i \leq n).
\]

For \( i \) with \( 1 \leq i \leq n \) we have
\[
P\left( X > \frac{ix}{n} \mid X \geq \frac{(i-1)x}{n} \right) = P\left( X \notin I_i \mid X \geq \frac{(i-1)x}{n} \right)
= 1 - P\left( X \in I_i \mid X \geq \frac{(i-1)x}{n} \right) \approx 1 - \frac{\lambda x}{n}.
\]

Thus,
\[
P\left( X > \frac{ix}{n} \right) = P\left( X > \frac{ix}{n} \right. \& \left. X \geq \frac{(i-1)x}{n} \right)
= P\left( X > \frac{ix}{n} \mid X \geq \frac{(i-1)x}{n} \right) P\left( X > \frac{(i-1)x}{n} \right) \approx \left( 1 - \frac{\lambda x}{n} \right) P\left( X > \frac{i-1}{n} x \right);
\]

The first equation holds because if \( X > i x / n \) then also \( X > (i-1) x / n \), so saying that both events occur just means that the first event occurs. As for the second equation, that holds in view of the definition of conditional probability in equation (6.1). Using these equations for \( i \) with \( 1 \leq i \leq n \), and noting that for \( i = 1 \) we have \( P(X > (i-1)x/n) = P(X > 0) = 1 \), we obtain
\[
P(X > x) \approx \left( 1 - \frac{\lambda x}{n} \right)^n.
\]

Making \( n \to \infty \), we obtain
\[
P(X > x) = \lim_{n \to \infty} \left( 1 - \frac{\lambda x}{n} \right)^n = e^{-\lambda x};
\]

see [14] for a discussion of the exponential function, where this limit is taken to be the definition of the exponential function. Thus, for \( x > 0 \), we have \( F_X(x) = P(X \leq x) = 1 - P(X > x) = 1 - e^{-\lambda x} \).

### 19.3.3 The expectation and the variance of the exponential distribution

To calculate the expectation of \( X \) is not difficult. We have
\[
E(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{0}^{\infty} \lambda x e^{-\lambda x} \, dx = \lim_{A \to \infty} \int_{0}^{A} \lambda x e^{-\lambda x} \, dx.
\]

[19,5] We used the logic symbol \& for “and” rather than the intersection symbol \( \cup \), since the events are described as relations rather than as sets.
Using integration by parts with \( f(x) = x \) and \( g'(x) = \lambda e^{-\lambda x} \) (cf. equation (10.3)), when we have \( f'(x) = 1 \), and \( g(x) = -e^{-\lambda x} \), we obtain

\[
E(X) = \lim_{A \to \infty} \left( -xe^{-\lambda x} \bigg|_{x=0}^{A} + \int_{0}^{A} e^{-\lambda x} \, dx \right) = \lim_{A \to \infty} \left( -Ae^{-\lambda A} - \frac{1}{\lambda} e^{-\lambda x} \bigg|_{x=0}^{A} \right)
\]

Noting that \( \lim_{A \to \infty} Ae^{-\lambda A} = 0 \) and \( \lim_{A \to \infty} e^{-\lambda A} = 0 \) we obtain that \( \lim_{A \to \infty} Ae^{-\lambda A} = \frac{1}{\lambda} \).

As for the second moment, we have

\[
E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_{0}^{\infty} \lambda x^2 e^{-\lambda x} \, dx = \lim_{A \to \infty} \int_{0}^{A} \lambda x^2 e^{-\lambda x} \, dx.
\]

Integrating by parts with \( f(x) = x^2 \) and \( g'(x) = \lambda e^{-\lambda x} \) (cf. equation (10.3)), when we have \( f'(x) = 2x \), and \( g(x) = -e^{-\lambda x} \), we arrive at

\[
E(X^2) = \lim_{A \to \infty} \left( -x^2 e^{-\lambda x} \bigg|_{x=0}^{A} + \int_{0}^{A} 2xe^{-\lambda x} \, dx \right) = \lim_{A \to \infty} A^2 e^{-\lambda A} + \int_{0}^{\infty} 2xe^{-\lambda x} \, dx
\]

The limit on the right-hand side is 0 and the integral is 2/\( \lambda^2 \) according to (19.1). Thus, \( E(X^2) = 2/\lambda^2 \). Hence,

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

### 20 The normal distribution

The normal distribution is the distribution most widely used in statistics. This is because of the Central Limit Theorem, to be discussed later, in Section 35.

#### 20.1 The standard normal distribution

The random variable \( X \) has standard normal distribution if its density function is

\[
f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.
\]

the scaling factor \( 1/\sqrt{2\pi} \) is there to ensure that \( \int_{-\infty}^{\infty} f_X(x) \, dx = 1 \). That this is indeed so, one can verify by using the formula

\[
\int_{-\infty}^{\infty} e^{-x^2} \, dx = \sqrt{\pi}.
\]

---

19.6 To establish the first limit, we have

\[
\lim_{A \to \infty} Ae^{-\lambda A} = \lim_{A \to \infty} \frac{A}{e^{\lambda A}} = \lim_{A \to \infty} \frac{1}{\lambda e^{\lambda A}} = 0,
\]

where we used l’Hospital’s rule to obtain the second equation.

19.7 Use l’Hospital’s rule twice for the limit \( \lim_{A \to \infty} A^2/e^{\lambda A} \).
This is easy to see, since
\[
\left( \int_{0}^{\infty} e^{-x^2} \, dx \right)^2 = \int_{0}^{\infty} e^{-x^2} \, dx \int_{0}^{\infty} e^{-y^2} \, dy = \int_{0}^{\infty} \int_{0}^{\infty} e^{-x^2} e^{-y^2} \, dy \, dx.
\]
\[
= \iint_{\{(x,y) : x \geq 0, y \geq 0\}} e^{-x^2-y^2} \, dy \, dx = \iiint_{\{(r,\theta) : r \geq 0, 0 \leq \theta \leq \pi/2\}} e^{-r^2} r \, dr \, d\theta.
\]
The last integral was obtained by transforming the double integral in Cartesian coordinates to polar coordinates. This last integral is easily evaluated by iterated integration; it equals
\[
\int_{0}^{\pi/2} \int_{0}^{\infty} e^{-r^2} r \, dr \, d\theta = \int_{0}^{\pi/2} \int_{0}^{\infty} e^{-\frac{1}{2} t} \frac{1}{2} d\theta = \int_{0}^{\pi/2} \frac{1}{2} d\theta = \frac{\pi}{4},
\]
the inner integral was evaluated by the substitution \( t = r^2 \). This implies that
(20.3) \[
\int_{-\infty}^{\infty} e^{-x^2} \, dx = 2 \int_{0}^{\infty} e^{-x^2} \, dx = 2 \sqrt{\frac{\pi}{4}} = \sqrt{\pi}.
\]

### 20.1.1 The expectation of the standard normal distribution

We have
\[
E(X) = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \int_{-\infty}^{0} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \, dx + \int_{0}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \, dx = 0;
\]
the last equation holds since the integrals after the second equation cancel each other, as can be seen by substituting \( t = -x \) in one of them \(^{[20.3]}\).

### 20.1.2 The variance of the standard normal distribution

For the variance, we have
\[
V(X) = E(X^2) = \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2} \, dx = 2 \int_{0}^{\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2} \, dx
\]
\[
= -\frac{2}{\sqrt{2\pi}} \lim_{A \to \infty} \int_{0}^{A} x \left( -xe^{-x^2/2} \right) dx.
\]
Using integration by parts with \( f(x) = x \) and \( g'(x) = -xe^{-x^2/2} \) (cf. equation \((10.3)\)), when we have \( f'(x) = 1 \), and \( g(x) = e^{-x^2/2} \), we obtain
\[
V(X) = E(X^2) = -\frac{2}{\sqrt{2\pi}} \lim_{A \to \infty} \left( xe^{-x^2/2} \right|_{x=0}^{A} - \int_{0}^{A} e^{-x^2/2} \, dx \right)
\]
\[
= -\frac{2}{\sqrt{2\pi}} \lim_{A \to \infty} \left( Ae^{-A^2/2} - \int_{0}^{A} e^{-x^2/2} \, dx \right) = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-x^2/2} \, dx
\]
\[
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx = \int_{-\infty}^{\infty} f_X(x) \, dx = 1;
\]
\(^{[20.1]}\)

We would caution against directly using the change of variable rule on improper integrals. While this would be possible to do, one needs to know the precise rules when this is allowed. It is usually preferable to rewrite the improper integral on an infinite integral as a limit of an integral on a finite interval. Note that when using the equation \( \int_{-\infty}^{\infty} \lim_{b \to \infty} \int_{a}^{b} \, dt \), \( a \) and \( b \) must vary independently; we definitely cannot assume that \( a = -b \). The integral can also be fully evaluated using the substitution \( t = -x^2/2 \) without splitting it up, but it is easier to point out the cancelation.
the fifth equation holds since the integral from 0 to ∞ is the same as the integral from −∞ to 0. The integrand after this equation is the same as the density function, and so the integral is 1.

20.1.3 The distribution function of the standard normal distribution

The distribution function of the standard normal distribution is often denoted by Φ:

\[ \Phi(x) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} \, dt. \]

In view of the importance of the standard normal distribution, most statistic books have tables of the function Φ; for example [3, p. 324].

20.2 The general normal distribution

Let \( Y \) be a random variable with standard normal distribution, and let \( \sigma > 0 \) and \( \mu \) be real numbers. Then the random variable \( X = \sigma Y + \mu \) is said to have general normal distribution denoted as \( \mathcal{N}(\mu, \sigma^2) \). We have \( \text{E}(X) = \mu \) and \( \text{V}(X) = \sigma^2 \) according to equations (13.1), (14.2), and (14.3).

20.2.1 The distribution function of the general normal distribution

We have

\[ F_X(x) = \text{P}(X \leq x) = \text{P}(\sigma Y + \mu \leq x) = \text{P} \left( Y \leq \frac{x - \mu}{\sigma} \right) = \Phi \left( \frac{x - \mu}{\sigma} \right) \]

\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2} \, dt = \lim_{A \to -\infty} \int_{A}^{(x-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt. \]

Making the change of variables \( u = \sigma t + \mu \), we have \( du = \sigma \, dt \), i.e., \( dt = (1/\sigma) \, dx \); further, we have \( t = (u - \mu)/\sigma \). As for the limits, for \( t = A \) we have \( u = \sigma A + \mu \), and for \( t = (x - \mu)/\sigma \), we have \( u = x \). Thus,

\[ F_X(x) = \lim_{A \to -\infty} \int_{A}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/(2\sigma^2)} \, \frac{1}{\sigma} \, du = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^{x} e^{-(u-\mu)^2/(2\sigma^2)} \, du. \]

20.2.2 The density function of the general normal distribution

For the density function of \( x \) we have

\[ f_X(x) = \frac{d}{dx} F_X(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-(x-\mu)^2/(2\sigma^2)}. \]

21 The gamma distribution

21.1 The gamma function

The gamma function is defined by the integral

\[ \Gamma(t) \overset{\text{def}}{=} \int_{0}^{\infty} x^{t-1} e^{-x} \, dx. \quad (t > 0); \]
This is an improper integral for two reasons. First, the upper limit is $+\infty$, second, for the values of $t$ with $0 < t < 1$ the integrand becomes infinite at $x = 0$. The easiest way to interpret this integral by splitting it up at, say $x = 1$ in case $0 < t < 1$:

$$
\Gamma(t) = \int_0^1 x^{t-1} e^{-x} \, dx + \int_1^\infty x^{t-1} e^{-x} \, dx = \lim_{\epsilon \to 0} \int_\epsilon^1 x^{t-1} e^{-x} \, dx + \lim_{A \to \infty} \int_1^A x^{t-1} e^{-x} \, dx
$$

here $\epsilon \searrow 0$ means that $\epsilon$ tends to zero from the right. The second integral on the right-hand side is convergent for all values of $t$. The first integral on the right-hand side is convergent for only $t > 0$. An important property of the gamma function is that

$$
(21.2) \quad \Gamma(t+1) = t\Gamma(t) \quad (t \neq 0, -1, -2, -3 \ldots).
$$

It is not hard to prove this formula for $t > 0$. Indeed, using the integration by parts formula

$$
(21.3) \quad \int_a^b f(x)g'(x) \, dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x) \, dx
$$

with $f(x) = e^{-x}$ and $g'(x) = tx^{t-1}$, when $f'(x) = -e^{-x}$ and $g(x) = x^t$. For small $\epsilon > 0$ and large positive $A$ we have

$$
\int_\epsilon^A e^{-x} tx^{t-1} \, dx = e^{-\epsilon}x^t|_x=\epsilon + \int_\epsilon^A e^{-x} x^t \, dt.
$$

When $\epsilon \searrow 0$, i.e. $\epsilon$ tends to $0$ from the right and $A \to \infty$, the left-hand side tends to $t\Gamma(t)$, the first term on the right tends to 0, and the second term tends to $\Gamma(t)$, verifying equation (21.2).

For $t = 1$, the integral in formula (21.1) is easily evaluated. For positive $A$, we have

$$
\Gamma(1) = \int_0^\infty e^{-x} \, dx = \lim_{A \to \infty} \int_0^A e^{-x} \, dx = \lim_{A \to \infty} \left(-e^{-x}\right)|_{x=0}^A = \lim_{A \to \infty} (-e^{-A} - (-e^0)) = 1.
$$

This shows that $\Gamma(1) = 1 = 0!$, and so, using equation (21.2) repeatedly we can see that for any $n \geq 1$ the equation $\Gamma(n+1) = n!$ holds. Thus, the gamma function can be considered an extension of the factorial for noninteger arguments.

### 21.2 The gamma distribution

The gamma distribution has two parameters. The continuous random variable $X$ with values $X \geq 0$ has $\Gamma(\alpha, \theta)$ distribution for positive real numbers $\alpha$ and $\theta$ if we have

$$
(21.4) \quad f_X(x) = \begin{cases} 
\frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta} & \text{if } x > 0, \\
0 & \text{if } x \leq 0.
\end{cases}
$$

21.1 The notation suggests that $\epsilon$ needs to decrease in order to reach 0.
21.2 This is not hard to show by using the comparison test for integrals, which is similar to the comparison test for series. For large values of $x$, the integrand is less than $e^{-x/2}$.
21.3 The comparison test used near 0 shows that the first integral is convergent if and only if

$$
\int_0^1 x^{t-1} \, dx
$$

is convergent, since $e^{-1} \leq e^{-x} < 1$ for $0 < x \leq 1$; this integral is convergent for $t > 0$, and it is divergent for $t \leq 0$.

The gamma function can be extended to all values of $t$, real or complex, with the exception of the values 0, $-1$, $-2$, $-4$, $\ldots$. If one considers complex values of $t$, then the integral is convergent whenever $\Re t > 0$. We will, however, not be concerned with complex values of $t$ in these notes.

21.4 In fact, defining $\Gamma(t)$ for all complex values of $t$ with $\Re t > 0$ by formula (21.1), one way to extend $\Gamma(t)$ for any other values in its domain by using formula (21.2).

21.5 We only need to take limit $\epsilon \searrow 0$ if $0 < t < 1$. For $t \geq 0$, we can simply take $\epsilon = 0$. 

59
Using the definition of the gamma function given in equation (21.1), we can see that \( \int_{1}^{\infty} f_X(x) \, dx = 1. \) The special case of \( \alpha = 1 \) gives the exponential distribution. See the Wikipedia page of the gamma distribution for graphs of the density function of the gamma distribution. Here \( \alpha \) is called the shape parameter, and \( \theta \) is called the scale parameter.

21.3 Sums of independent gamma distributions of the same scale

We have the following

**Theorem 21.1.** Let \( n \) be a positive integer, and let \( \theta > 0 \) and \( \alpha_i > 0 \) be real numbers for \( 1 \leq i \leq n \), and let \( X_i \) be independent random variables for \( 1 \leq i \leq n \), and assume that \( X_i \) has distribution \( \Gamma(\alpha_i, \theta) \). Then the random variable \( X = \sum_{i=1}^{n} X_i \) has \( \Gamma\left(\sum_{i=1}^{n} \alpha_i, \theta\right) \) distribution.

The proof of this theorem will be given in Subsection 26.2.

22 The Jacobian: change of variables in multiple integrals

22.1 Change of variables: revisiting the single-variable case

Before we embark on discussing change of variables in the multi-variable case, we want to revisit the single variable case. Suppose that in the integral

\[ \int_{A}^{B} f(x) \, dx \]

we want to replace \( x \) with \( h(t) \) for some increasing function \( h \). The formula as to how to do this described in formula (10.4) in terms of Stieltjes integrals with \( g(x) = x \). Then Theorem stieltjes: dg when \( g \) exists shows the way to rewrite the resulting Stieltjes integral. If one follows the calculation involved, one can see a proof of the change of variable formula that is more enlightening than the usual derivation of the change of variable formula as a reversal of the chain rule of differentiation. What is clear from this calculation is that, approximating the integral integral

\[ \int_{a}^{b} f(h(x))h'(x) \, dx = \int_{a}^{b} f(h(x)) \, dh(x) = \int_{h(a)}^{h(b)} f(t) \, dt \]

Here the first equation holds according to (10.2), and the second one holds in view of (10.4). This shows that in view of the change of variable formula \( h'(x) \, dx \) corresponds to \( dt \).

22.2 Change of variables in double integrals

Consider the double integral

\[ I = \iint_{R} f(x, y) \, dx \, dy \]

21.6 The meaning of the parameters is that by changing the shape parameter, the shape of the density function changes, while the change of the scale parameter essentially spreads out the same shape by stretching it in the \( x \) direction, and contracting it in the \( y \) direction as \( \theta \) increases.

22.1 This is more than formal correspondence, as can be seen by following the proofs of the quoted formulas, which gives a proof of the change of variable formula without the chain rule of differentiation. A direct proof of the change of variable formula is given along these lines in [13 §30, p. 66–69]. In the proof, it is not assumed that \( h(x) \) is monotonic.
on a region \( R \) in the plane. Let \( S \) be a region in the plane and let \( T : S \to R \) be a one-to-one mapping of \( S \) onto \( R \). We can describe the mapping \( T \) by a pair of functions \( x = \phi(u,v) \) and \( y = \psi(u,v) \). We want to describe the integral \( I \) in terms of an integral of an area of \( S \) such as

\[
I = \int_S f(\phi(u,v),\psi(u,v)) \, du \, dv;
\]

The question mark indicate that we need to figure out what to replace the area element \( dx \, dy \) with. The integral \( I \) written in terms of \( u \) and \( v \) is is approximated by sums of terms

\[
f(\phi(\alpha_{ij},\beta_{ij}),\psi(\alpha_{ij},\beta_{ij}))(u_{i-1} - u_i)(v_{j-1} - v_j) = f(\phi(\alpha,\beta),\psi(\alpha,\beta)) \Delta u \Delta v
\]

where we have \( u_{i-1} \leq \alpha_{ij} \leq u_i \) and \( v_{j-1} \leq \beta_{ij} \leq v_j \); on the right-hand side, we dropped the subscripts for the sake of simplicity. Now the rectangle with vertices \((u,v), (u+\Delta u,v), (u,v+\Delta v), (u+\Delta u,v+\Delta v)\), is mapped onto a near parallelogram in the \((x,y)\) plane by the mapping \( T = (\phi,\psi) \). The factor \( \Delta u \Delta v \) needs to be so chosen that the area \( \Delta u \Delta v \) is (nearly) equal to the area of this image in the \((x,y)\) plane. Assume that \( \phi \) and \( \psi \) are differentiable functions, i.e., that their total differentials exist. Then

\[
dx = \phi_u \, du + \phi_v \, dv \quad \text{and} \quad dy = \psi_u \, du + \psi_v \, dv.
\]

So, writing \( x = \phi(u,v) \) and \( y = \psi(u,v) \) for the specific \( u \) and \( v \), \( T(u,v) = (x,y) \), \( T(u+\Delta u,v) = (x + \phi_u \Delta u, y + \psi_u \Delta u) \), \( T(u,v+\Delta v) = (x + \phi_v \Delta v, y + \psi_v \Delta v) \), \( T(u+\Delta u,v+\Delta v) = (x + \phi_u \Delta u + \psi_v \Delta v, y + \phi_v \Delta v + \psi_u \Delta v) \).

Thus, the sides of this image parallelogram can be described as the vectors \( r_u = \phi_u \Delta u \mathbf{i} + \psi_u \Delta u \mathbf{j} \) and \( r_v = \psi_v \Delta v \mathbf{i} + \phi_v \Delta v \mathbf{j} \). We have \( r_u \times r_v = \det J \Delta u \Delta v \mathbf{k} \), where \( \det J \) is the determinant of the Jacobian matrix of \((x,y)\) with respect to \((u,v)\); \( \det J \) stands for the determinant of the matrix \( J \). To approximate the area of the image parallelogram, we need to take the length of the vector \( |r_u \times r_v| = |\det J| \Delta u \Delta v \). That is, in the above formula, \( |\det J| \) will be the factor \( \Delta u \Delta v \) that we wanted to determine. That is, we have

\[
\iint_R f(x,y) \, dx \, dy = \iint_S f(\phi(u,v),\psi(u,v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right| \, du \, dv.
\]

A similar formula can be written for higher dimensional integrals. For more about the Jacobian, see Subsection 25.2 in [15] p. 101.

The fact that we need to take the absolute value of the Jacobian determinant in the above formula is a sign of trouble. In a theory of integration that is consistent with coordinate transormation the fiction that the area is always positive cannot be maintained, since one expect that in a linear coordinate transfromation, areas and volumes transform as multilinear (i.e., linear in each variable) functions. In a consistent theory, instead of \( dx \, dy \) one write \( dx \wedge dy \), with the stipulation that \( dx \wedge dy = -dy \wedge dx \). In such a theory, areas and volumes need to be anti-symmetric tensors, described in terms of differential forms. In the theory, integration by parts, Greene’s Theorem, Stokes’s Theorem, and the Divergence Theorem receives a unified treatment. While this is above the level of the present course, the ambitious student may successfully consult the book [1] for the subject.

22.2 Named after Carl Gustav Jacob Jacobi. Some authors use \( \partial(u,v)/\partial(x,y) \) to denote the determinant of the Jacobian matrix, and simply call it Jacobian.
23 The beta distribution

The random variable \( X \) with parameters \( a, b > 0 \) has values in the interval \((0, 1)\) and its density function is

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

where

\[
B(a,b) \overset{\text{def}}{=} \int_0^1 x^{a-1}(1-x)^{b-1} \, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)};
\]

the second equation here needs a proof, which we are about to give.

Proof of second equation in (23.2). We have

\[
\Gamma(a)\Gamma(b) = \int_0^\infty u^{a-1}e^{-u} \, du \int_0^\infty v^{b-1}e^{-v} \, dv = \iint_S u^{a-1}v^{b-1}e^{-(u+v)} \, du \, dv,
\]

where the region \( S \) is given as

\[
S = \{(u,v): u > 0, \quad v > 0\}.
\]

We define the mapping \( \phi(u,v) = (x,y) \) by \( x = u/(u+v) \) and \( y = u+v \); this maps \( S \) onto the region

\[
T = \{(x,y): 0 < x < 1, \quad y > 0\}.
\]

This mapping is one-to-one, and we have \( u = xy \) and \( v = y - xy \). Writing \( \text{det} \, A \) for the determinant of the square matrix \( A \), the Jacobian determinant of this mapping is

\[
\text{det} \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial v}{\partial x} \\ \frac{\partial u}{\partial y} & \frac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} y & -y \\ x & 1-x \end{vmatrix} = y(1-x) + xy = y.
\]

Thus, rewriting the double integral above in terms of the variables \( x \) and \( y \), we have

\[
\Gamma(a)\Gamma(b) = \iint_T (xy)^{a-1}(y-x)^{b-1}e^{-y} \, dx \, dy
\]

\[
= \int_0^1 x^{a-1}(1-x)^{b-1} \, dx \int_0^\infty y^{a+b-1}e^{-y} \, dy = B(a,b)\Gamma(a+b),
\]

establishing the second equation in (23.2). \( \square \)

24 The chi squared and the lognormal distributions

Chi refers to the Greek letter \( \chi \)\(^{24.1}\). As we will show below, in Subsection 24.1 if \( X \) has \( \mathcal{N}(0,1) \), i.e., standard normal, distribution, then \( X^2 \) has \( \Gamma(1/2,2) \) distribution; this distribution is also called \( \chi^2(1) \) distribution, that is, chi squared distribution with degree of freedom 1. If \( X_i \) are independent standard normal variables, then the distribution of \( \sum_{i=1}^n X_i^2 \) is called \( \chi^2(n) \) distribution, that is, the chi squared distribution of degree of freedom \( n \). According to Theorem 21.1 this distribution,

\(^{24.1}\)This is the lower case chi. The upper case letter looks the same as the capital X.
being the sum of \( n \) independent random variables each having \( \Gamma(1/2, 2) \) distribution, is identical to the \( \Gamma(n/2, 2) \) distribution.

In using Theorem 24.1 we made use of the result that if the random variables \( X_1, X_2, \ldots, X_n \) are independent, and \( f_1, f_2, \ldots, f_n \) are nice real-valued functions on the set of real numbers, then the random variables \( f_1(X_1), f_2(X_2), f_3(X_3), \ldots, f_n(X_n) \) are also independent.\(^{24.2}\) We only used this in the case \( f_i(x) = x^2 \) for all \( i \), but the more general statement is also true.

### 24.1 The square of the standard normal distribution

Let \( X \) be a standard normal variable, i.e., a variable having distribution \( \mathcal{N}(0, 1) \). We will calculate the density function of \( Y = X^2 \). Recall that the density function of \( X \) is

\[
f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.
\]

Next we will calculate the distribution function of \( Y = X^2 \). For an arbitrary \( t \geq 0 \) we have

\[
F_Y(t) = P(Y \leq t) = P(X^2 \leq t) = P(-\sqrt{t} \leq X \leq \sqrt{t})
\]

\[
= P(-\sqrt{t} \leq X \leq 0) + P(0 \leq X \leq \sqrt{t}) = 2P(0 \leq X \leq \sqrt{t});
\]

the last equation holds since the density function of \( X \) is even, i.e., \( f_X(t) = f_X(-t) \). Thus,

\[
F_Y(t) = 2 \int_0^\sqrt{t} f_X(x) \, dx = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{t}} e^{-x^2/2} \, dx = \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^u e^{-x^2/2} \, dx,
\]

where \( u = \sqrt{t} \).

For the density function of \( Y \) we have

\[
f_Y(t) = \frac{d}{dt} F_Y(t) = \frac{du}{dt} \frac{du}{du} F_Y(u^2) = \frac{d\sqrt{t}}{dt} \frac{d}{du} \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^u e^{-x^2/2} \, dx,
\]

where in the second equations we used that \( t = u^2 \), and we also used the chain rule of differentiation. Noting that, according to the Fundamental Theorem of Calculus, the derivative of an integral with respect to its upper limit is the integrand at the upper limit, we obtain

\[
f_Y(t) = \frac{1}{2\sqrt{t} \sqrt{\pi}} e^{-u^2/2} = \frac{1}{\sqrt{\pi} 2^{1/2}} t^{1/2} e^{-t/2} = \frac{1}{\sqrt{\pi} 2^{1/2}} t^{1/2-1} e^{-t/2};
\]

we wrote the right-hand side in a form where one can recognize that this density is identical to the density function of the \( \Gamma(1/2, 2) \) distribution given in formula (21.4) with \( \alpha = \theta = 1/2 \), except for the value of \( \Gamma(1/2) \).

Since the constant in a density function must be so chosen that the integral of the density function on \( (-\infty, \infty) \) is 1, the constants in the two density functions must be equal; so we can conclude that \( \Gamma(1/2) = \sqrt{\pi} \) and that the square \( Y \) of a standard normal distribution has \( \Gamma(1/2, 2) \) distribution. This distribution is also called the \( \chi^2(1) \) distribution.

\(^{24.2}\) The function \( f_i \) needs to be Borel measurable to ensure that \( f(X_i) \) is a random variable; however, a definition of Borel measurable is beyond the scope of this course. Suffice it to say that assuming the functions \( f_i \) are continuous is certainly sufficient.
24.2 The lognormal distribution

The argument used in finding the density function of the square of the normal distribution can be extended to the case when we want to find the density function of $g(X)$ of a random variable $X$ in case the function $g$ consists of a few monotonic pieces; for example, the function $g$ with $g(x) = x^2$ for all real $x$ consists of two monotonic pieces: decreasing for $x < 0$ and increasing for $x > 0$. The method is not usable in practice in the general case. There are so many different situations that it is hardly worth formulating a general result; instead, we only give an example of practical importance.

Assume $X$ is a normal variable, and let $Y = e^X$. The variable $Y$ is said to have lognormal distribution. If $S_0$ is a price of a certain stock at the beginning of a time period, and $S_1$ is its price at the end, then $S_1/S_0$ is usually considered to have lognormal distribution. The reason is that when an investor wants to decide how much to invest in a stock, he or she is interested in the proportional increase of the value of the stock, rather than the increase in the price of a single stock; that is, what matters is $S_1/S_0$ rather than $S_1 - S_0$.

Assuming that $X$ has distribution $\mathcal{N}(\mu, \sigma^2)$ and $Y = e^X$, for any $y > 0$ we have \(^{24,3}\) As for the distribution function $F_Y$ of $Y$, we have

$$F_Y(y) = P(Y \leq y) = P(e^X \leq y) = P(X \leq \log y) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\log y} e^{-\left(t-\mu\right)^2/(2\sigma^2)} dt.$$  

Differentiating this with respect to $y$, we will obtain the density function $f_Y(y)$ of $Y$. In using the chain rule of differentiation, we will write $x = \log y$. We have

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{dx}{dy} \frac{d}{dx} \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-\left(t-\mu\right)^2/(2\sigma^2)} dt$$

$$= \frac{1}{y} e^{-\left(x-\mu\right)^2/(2\sigma^2)} \frac{1}{\sqrt{2\pi\sigma y}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right) \quad (y > 0),$$

where $\exp x = e^x$. \(^{24,4}\) For $y \leq 0$, we have $f_Y(y) = 0$.

24.3 Reading

[19] Chapter 5 except Subsections 5.6.2 and 5.6.5, pp. 189–236.

24.4 Homework

[19] Chapter 5, pp. 228], Problems 1, 2, 3, 4, 6, 8, 11, 15, 17, 32, 33, 35, 39, 42.

(if you do not use mathematical software to do this, you need a table of the standard normal variable for some problems, easily found online).


\(^{24,3}\)Here $\log u$ denotes the natural logarithm of $u$. In mathematics, one usually uses $\log$ rather that $\ln$, which was commonly used in calculus courses.

\(^{24,4}\)The symbol $\exp x$ or $\exp(x)$ is common in mathematics to indicate $e^x$. It is especially useful when the exponent is a complicated expression, so one does not have to write a long exponent in tiny letters.
25 Joint distribution functions

Given two random variables $X$ and $Y$, their joint distribution function is defined as

\[(25.1) \quad F_{X,Y}(x,y) = P\{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\} = P(X \leq x, Y \leq y),\]

where the expression on the right is the commonly used more concise notation. In case $X$ and $Y$ are discrete random variables assuming the values values $x_i$ and $y_j$, one usually prefers to use the joint probability, defined as

\[(25.2) \quad p_{X,Y}(x,y) = P\{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\} = P(X = x, Y = y),\]

instead of the joint distribution function. Clearly, this function is 0 unless $X$ assumes the value $x$ and $Y$ assumes the value $y$; even then, it might still be 0. It is possible to express the joint probability function with the aid of the joint distribution function; how to do this is not of immediate interest to us.

25.1 Marginal distributions

Given the joint distribution function $F_{X,Y}$ of the random variables $X$ and $Y$, the distribution functions $F_X$ and $F_Y$ of $X$ and $Y$ can be recovered as follow:

\[(25.3) \quad F_X(x) = \lim_{y \to \infty} F_{X,Y}(x,y) \quad \text{and} \quad F_Y(x) = \lim_{x \to \infty} F_{X,Y}(x,y).\]

To verify the first of these equations, note that

\[
\{\omega \in \Omega : X(\omega) \leq x\} = \bigcup_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq n\} \\
= \{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq 1\} \\
\cup \bigcup_{n=1}^{\infty} \{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq n + 1\} \setminus \{\omega \in \Omega : X(\omega) \leq x \text{ and } Y(\omega) \leq n\}.
\]

The sets on the right-hand side are pairwise disjoint, so their probabilities add up. This leads to the equation

\[
F_X(x) = F_{X,Y}(x,1) + \sum_{n=1}^{\infty} (F_{X,Y}(x,n+1) - F_{X,Y}(x,n)) \\
= F_{X,Y}(x,1) + \lim_{N \to \infty} \sum_{n=1}^{N} (F_{X,Y}(x,n+1) - F_{X,Y}(x,n)) \\
= F_{X,Y}(x,1) + \lim_{N \to \infty} (F_{X,Y}(x,N+1) - F_{X,Y}(x,1)) \\
= \lim_{N \to \infty} F_{X,Y}(x,N+1) = \lim_{y \to \infty} F_{X,Y}(x,y);
\]

here the second equation amounts to saying that the sum of a series equals the limit of its partial sums, and the last equation holds since $(F_{X,Y}(x,y))$ for a fixed $x$ is a nondecreasing function of $y$. This verifies the first equation in (25.3). The second equation can be verified similarly.
25.2 Probability of a rectangle

Assume $X$ and $Y$ are random variables with joint distribution function $F_{X,Y}$. Given real numbers $a$, $b$, $c$, and $d$, with $a < b$ and $c < d$, we have

\[ P(a < X \leq b, c < Y \leq d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c) \]

To see this, note that

\[ [X \leq b, Y \leq d] \setminus [(a < X \leq b, c < Y \leq d)] = [X \leq a, Y \leq d] \cup [X \leq b, Y \leq c]. \]

The events on the right-hand side are not disjoint, so we need formula (4.2) in Lemma 4.1 to calculate the probability on the right-hand side. Noting that

\[ [X \leq a, Y \leq d] \cap [X \leq b, Y \leq c] = [X \leq a, Y \leq c], \]

this gives

\[ P([X \leq b, Y \leq d] \setminus [(a < X \leq b, c < Y \leq d)]) = F_{X,Y}(a,d) + F_{X,Y}(b,c) - F_{X,Y}(a,c) \]

Since

\[ [X \leq b, Y \leq d] = ([X \leq b, Y \leq d] \setminus [(a < X \leq b, c < Y \leq d)]) \cup [(a < X \leq b, c < Y \leq d)] \]

is a disjoint union, their probabilities add up, so

\[ F_{X,Y}(b,d) = P([X \leq b, Y \leq d] \setminus [(a < X \leq b, c < Y \leq d)]) + P([(a < X \leq b, c < Y \leq d)]). \]

This equation together with (25.5) gives equation (25.4).

Using arguments similar to the ones used in Subsection 25.1, we can show that

\[ P(a \leq X \leq b, c \leq Y \leq d) = \lim_{x \rightarrow a \atop y \rightarrow c} P(x \leq X \leq b, y \leq Y \leq d). \]

This allows us to express the joint probability function of two discrete variables given in equation (25.2) in terms of their joint distribution functions, since

\[ p_{X,Y}(x,y) = P(\{\omega \in \Omega : X(\omega) = P(X = x, Y = y) = P(x \leq X \leq x, y \leq Y \leq y). \]

We omit the details.

25.3 Joint density functions

If the mixed second derivatives of the joint distribution function $F_{X,Y}$ of the random variables $X$ and $Y$ exist, we can define their joint density function as

\[ f_{X,Y}(x,y) \overset{\text{def}}{=} \frac{\partial}{\partial x} \frac{\partial}{\partial y} F_{X,Y}(x,y). \]

The assumption of differentiability is not quite enough, and the comment made after the definition of the density function in formula (12.4) applies also here. For the joint density function to be of any use, we need to have

\[ F_{X,Y}(x,y) = \int_{\{(u,v): u \leq x \text{ and } v \leq y\}} f_{X,Y}(s,t) \, ds \, dt. \]
This is certainly true if the right-hand side of equation (25.6) is continuous everywhere, but this is not a necessary condition.

25.4 Sums of two random variables

Assume $X$ and $Y$ are two random variables that have a joint density function $f_{X,Y}$. We can calculate the distribution function $X + Y$ as follows.

$$F_{X+Y}(t) = P\{\omega : X(\omega) + Y(\omega) \leq t\} = \int_{\{(u,v): u+v \leq t\}} f_{X,Y}(x,y) \, dx \, dy,$$

(25.8)

$$= \int_{-\infty}^{\infty} dx \int_{-\infty}^{t-x} f_{X,Y}(x,y) \, dy,$$

where in the last step we used successive integration to evaluate the double integral.

Differentiating this equation with respect to $t$ allows us to find the density function of $X + Y$.

We have

$$f_{X+Y}(t) = \frac{d}{dt} \int_{-\infty}^{\infty} dx \int_{-\infty}^{t-x} f_{X,Y}(x,y) \, dy = \int_{-\infty}^{\infty} dx \frac{\partial}{\partial t} \int_{-\infty}^{t-x} f_{X,Y}(x,y) \, dy,$$

(25.9)

$$= \int_{-\infty}^{\infty} f_{X,Y}(x,t-x) \, dx,$$

where the second equation assumes that the integration and the differentiation are interchangeable.

The question whether the interchange of the integration and the differentiation in this calculation is correct is, however, moot, since one can show that the density function $f_{X+Y}$ obtained by the calculation is the correct density function of $X + Y$ if $f_{X,Y}$ is the correct joint density function of $X$ and $Y$. Indeed, for the putative density function obtained we have

$$\int_{-\infty}^{u} f_{X,Y}(t) \, dt = \int_{-\infty}^{u} dt \int_{-\infty}^{\infty} f_{X,Y}(s,t-s) \, ds$$

$$= \int \int_{\{(s,t): t \leq u\}} f_{X,Y}(s,t-s) \, ds \, dt.$$

Using the mapping $\phi(s,t) = (x,y)$ with $x = s$ and $y = t-x$, this maps the region

$$S = \{(s,t) : t \leq u\}$$

25.1 This condition also guarantees that the mixed derivative $\frac{\partial}{\partial y} \frac{\partial}{\partial x} F_{X,Y}$ also exists and the two mixed partial derivatives are equal; see [11, Theorem 2, p. 2].

25.2 Similarly to the situation concerning the density function of a single random variable, from the viewpoint of modern integration theory equation (25.7) has primacy over equation (25.6) in defining the joint density function. See footnote [22.1] on page 67.

25.3 Successive integration is allowed since the integrand is nonnegative and its double integral on the whole plane exists (in fact, it is 1, since the integrand is a joint density function); this is true according to Lebesgue integration theory, but a discussion of this is beyond the level of this course.

25.4 Whether this is allowed is a somewhat delicate question in view of the infinite interval of integration, and it is beyond the level of this course. One way of justifying this interchange would be via the Lebesgue bounded convergence theorem, which would require that for a fixed $t$ the difference quotients corresponding to the partial differentiation are uniformly bounded by a Lebesgue integrable function of $f$ (note that these difference quotients are all nonnegative, since a density function is always nonnegative.)
onto the region

\[ T = \{(x, y) : x + y \leq u \}. \]

This mapping is one-to-one, with \( s = x \) and \( t = x + y \). The Jacobian determinant of this mapping is

\[
\det \frac{\partial(s, t)}{\partial(x, y)} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1.
\]

Thus, rewriting the double integral in terms of the variable \( s \) and \( y \), we obtain

\[
\int_{-\infty}^{u} f_{X,Y}(t) \, dt = \iint_T f_{X,Y}(x, y) \, dx \, dy = P(X + Y \leq u);
\]

the last equation holds since \( f_{X,Y} \) was assumed to be the joint density function of \( X \) and \( Y \). The equality of the sides shows that the function \( f_{X+Y} \) given by equation (25.9) is the indeed the density function of \( X + Y \), regardless whether the interchange of the integration and the differentiation used in this equation can be justified rigorously.

26 Independent random variables. Sums of continuous variables

Independent random variables were defined in Subsection 13.3. Assuming \( X \) and \( Y \) are independent, we have

\[
F_{X,Y}(x, y) = P(X \in (-\infty, x] \text{ and } Y \in (-\infty, y])
= P(X \in (-\infty, x]) \cdot P(Y \in (-\infty, y]) = F_X(x)F_Y(y),
\]

where the second equation holds according to the definition of independence given in Subsection 13.3. As we remarked after that definition, in order that \( X \) and \( Y \) be independent, we may restrict the assumption described there to intervals of type \( I_X = (-\infty, x] \) and \( I_Y = (-\infty, y] \). This means that the validity of this equation guarantees that \( X \) and \( Y \) are independent; that is, the equality of the sides for all \( x, y \in \mathbb{R} \) in this equation is a necessary and sufficient condition for \( X \) and \( Y \) to be independent.

If \( X \) and \( Y \) are nice continuous random variables so that they have density functions, differentiating the above equation with respect to \( x \) and \( y \), we obtain

\[
f_{X,Y}(x, y) = f_X(x)f_Y(y),
\]

26.1 Sum of two independent continuous random variables; convolution

If \( X \) and \( Y \) are independent random variables, each having a density function, using equations (25.9) and (26.2) we have

\[
f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t - x) \, dx.
\]

The integral on the right is called the convolution of the functions \( f_X \) and \( f_Y \), and it is usually denoted by the symbol \( f_X * f_Y \); that is, \( f_{X+Y} = f_X * f_Y \).

25.5 Much of the lesson shown by pursuing these calculations can be appreciated only in the light of Lebesgue integration theory. Namely, that integration is much easier to handle technically than differentiation; that is, there are no serious difficulties involved in handling multiple integrals, the issues of interchanging integration and differentiation are more delicate. All this underlines the importance of the viewpoint, made in footnote 22 on p. 37 that the density function should be considered as a Radon-Nikodym derivative, not a pointwise derivative.
26.2 Sums of independent gamma variables

At this point, we are in a position to give a proof of Theorem 21.4. It is sufficient to give the proof in case of two gamma variables. So, assume that $X$ has distribution $\Gamma(\alpha, \theta)$ and $Y$ has distribution $\Gamma(\beta, \theta)$, where $\alpha$, $\beta$, and $\theta$ are positive real numbers; assume, further, that $X$ and $Y$ are independent. Let $Z = X + Y$. Given that $X \geq 0$ and $Y \geq 0$, we have $Z \geq 0$ Using equation (26.3), for $t \geq 0$ we have

$$f_Z(t) = \int_{-\infty}^{\infty} f_X(x)f_Y(t-x) \, dx = \int_0^t \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1}e^{-x/\theta} \frac{1}{\Gamma(\beta)\theta^\beta} (t-x)^{\beta-1}e^{-(t-x)/\theta} \, dx.$$  

On the right-hand side, we substituted the density functions of $f_X$ and $f_Y$ as given by formula (21.4); the limits of the integral explained by noting that the integrand is 0 outside the interval $(0, t)$, since $f_X(x) = 0$ is $x \leq 0$ and $f_Y(t-x) = 0$ is $t-x \leq 0$, i.e., when $x \geq t$.

Simplifying the integral on the right-hand side, we obtain

$$f_X(t) = \frac{e^{-t/\theta}}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} \int_0^t x^{\alpha-1}(t-x)^{\beta-1} \, dt.$$  

Using the change of variable $x = tu$ in the integral for fixed $t$, when $dx = t \, du$, we obtain

$$f_X(t) = \frac{e^{-t/\theta}t^{(\alpha-1)+(\beta-1)+1}}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} \int_0^1 u^{\alpha-1}(1-u)^{\beta-1} \, du = \frac{\Gamma(\alpha+\beta-1)e^{-t/\theta}}{\Gamma(\alpha+\beta)\theta^{\alpha+\beta}},$$  

where the last equation holds in view of formula (23.2). According to formula (21.4), the function on the right-hand side is the density function of the $\Gamma(\alpha + \beta, \theta)$ distribution (note that we assumed $t \geq 0$ throughout these calculations). This completes the proof 26.1.

26.3 Sums of independent normal variables

We have the following lemma.

**Lemma 26.1.** The sum of two independent normal random variables is normal.

**Proof.** Let $X$ and $Y$ be independent normal variables; in proving that $X + Y$ is normal, we may assume that $X$ is an $N(0, 1)$ and $Y$ is an $N(0, \sigma^2)$ variable for some $\sigma > 0$. Indeed, assume that $U$ and $V$ are arbitrary independent random variables with distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, where $\sigma_1, \sigma_2 > 0$. Then writing $X = (U - \mu_1)/\sigma_1$ and $Y = (V - \mu_2)/\sigma_1$, then $X$ is an $N(0, 1)$ variable and $Y$ is an $N(0, \sigma^2)$ with $\sigma = \sigma_2/\sigma_1$. We have

$$U + V = \sigma_1(X + Y) + \mu_1 + \mu_2.$$  

If we show that $X + Y$ is normal, it will follow that $U + V$ is also normal according to Subsection 20.2.

So, assume $X$ and $Y$ are independent with distributions $N(0, 1)$ and $N(0, \sigma^2)$, respectively; that is,

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \quad \text{and} \quad f_Y(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-x^2/(2\sigma^2)}$$

26.1 Actually, we did not need to use formula (23.2) in the proof, since the integral of a density function must be 1. For this reason, if everything else in the expression on the right-hand side matches with the density function of the $\Gamma(\alpha + \beta, \theta)$ distribution, the constants must also match. This is possible only if second equation in (23.2) holds. Our arguments here therefore supply a new proof of this equation.
according to (20.1) and (20.5). Hence, using equation (26.3), we have

\[ f_X(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x) \, dx = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left( -\frac{x^2}{2} - \frac{(x-t)^2}{2\sigma^2} \right) \, dx \]

\[ = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left( -\sigma^2 x^2 + x^2 - 2xt + t^2 \right) \, dx \]

\[ = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left( -\frac{1}{2\sigma^2} \left((\sigma^2 + 1) \left( x^2 - 2x \frac{t}{\sigma^2 + 1} \right) + t^2 \right) \right) \, dx \]

Completing the square on the right-hand side

\[ f_X(t) = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left( -\frac{1}{2\sigma^2} \left( (\sigma^2 + 1) \left( x - t \frac{1}{\sigma^2 + 1} \right)^2 + t^2 \left( 1 - \frac{1}{\sigma^2 + 1} \right) \right) \right) \, dx \]

\[ = \frac{1}{2\pi\sigma} \int_{-\infty}^{\infty} \exp\left( -\frac{1}{2\sigma^2} \left( \left( x - t \frac{1}{\sigma^2 + 1} \right)^2 - t^2 \frac{2\sigma^2}{(\sigma^2 + 1)} \right) \right) \, dx \]

The last term in the exponent can be moved before the integral sign, and then we can substitute \( u = x - t/(\sigma^2 + 1) \) for \( x \) in the integral:

\[ f_X(t) = \frac{1}{2\pi\sigma} \exp\left( -\frac{t^2}{2(\sigma^2 + 1)} \right) \int_{-\infty}^{\infty} \exp\left( -\frac{u^2}{2\sigma^2/\left(\sigma^2 + 1\right)} \right) \, du \]

\[ = \frac{1}{\sqrt{2\pi\sqrt{\sigma^2 + 1}}} \exp\left( -\frac{t^2}{2(\sigma^2 + 1)} \right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma/\sqrt{\sigma^2 + 1}}} \exp\left( -\frac{u^2}{2\sigma^2/\left(\sigma^2 + 1\right)} \right) \, du \]

\[ = \frac{1}{\sqrt{2\pi\sqrt{\sigma^2 + 1}}} \exp\left( -\frac{t^2}{2(\sigma^2 + 1)} \right) ; \]

the last equation holds because the integral on the second line is 1, the integrand being the density function of an \( \mathcal{N}(0, \sigma^2/(\sigma^2 + 1)) \) distribution. The right-hand side is the density function of an \( \mathcal{N}(0, \sigma^2 + 1) \) distribution, completing the proof of the lemma.

**Corollary 26.1.** Let \( n > 1 \) be an integer and assume that \( X_1, X_2, \ldots, X_n \) are independent random variables with \( X_i \) having an \( \mathcal{N}(\mu_i, \sigma_i^2) \) distribution with \( \sigma_i > 0 \) for each \( i \) with \( 1 \leq i \leq n \). Then the sum \( X = \sum_{i=1}^{n} X_i \) has an \( \mathcal{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2) \) distribution.

**Proof.** The assertion that \( X \) has a normal distribution is an immediate consequence of Lemma 26.1. The rest of the assertion follows since for independent random variables the means and the variances add up.

## 27 Sums of independent discrete random variables

In the preceding section we considered the consequences of formula (26.3) to discuss the density function of sums of certain independent continuous random variables. A similar formula for discrete random variables would take a much more complicated form in view of the different sets of values assumed by discrete random variable. While a general formula could be discussed in terms of multivariate Stieltjes integrals, to be discussed below in Section 28, that are valid whether the variables are discrete or continuous or otherwise, practical calculations for discrete random variables
are often done with the aid of sums, the use of such a formula is too complicated for specific calculations. We find it more helpful to discuss some special cases to illustrate how such calculations are done.

27.1 Sums of independent Poisson variables

Let $X_1$ and $X_2$ be two independent Poisson random variables with parameters $\lambda_1$ and $\lambda_2$, and let $X = X_1 + X_2$. Intuitively, it is easy to see that then the variable $X$ has a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$. Indeed, $X_1$ counts the number of independent occurrences of an event in the interval $[0, 1]$ if the probability of an event occurring in a short interval of length $\delta$ is approximately $\lambda_1 \delta$. Similarly, $X_2$ counts the number of occurrences if the probability in the same interval is $\lambda_2 \delta$. Whether these events are of the same (kind) or not is immaterial. Then, for one of the events counted by $X_1$ or $X_2$ to occur in the same interval is $(\lambda_1 + \lambda_2) \delta$. This is because it is very unlikely for both events to occur in the same interval, so we can count these occurrences as approximately mutually exclusive. What follows is a formal calculation replacing this intuitive argument; our main goal is to show how to do such calculations. Otherwise, the intuitive argument should suffice by itself, since it would not be too difficult to turn it into a rigorously correct argument. For any $k \geq 0$ we have

$$P(X_1 = k) = \frac{\lambda_1^k}{k!} e^{-\lambda_1} \quad \text{and} \quad P(X_2 = k) = \frac{\lambda_2^k}{k!} e^{-\lambda_2}.$$ 

Hence, given $n \geq 0$ we have

$$P(X = n) = P(X_1 + X_2 = n) = \sum_{k=0}^{n} P(X_1 = k, X_2 = n - k) = \sum_{k=0}^{n} P(X_1 = k) P(X_2 = n - k)$$

$$= \sum_{k=0}^{n} \frac{\lambda_1^k}{k!} e^{-\lambda_1} \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2} = \frac{1}{n!} e^{-(\lambda_1 + \lambda_2)} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k}.$$ 

Using equation (3.3), the right-hand side can be written as

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

where the last equation follows by the binomial theorem. This equation shows that $X$ has a Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2$.

27.2 Sums of independent binomial variables

Let $X$ and $Y$ be independent discrete binomial variables $\text{Bin}(m, p)$ and $\text{Bin}(n, p)$, respectively, where $m$ and $n$ are positive integers and $0 < p < 1$. Then $X + Y$ has a $\text{Bin}(m + n, p)$ distribution. This is immediate by the following consideration. $P(X = k)$ means that if an experiment with probability $p$ of success is repeated $m$ times, then $k$ successes occur. $P(Y = l)$ means that if an experiment with probability $p$ of success is repeated $l$ times, then $l$ successes occur. Since it is irrelevant whether $X$ and $Y$ involve the same experiment or a different one, given that the probability of each success is the same, one can think about $X + Y = k + l$ to mean that if an experiment is repeated $m + n$ times, then a total of $k + l$ successes occur. So the result we will obtain is not in doubt; our aim is only to show how to do the calculation directly.
Write \( q = 1 - p \), given an arbitrary integer \( r \) with \( 0 \leq r \leq m + n \) we have

\[
\Pr(X + Y = r) = \sum_{k=0}^{r} \Pr(X = k, Y = r - k) = \sum_{k=0}^{r} \Pr(X = k) \Pr(Y = r - k)
\]

\[
= \sum_{k=0}^{r} \binom{m}{k} p^k q^{m-k} \binom{n}{r-k} p^{r-k} q^{n-(r-k)},
\]

note that we will not get in trouble if \( k > m \), since we have \( \Pr(X = k) = 0 \) and we also have \( \binom{m}{k} = 0 \) in that case according to equation (18.1). Similarly, it is not a problem if \( r - k > n \). The right-hand side further equals

\[
\sum_{k=0}^{r} p^r q^{m+n-r} \binom{m}{k} \binom{n}{r-k} = p^r q^{m+n-r} \sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-m}
\]

The right-hand side equals

\[
\binom{m+n}{r} p^r q^{m+n-r}
\]

as expected in view of the identity

(27.1)

\[
\sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-k} = \binom{m+n}{r}.
\]

A combinatorial interpretation of this identity is as follows. Let \( A \) and \( B \) be disjoint sets, \( A \) having \( m \) elements and \( B \) having \( n \) elements, to count the \( r \) elements subsets of \( A \cup B \), one needs pair every \( k \) element subset of \( A \) for \( 0 \leq k \leq r \) (if \( r > m \), there will be no such subset, but that is no problem) with every \( r - k \) element subset of \( B \).

Identity (27.1) can be extended as follows. If \( \alpha \) is real and \( k \geq 0 \) is an integer, recalling the definition of \( \binom{\alpha}{k} \) given in (18.2), for any two reals \( \alpha \) and \( \beta \) and for every integer \( r \geq 0 \) we have

(27.2)

\[
\binom{\alpha + \beta}{r} = \sum_{k=0}^{r} \binom{\alpha}{k} \binom{\beta}{r-k}.
\]

This can be verified by equating the coefficient of \( x^r \) on the two sides of the identity

\[
(1 + x)^{\alpha + \beta} = (1 + x)^{\alpha}(1 + x)^{\beta}
\]

written out in terms of the binomial series given in (18.3). Indeed, writing this identity in terms of the binomial series leads to the equation

\[
\sum_{r=0}^{\infty} \binom{\alpha + \beta}{r} x^r = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k \sum_{l=0}^{\infty} \binom{\beta}{l} x^l = \sum_{r=0}^{\infty} x^r \sum_{k=0}^{\infty} \binom{\alpha}{k} \binom{\beta}{r-k};
\]

the right-hand side results by multiplying together the two series in the middle. All the power series here converge if \( |x| < 1 \). Comparing the coefficients of \( x^r \) on the sides gives identity (27.2).

### 27.3 Reading

28.4 Homework

Chapter 6, pp. 291–296, Problems 2, 7, 10, 19, 16, 22, 23, 29 (need a table of the standard normal distribution for this; you can find one online), 32. Chapter 6, pp. 296–299, Theoretical Exercises 9, 10.

28 Multivariate Stieltjes integral

In order to discuss the expectation of functions of two random variables $X$ and $Y$ without necessarily assuming the existence of their density functions we need to extend the definition of Stieltjes integral to two dimensions. The extension to more than two dimensions is fairly simple, but far reaching generalizations are possible; see [19]. For a traditional treatment of multivariate Riemann–Stieltjes integrals, see [9]. Chapter III, pp. 101–140.

28.1 The double Stieltjes integral

Given real numbers $a < b$ and $c < d$ and integers $m,n > 0$ let $f$ and $g$ be functions on the two-dimensional interval $R = [a, b] \times [c, d]$. We want to define the integral

\[
\iint_S f(x,y) \, dx \, dy
\]

as a Riemann–Stieltjes integral. To this end, consider partitions (see Definition 10.1) of these intervals:

\[
P_x : a = x_0 < x_1 < x_2 < \ldots < x_m = b
\]

and

\[
P_y : c = y_0 < y_1 < y_2 < \ldots < y_n = d.
\]

These two partitions create a partition $P$ of the rectangle $R$ into smaller rectangles $R_{ij} = [x_{i-1}, x_i] \times [y_{j-1}, y_j]$. For $i,j$ with $1 \leq i \leq m$ and $1 \leq j \leq n$. Define the norm of the partition $P$ as

\[
\|P_x\| = \max\{x_i - x_{i-1} : 1 \leq i \leq m\}
\]

\[
\|P_y\| = \max\{y_j - y_{j-1} : 1 \leq j \leq n\}.
\]

We introduce two difference operators associated with these partitions. For any function $h(x,y)$, and ant $i,j$ with $1 \leq i \leq m$ and $1 \leq j \leq n$, we write

\[
\Delta_x h(x_i, y) \overset{def}{=} h(x_i, y) - h(x_{i-1}, y)
\]

\[
\Delta_y h(x, y_j) \overset{def}{=} h(x, y_j) - h(x, y_{j-1}).
\]

Then, for any such $i,j$, define the $g$-size of the rectangle $R_{ij}$ as

\[
\Delta_y \Delta_x g(x_i, y_j) = \Delta_x \Delta_y g(x_i, y_j)
\]

\[
= g(x_i, y_j) - g(x_{i-1}, y_j) - g(x_i, y_{j-1}) + g(x_{i-1}, y_{j-1});
\]

Note that in defining these difference operators, only some of the dependencies are indicated. For each such $j$ and $j$ pick a point $P_{ij} \in R_{ij}$ called a tag. The partition $P$ with these tags is called a

\[28.1\] The operator $\Delta x$ also depends i and the partition $P_{x}$; it is not worth clearing up the ambiguities in these symbols, since they are used more for motivation, rather than clearly describing expressions. For example, when writing $\Delta y \Delta x$ below, we would need to explain why $\Delta y$ affects only $y_j$, and not on $x_{i-1}$, for example.
tagged partition of the two-dimensional interval $R$. The Riemann-Stieltjes sum associated with this tagged partition is defined as

$$S = \sum_{i=1}^{m} \sum_{j=1}^{n} f(P_{ij}) \Delta y \Delta x g(x_i, y_i).$$

If these sums converge to a number $A$ as $\|P\| \to 0$ for all choices of the tags $P_{ij}$, then we call this number $A$ the Riemann–Stieltjes integral of $f$ with respect to $g$ on $R$:

$$\int_{R} \int_{R} f(x, y) \, dy \, dx \, g(x, y) = \int_{R} \int_{R} f \, dy \, dx \, g \overset{def}{=} A.$$

### 28.2 Iterated integrals: the difficulties

Writing a double Stieltjes integral as an iterated integral is not as straightforward as in case of the usual double Riemann integral. If we want to treat the summations in the double Riemann-Stieltjes sum given in (28.3), we need to choose tags where the first coordinate depends only on $i$ and the second coordinate depends only on $j$. That is, we need to take $P_{ij} = (\xi_i, \eta_j)$, where $x_{i-1} \leq \xi_i \leq x_i$ and $y_{j-1} \leq \eta_j \leq y_j$. For this choice we have

$$S_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} f(\xi_i, \eta_j) \Delta y \Delta x g(x_i, y_i).$$

If the Riemann–Stieltjes integral exists then this integral will certainly converge as $\|P\| \to 0$, but since the tags here are more restricted than in (28.3), the convergence of these sums is not sufficient for the existence of the integral. If we make $\|P_y\| \to 0$ without changing the partition $P_x$ and the corresponding tags $\xi_i$, the interval sums converge to a one-dimensional Stieltjes integral, and we are left with the sum

$$S_1 = \sum_{i=1}^{m} \int_{y=c}^{d} f(\xi_i, y) \, d(g(x_i, y) - g(x_{i-1}, y)).$$

In the lower limit of the integral, we wrote $y = c$ to indicate that the variable in this one-variable Stieltjes integral, $y$ is the variable.

If we now make $\|P_x\| \to 0$, the outside sum will also converge to a kind of integral, but not any kind of integral we discussed before. To this end, we will next introduce the Burkill integral. While we want to avoid going into deep discussions of these integral concepts, we need them to understand at some level in later discussions.

### 28.3 The Burkill integral

Let $\phi$ be a function that assigns subsets of $\mathbb{R}$ as values to subinterval $[u, v]$ ($u \leq v$) of a given interval $[a, b]$. We want to define the integral $\int_{a}^{b} \phi$. Consider a partition

$$P : a = x_0 < x_1 < x_2 < \ldots < x_n = b$$

of the interval $[a, b]$. For each $i$ with $1 \leq i \leq n$ let $c_i \in \phi([x_{i-1}, x_i])$; the sum

$$S = \sum_{i=1}^{n} c_i$$

74
will be called a Riemann sum for the integral $\int_a^b \phi$. If the limit of these Riemann sums exists when $\|P\| \rightarrow 0$, this limit is called the Burkill integral of $\phi$ on $[a, b]$ and it is written as $\int_a^b \phi(dx) = \int_a^b \phi$.

The Riemann integral $\int_a^b f(x) \, dx$ can be considered as a special case of the Burkill integral with

$$\phi([u, v]) = \{(v - u)f(\xi) : \xi \in [u, v]\}.$$ 

Taking limit $\|P\| \rightarrow 0$ in the sum in equation (28.5), we obtain the double Stieltjes integral in (28.1) as a Burkill integral

$$\int\int_S f(x, y) \, dx \, dy = \int_a^b \phi(dx),$$

where

$$\phi([u, v]) = \left\{ \int_{y=c}^d f(\xi, y) \, d(g(v, y) - g(u, y)) : \xi \in [u, v] \right\}.$$ 

### 28.4 Improper integrals

Extending these integrals to infinite regions can be accomplished by taking limits when $b, d \rightarrow +\infty$ and $a, c \rightarrow -\infty$.

### 29 Expectation of functions of two random variables

Given a pair of random variables $X$ and $Y$ with a joint distribution function $F_{X,Y}$ and a function $g$ of two variables, we have

$$E(g(X, Y)) = \int\int_{\mathbb{R}^2} g(x, y) \, dx \, dy \, F_{X,Y}(x, y),$$

assuming this integral is absolutely convergent, i.e., if the integral is convergent even if we replace $g(x, y)$ with $|g(x, y)|$. This can be explained in a way formula (11.4) was explained in Subsection 11.1 where the reason for requiring absolute convergence was also explained. To this end, note that according to formulas (25.1) and (28.2) we have

$$\Delta_x \Delta_y F_{X,Y}(x_i, y_j) = P(x_{i-1} < X \leq x_i \text{ and } y_{j-1} < Y \leq y_j),$$

and sum corresponding to the approximating sum in equation (28.3) to this integral becomes

$$S = \sum_{i=1}^m \sum_{j=1}^n g(\xi_{ij}, \eta_{ij}) \, P(x_{i-1} < X \leq x_i \text{ and } y_{j-1} < Y \leq y_j),$$

where we wrote $P_{ij} = (\xi_{ij}, \eta_{ij})$; note that $x_{i-1} \leq \xi_{ij} \leq x_i$ and $y_{j-1} \leq \eta_{ij} \leq y_j$. This sum restricts the integral to the rectangle $[a, b] \times [c, d]$, so after using these sums to approximate the integral on this rectangle, we need to make $a, c \rightarrow -\infty$ and $b, d \rightarrow \infty$ to obtain the improper integral in (29.1). It is simpler to discuss the basic properties of the expectation described in this integral in terms of the approximating sum than to discuss the general properties of the multi-dimensional Stieltjes
integral, especially in view of the technical difficulties involved in writing such integrals as iterated integrals.

The same comment applies here as in the one-dimensional case. From the viewpoint of modern integration theory, the pair of random variables \((X, Y)\) is a measurable transformation from \(\Omega\) into \(\mathbb{R} \times \mathbb{R}\), and the differential \(d_x d_y F_{X,Y}\) is just the measure \(d\{(P(X,Y)^{-1})\). Our formula \((29.1)\) is just the result of application of [8] §39 (Measurable Transformations), Theorem C, p. 163] to this measurable transformation. The technical difficulties implied by involving two-variable Stieltjes integrals illustrate the difficulties of our present approach not being able to take advantage of modern integration theory.

Instead of the approximating sum given in \((29.2)\) it will be simpler to work with the sum in \((28.4)\), where we took \(P_{ij} = (\xi_i, \eta_j), \) where \(x_{i-1} \leq \xi_i \leq x_i\). In this case, the sum \((29.2)\) becomes

\[
S_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} g(\xi_i, \eta_j) P(x_{i-1} < X \leq x_i \text{ and } y_{j-1} < Y \leq y_j),
\]

While it is convenient to work with this sum instead of the sum \((29.2)\), it is important to remember the convergence of the sum in \((29.2)\) is required for the existence of the integral in \((29.1)\); the convergence of the sum in \((29.3)\) is not sufficient.

### 29.1 Expectation when \(g(X, Y)\) only depends on \(X\)

In case \(g(x, y)\) depends only on \(x\), writing \(g(x) = g(x, y)\), the approximating sum \((29.3)\) becomes

\[
S_1 = \sum_{i=1}^{m} g(\xi_i) \sum_{j=1}^{n} P(x_{i-1} < X \leq x_i \text{ and } y_{j-1} < Y \leq y_j)
\]

\[
= \sum_{i=1}^{m} g(\xi_i) P(x_{i-1} < X \leq x_i \text{ and } y_0 < Y \leq y_n)
\]

\[
\to \sum_{i=1}^{m} g(\xi_i) P(x_{i-1} < X \leq x_i),
\]

where the limit on the right-hand side is obtained when \(c = y_0 \to -\infty\) and \(d = y_n \to \infty\). Noting that \(P(x_{i-1} < X < x_i) = F_X(x_i) - F_X(x_{i-1})\), the sum on the right-hand side is the approximating sum to the one-dimensional Riemann-Stieltjes integral

\[
\int_{a}^{b} g(x) \, dF_X(x) \to \mathbb{E}(g(X)),
\]

where the limit on the right-hand side is taken when \(a \to -\infty\) and \(b \to \infty\); cf. equation \((11.3)\). That is,

\[
(29.4) \quad \mathbb{E}(g(X)) = \iint_{\mathbb{R}^2} g(x) \, d_x d_y F_{X,Y}(x, y) = \int_{-\infty}^{\infty} g(x) \, dF_X(x);
\]

the integral on the right-hand side is a one-dimensional Stieltjes integral; in fact, it is the same Stieltjes integral that occurs in formula \((11.3)\). This is of course what one would naturally expect; the expectation \(\mathbb{E}(g(X))\) should not depend on the presence of the random variable \(Y\), given that \(Y\) has no influence here.
29.2 Expectation of a sum

The expectation of a sum was discussed above, in Subsection 13.1. Here we revisit the issue in the light of the joint distribution function. According to equation (29.1) we have

$$E(X + Y) = \int\int_{\mathbb{R}^2} (X + Y) \, dx \, dy \, F_{X,Y}(x,y)$$

$$= \int\int_{\mathbb{R}^2} X \, dx \, dy \, F_{X,Y}(x,y) + \int\int_{\mathbb{R}^2} Y \, dx \, dy \, F_{X,Y}(x,y) = E(X) + E(Y)$$

according to equation (29.4).

29.3 Expectation of product of independent variables

The expectation of the product of two independent variables was discussed in Theorem 13.1. Here we revisit the issue in the light of formula (29.1). Let $\phi$ and $\psi$ be two nice functions on $\mathbb{R}$; we will consider $E(\phi(X)\psi(Y))$; we will explain the reason why we do not restrict our discussion to $E(XY)$. If $g(x,y) = xy$ and the variables $XY$ are independent, we have

$$P(x_{i-1} < X \leq x_i \text{ and } y_{j-1} < Y \leq y_j) = P(x_{i-1} < X \leq x_i) \, P(y_{j-1} < Y \leq y_j),$$

and so the sum (29.3) approximating the integral expressing $E(\phi(X)\psi(Y))$ becomes

$$S_1 = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(\xi_i) \psi(\eta_j) \, P(x_{i-1} < X \leq x_i) \, P(y_{j-1} < Y \leq y_j),$$

(29.5)

$$= \left( \sum_{i=1}^{m} \phi(\xi_i) \, P(x_{i-1} < X \leq x_i) \right) \left( \sum_{j=1}^{n} \psi(\eta_j) \, P(y_{j-1} < Y \leq y_j) \right).$$

Taking limits in these sums, we can the analogous equation for integrals:

$$\int\int_{[a,c] \times [b,d]} \phi(x)\psi(y) \, dx \, dy = \int_{a}^{b} \phi(x) \, dF_{X}(x) \int_{c}^{d} \psi(y) \, dF_{Y}(y).$$

Making $a,c \to -\infty$ and $b,d \to \infty$, we obtain the equation

$$E(\phi(X)\psi(Y)) = E(\phi(X))E(\psi(Y)).$$

(29.6)

It would seem that involving the functions $\phi$ and $\psi$ in this equation is superfluous, since if $X$ and $Y$ are independent, $\phi(X)$ and $\psi(Y)$ are also independent, so it would appear to be enough to state the equation $E(XY) = E(X)E(Y)$ only. However, we have only discussed the independence of real-valued random variables, and we will need equation (29.5) specifically for complex-valued random variables when working with characteristic functions of random variables below. We could do this by basing our discussion on complex-valued random variables, but the tools we have in our possession are limited for this purpose. If we were able to use equation (11.2) to define the expectation of a random variable $X$, then it would make little difference whether $X$ is real-valued or complex-valued. But being forced to use equation expect: RS int defining $E$ of $X$, we are in a much more difficult situation when dealing with complex-valued random variables, since distribution functions only work for real-valued random variables.
30 Covariance and correlation

30.1 Random variable with zero variance

Lemma 30.1. If $X$ be a real-valued random variable such that $E(X^2) = 0$ then $P(X = 0) = 1$.

Proof. Let the events $A$ be the event $A = (X \neq 0)$; further, let $A_1 = (X \geq 1)$, and for $n > 1$ let $A_n = (1/(n-1) > X \geq 1/n)$. We have

$$A = \bigcup_{n=1}^{\infty} A_n.$$ 

As the events $A_m \cap A_n = 0$ if $m \neq n$, this implies that

$$P(A) = \sum_{n=1}^{\infty} P(A_n).$$

As $P(A) > 0$, there is an $n \geq 1$ such that $P(A_n) > 0$. With this $n$ we have

$$E(X^2) \geq P(A_n) \cdot \frac{1}{n^2} > 0.$$  \hfill \square

The following corollary is important to mention, even though we will not have an immediate use for it. The term almost surely used in its formulation refers to an event that has probability 1.

Corollary 30.1. If the random variable $X$ has variance 0, then it is almost surely constant.

Proof. Indeed, by Lemma 30.1 we have $P(X = E(X)) = 1$.  \hfill \square

30.2 Schwarz’s inequality

The next lemma also has a version for complex-valued random variables; even though we deal only with real-valued random variables in this notes, we want to emphasize the point that we are only discussing the real-valued version. The proof of the complex-valued version is somewhat more complicated.

Lemma 30.2 (Schwarz’s inequality). For any two real-valued random variables $X$ and $Y$, the inequality

$$(E(XY))^2 \leq E(X^2) E(Y^2).$$

holds.

Proof. We may assume that $P(X \neq 0) > 0$, since otherwise $E(XY) = 0$, so the inequality to be proved clearly holds. Then, according to Lemma 30.1 we have $E(X^2) > 0$. Let $\lambda$ be an arbitrary real number. Then, again by Lemma 30.1 the equation

$$E((\lambda X + Y)^2) = 0$$

\footnote{The first inequality here can be justified with the aid of Markov’s inequality \ref{31.1} given below. If the reader needs a formal justification, she is free to look ahead.}
can hold only if \( \lambda X + Y = 0 \) almost surely. As \( X \neq 0 \) with positive probability, this equation can only hold for a single value of \( \lambda \).

Now,
\[
E((\lambda X + Y)^2) = \lambda^2 E(X^2) + 2\lambda E(XY) + E(Y^2).
\]
Considering
\[
\lambda^2 E(X^2) + 2\lambda E(XY) + E(Y^2) = 0
\]
as a quadratic equation for \( \lambda \) with the various expectation as coefficients, this equation has at most one real solution, hence its discriminant cannot be positive. That is,
\[
(2 E(XY))^2 - 4 E(X^2) E(Y^2) \leq 0.
\]
Rearranging this, we obtain the inequality to be proved.

### 30.3 Covariance and correlation

Given two random variables \( X \) and \( Y \), their covariance measures how much these variables change together. It is defined as
\[
(30.1) \quad \text{Cov}(X, Y) \overset{\text{def}}{=} E\left((X - E(X))(Y - E(Y))\right).
\]
The correlation compares the joint variability of these random variables to their variances. It is defined as
\[
(30.2) \quad \text{Corr}(X, Y) \overset{\text{def}}{=} \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}.
\]

We have

**Lemma 30.3.** Given two real-valued random variables \( X \) and \( Y \) with nonzero variances, we have
\[
-1 \leq \text{Corr}(X, Y) \leq 1.
\]

**Proof.** We have
\[
\text{Corr}(X, Y) = \frac{E\left((X - E(X))(Y - E(Y))\right)}{\sqrt{E\left((X - E(X))^2\right) E\left((Y - E(Y))^2\right)}}.
\]
This is between \(-1\) and \(1\) in view of Schwarz’s inequality (cf. Lemma 30.2).

### 30.4 Covariance of two sums

Let \( m, n \geq 0 \) be integers, and let \( X_i \) and \( Y_j \) be random variables for \( i, j \) with \( 1 \leq i \leq n \) and \( 1 \leq j \leq m \). Write \( \mu_i = E(X_i) \) and \( \nu_j = E(Y_j) \) for the corresponding expectations. We have
\[
\text{Cov} \left( \sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j \right) = E \left( \sum_{i=1}^{n} (X_i - \mu_i) \sum_{j=1}^{m} (Y_j - \nu_j) \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} E((X_i - \mu_i)(Y_j - \nu_j)) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j).
\]

\[\text{This equation is a genuine quadratic equation, since } E(X^2) \neq 0, \text{ that is, the coefficient of } \lambda^2 \text{ is not zero, according to what we said above.}\]
Noting that \( V(X) = \text{Cov}(X, X) \) for any random variable, we in particular have
\[
V\left( \sum_{i=1}^{n} X_i \right) = \text{Cov}\left( \sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j)
\]
\[
= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} V(X_i) + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \text{Cov}(X_i, X_j);
\]
for the last equation, we made use of the fact that \( \text{Cov}(X_i, X_i) = \text{Cov}(X_j, X_i) \), so in the second sum, you included only terms for which \( j < i \), but we took these terms twice.

30.5 Reading

30.6 Homework

31 Chebyshev inequality and the weak law of large numbers

31.1 Markov’s inequality

The inequality in the next lemma is named after Andrey Markov, even though the inequality appears in an earlier work of his teacher, Pafnuty Chebyshev.

Lemma 31.1 (Markov’s inequality). Let \( X \geq 0 \) be a random variable such that \( E(X) \) exists, and let \( a > 0 \) be a real number. We have
\[
(31.1) \quad P(X \geq a) \leq \frac{E(X)}{a}.
\]

Proof. Define the random variable \( Y \) as
\[
Y = \begin{cases} 
  a & \text{if } X \geq a, \\
  0 & \text{otherwise}.
\end{cases}
\]

Define the \( f : \mathbb{R} \to \mathbb{R} \) as
\[
f(x) = \begin{cases} 
  x & \text{if } x < 0, \\
  0 & \text{if } 0 \leq x < a, \\
  a & \text{if } x \geq a.
\end{cases}
\]

We then have and \( Y = f(X) \). Hence
\[
a \cdot P(X \geq a) = E(Y) = E(f(X)) = \int_{-\infty}^{\infty} f(x) \, dF_X(x) \leq \int_{-\infty}^{\infty} x \, dF_X(x) = E(X);
\]
the inequality holds since \( f(x) \leq x \) for all \( x \) and \( P_X \) is a nondecreasing function. This establishes the desired inequality.

As a corollary, we have the following
Corollary 31.1 (Chebyshev’s inequality). Let $X$ be a random variable with variance $V(X) = \sigma^2$. Writing $m = \text{E}(X)$, we have

$$P(|X - m| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2},$$

for all $\epsilon > 0$.

Proof. We have

$$\sigma^2 = V(X) = \text{E}((X - m)^2) \geq \epsilon^2 P(|X - m| \geq \epsilon);$$

the last inequality holds in view of the Lemma 31.1 since $(X - m)^2$ is always nonnegative. Rearranging this inequality, the result follows. \qed

31.2 The weak law of large numbers

We have the following

Theorem 31.1 (The weak law of large numbers). Let $m$ and $\sigma \geq 0$ be real numbers, and let $X_1, X_2, X_3, \ldots$, be pairwise independent random variables with $\text{E}(X_i) = m$ and $V(X_i) = \sigma^2$ for $i$ with $1 \leq i < \infty$, and let $\epsilon > 0$. Writing

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

we have

$$\lim_{n \to \infty} P(|\bar{X}_n - m| \geq \epsilon) = 0.$$

Proof. We have

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

where the second equation holds because the variables $X_i$ are pairwise independent (one does not need full independence to ensure this equation). Hence, recalling that for a random variable $X$ and for a real number $\alpha$ we have $V(\alpha X) = \alpha^2 V(X)$, we have

$$V(\bar{X}_n) = V\left(\frac{1}{n} n\bar{X}_n\right) = \frac{1}{n^2} V(n\bar{X}_n) = \frac{1}{n} \sigma^2.$$

Hence, according to Chebyshev’s inequality (31.2) we have the following

$$P(|\bar{X}_n - m| \geq \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2} \to 0$$

as $n \to \infty$, completing the proof. \qed

31.1 That is, any two of them are independent. The book [3, Theorem 5, p. 121] makes the stronger assumption that these random variables are independent, but that assumption is never used. The assumption of pairwise independence is enough to ensure that equation (31.4) holds.
32 Statistical theory: point estimation

A good introduction to statistical theory is given in [3, Chapter 10–11, pp. 169–178]. The description is definitely worth reading, but it is not mathematically deep. Here we concentrate on the mathematical aspects.

32.1 Point estimation: general theory

In a statistical investigation one wants to find the value of a nonrandom quantity \( \theta \) that influences the observation; for example, the quantity \( \theta \) may be a physical constant, but more examples are given in [3, loc. cit.]. The results of each observation is a list \( x \) of numbers of a given length (that is, each measurement gives \( k \) numbers (such as, say, the weight and height of an object gives a list of length 2). If \( n \) independent observations are made, we have a sequence \( x_1, x_2, x_3, \ldots, x_n \), observations. This sequence is called a sample. In the mathematical model of the observation, \( x_i \) is the result of evaluating a vector-valued random variable \( X_i \). The reason for considering \( X_i \) a random variable is so that we can use methods of probability theory to analyse the experiment. The nature of randomness in evaluating \( X_i \) may lie in the behavior of the measuring instrument, or in the selection of the item to be measured from a large number of similar items, etc. We regard the observations as independent, since it is hardly reasonable to make a large number of observations, if an observation depends on another observation. So, in the mathematical model, the parameter \( \theta \) is estimated as a function \( \theta^* = \theta^*(X_1, X_2, X_3, \ldots, X_n) \). Instead of a single parameter \( \theta \), the experiment may involve a list of parameters \( \theta \), and one may consider a list \( \theta^* \) of functions approximating the list of parameters.

32.2 Estimation of the mean

All this is a bit too abstract, so we consider a concrete situation. In a large region, let \( m \) denote the average weight of all adult males. This is a fixed quantity, at least at a given time. In order to estimate the \( m \) it is usually not feasible to weigh all adult males and take the average (i.e., arithmetic mean). Instead, one randomly selects \( n \) individuals, and find that their weights are \( x_1, x_2, x_3, \ldots, x_n \), and one takes

\[
(32.1) \quad m^*(x_1, x_2, \ldots, x_n) = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i,
\]

as an approximation of \( m \). In order to know how good this way of approximating \( \theta \) is, one uses a probabilistic model of the experiment in which \( x_i \) is replaced with a random variable \( X_i \). It is assumed that \( \text{E}(X_i) = m \); in fact, one assumes that these variables are independent and identically distributed. One can study the behavior of the random variable

\[
(32.2) \quad m^*(X_1, X_2, \ldots, X_n) = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]

Such use of \( x_i \) referring to the measurement and \( X_i \) to the corresponding random variable is common.

---

32.1 Latin, loco citato, translated as in the place cited.
32.2 That is, the value if a list of numbers, rather just a single number.
32.3 In what follows, we will consider only a simplified model, in which each \( X_i = X_1 \) is a real-valued random variable.
32.4 The notation \( \bar{x}_n \) is customary in statistics for the expression on the right-hand side.
One finds that

\[ E(m^*(X_1, X_2, \ldots, X_n)) = E(\bar{X}_n) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) \]

(32.3)

\[ = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} \sum_{i=1}^{n} m = \frac{1}{n} nm = m. \]

Such an estimate is called unbiased. In general, an estimate \( \theta^*(X_1, X_2, \ldots, X_n) \) is called unbiased if

\[ E(\theta^*(X_1, X_2, \ldots, X_n)) = \theta. \]

Assuming that \( V(X_i) = \sigma^2 \) exists, we have for every \( \epsilon > 0 \),

\[ \lim_{n \to \infty} P \left( |m^*(X_1, X_2, \ldots, X_n) - m| \geq \epsilon \right) = \lim_{n \to \infty} P \left( \left| \frac{1}{n} \sum_{i=1}^{n} X_n - m \right| \geq \epsilon \right) = 0 \]

according to Theorem 31.1. Such an estimate is called consistent. In general, \( \theta^*(X_1, X_2, \ldots, X_n) \) is called consistent if for every \( \epsilon > 0 \) the relation

\[ \lim_{n \to \infty} P \left( |\theta^*(X_1, X_2, \ldots, X_n) - \theta| \geq \epsilon \right) = 0. \]

holds.

### 32.3 Estimation of the variance when the mean is known

We recall that the variance can be expressed as

\[ V(X) = E \left( (X - E(X))^2 \right) \]

(32.4)

Assuming that the value of \( m = E(X) \), estimating the variance comes down to estimating the value of

\[ V(X) = E \left( (X - m)^2 \right). \]

That is, we need to estimate the mean of the random variable \( (X - m)^2 \). This is the same problem we discussed in Subsection 32.2. So, we can just use the solution described there. That is,

\[ \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2 \]

is an unbiased and consistent estimate of \( \sigma^2 \).

### 32.4 Estimation of the variance when the mean is not known

When the variance is not known, one is tempted to use equation (32.5) with \( m \) replaced by by the estimate given for \( m \) in formula (32.2), but it turns out that this would lead to an estimate that is biased. The following is an unbiased estimate:

\[ S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2, \]

(32.6)
where $\bar{X}_n$ is given in equation (32.2) to estimate $m$. To see this, we will evaluate the expectation of $(n - 1)S^2$. Writing $m = E(X)$ we have

\[(n - 1)S^2 = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} ((X_i - m) - (\bar{X}_n - m))^2 \]

\[= \sum_{i=1}^{n} (X_i - m)^2 - \sum_{i=1}^{n} 2(\bar{X}_n - m)(X_i - m) + \sum_{i=1}^{n} (\bar{X}_n - m)^2 \]

\[= \sum_{i=1}^{n} (X_i - m)^2 - 2(\bar{X}_n - m)\sum_{i=1}^{n} (X_i - m) + n(\bar{X}_n - m)^2, \]

where the last equation holds since in the expression preceding it, the factor $2(\bar{X}_n - m)$ in the second sum does not depend on $i$, and in the third sum the terms do not depend on $i$. Noting that the second sum on the right-hand side equals $n(\bar{X}_n - m)$ according to equation (32.2), we have

\[(n - 1)S^2 = \sum_{i=1}^{n} (X_i - m)^2 - 2n(\bar{X}_n - m)^2 + n(\bar{X}_n - m)^2, \]

Hence

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

\[(32.7) \]

\[= \sum_{i=1}^{n} E((X_i - m)^2) - n E((\bar{X}_n - m)^2) = \sum_{i=1}^{n} V(X_i) - n V(\bar{X}_n); \]

the last equation holds since $E(X_i) = E(\bar{X}_n) = m$ (cf. (32.4)), Given that the variables $X_i$ are independent, we have

\[V(n\bar{X}_n) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2, \]

where the third equation holds since $V(X_i) = \sigma^2$. Therefore,

\[(32.8) \]

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

Substituting this into the right-hand side of (32.7), we obtain that

\[E((n - 1)S^2) = n\sigma^2 - \sigma^2 = (n - 1)\sigma^2. \]

Hence, $E(S^2) = \sigma^2$, showing that $S^2$ is indeed an unbiased estimate for $\sigma^2$. 

\[32.5\] The fact that $m$ is not known will not affect this calculation.

\[32.6\] The quantity being summed.

\[32.7\] Pairwise independence is enough for this.
32.5 Estimation of the standard deviation when the mean is not known

When \( \text{E}(X_i) \) is not known, one uses the square root of formula (32.6) to estimate the standard deviation:

\[
S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2},
\]

If \( U \) and \( V \) are random variables, and \( U = \sqrt{V} \), then \( \text{E}(U) \neq \sqrt{\text{E}(V)} \) unless \( U \) is almost surely constant. To see this, one needs to express the same statement in a different way:

\[
\text{If } U \text{ and } V \text{ are random variables and } V = U^2, \text{ then } \text{E}(V) \neq (\text{E}(U))^2 \text{ unless } U \text{ is almost surely constant.}
\]

Indeed, if \( \text{E}(V) = (\text{E}(U))^2 \) then

\[
V(U) = \text{E}(U^2) - (\text{E}(U))^2 = \text{E}(V) - \text{E}(V) = 0,
\]

which can happen only when \( U \) is almost surely constant. Indeed, \( V(U) = \text{E}((U - \text{E}(U))^2) \), and so, we need only to observe that, with \( Z = U - \text{E}(U) \), if \( \text{E}(Z^2) = 0 \) then \( Z = 0 \) almost surely (cf. Lemma 30.1).

33 The moment generating function of a random variable

Given a real-valued random variable \( X \), its moment generating function is the function

\[
M_X(t) \overset{def}{=} \text{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF_X(x).
\]

The reason for the name is that we have

\[
M_X(t) = \text{E}(e^{tX}) = \text{E}\left(\sum_{n=0}^{\infty} \frac{1}{n!} t^n X^n\right) = \sum_{n=0}^{\infty} \frac{1}{n!} \text{E}(X^n) t^n.
\]

This often allows one to quickly find the moments of a random variable.

33.1 The moment generating function of sums of independent random variables

Assuming \( X \) and \( Y \) are independent random variables having moment generating functions, we have

\[
M_{X+Y}(t) = \text{E}(e^{t(X+Y)}) = \text{E}(e^{tX} e^{tY}) = \text{E}(e^{tX}) \text{E}(e^{tY}) = M_X(t) M_Y(t),
\]

where the fourth equation holds since \( X \) and \( Y \) are independent – cf. (29.6).

33.2 The moment generating function of the exponential distribution

Assume \( X \) has the exponential distribution \( \text{Exp}(\lambda) \), where \( \lambda > 0 \). We are going to find its moment generating function, and use it to find the moments of \( X \).

\[32.8\text{“Almost surely” is a common expression in probability theory, used to mean “with probability 1.”}\]
We have
\[ f_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0, \\ \lambda e^{-\lambda x} & \text{if } x \geq 0. \end{cases} \]

\[ M_X(t) = \int_{-\infty}^{\infty} e^{xt} dF_X(x) = \int_{-\infty}^{\infty} e^{xt} f_X(x) \, dx = \int_{0}^{\infty} \lambda e^{xt} e^{-\lambda x} \, dx \]
\[ = \int_{0}^{\infty} \lambda e^{x(t-\lambda)} \, dx = \frac{\lambda}{\lambda - t}; \quad (t < \lambda). \]

for \( t \geq \lambda \) the integral is divergent. The right-hand side can be written as a power series centered at 0 and having radius of convergence \( \lambda \):
\[ \frac{\lambda}{\lambda - t} = 1 - \frac{t}{\lambda} \sum_{n=0}^{\infty} \frac{1}{\lambda^n} t^n \quad (t < \lambda). \]

Comparing this with equation (33.2), the coefficients must agree. Thus, we have
\[ (33.4) \quad E(X^n) = \frac{n!}{\lambda^n} \]

### 33.3 The moment generation function of the binomial distribution

Let \( X \) have a binomial distribution \( \text{Bin}(n, p) \). We are going to find its moment generating function, and use it to determine its first and second moments.

\( X \) assumes values \( k \) for \( 0 \leq k \leq n \) with probability \( ^n \! \mathrm{C}_k p^k (1-p)^{n-k} = P(X = k) = P(X \leq k) - P(X < k) = F_X(k) - F_X(k-0). \)

That is, at the places \( x = k \), the distribution function \( F_X(x) \) has jumps of the given size; elsewhere it is constant. Thus, according to Problem 10.1 we have
\[ M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF_X(x) = \sum_{k=0}^{n} e^{tk} P(X = k) \]
\[ = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = (e^t p + 1 - p)^n. \]

Assuming \( n \geq 2 \), we have
\[ M'_X(t) = pe^n (pe^t + 1 - p)^{n-1} \quad \text{and} \quad M''_X(t) = pe^n (pe^t + 1 - p)^{n-1} + p^2 e^n (n-1)(pe^t + 1 - p)^{n-2}. \]

Comparing coefficients in equation (33.2) with the coefficients in \( M_X(t) = M_X(0) + M'_X(0)t + \frac{M''_X(0)}{2} t^2 + \ldots \), we can see that \( E(X) = M'_X(0) = np \) and \( E(X^2) = M''_X(0) = np + n(n-1)p^2. \)

\[ ^{33.1}\text{Next, we use the notation } f(x-0) \overset{\text{def}}{=} \lim_{t \searrow x} f(t) \text{ and } f(x+0) \overset{\text{def}}{=} \lim_{t \nearrow x} f(t), \text{ commonly used in mathematics.} \]
33.4 The moment generating function of the Poisson distribution

Assume $X$ has Po($\lambda$) distribution. We are going to find its moment generating function.

For $k \geq 0$ we have

$$P(X = k) = \frac{\lambda^k}{k!}.$$  

Hence, similarly to the calculation in equation (33.5), we have

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} dF_X(x) = \sum_{k=0}^{\infty} e^{tk} P(X = k)$$

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

(33.6)

33.5 The moment generating function of a standard normal variable

Let $X$ be a standard normal variable. Then we have

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

Hence

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left(tx - \frac{t^2}{2}\right) \, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left(-\frac{(t+x)^2 - t^2}{2}\right) \, dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2} \, du,$$

where, in the last step, after moving the factor $e^{t^2/2}$ outside the integral, we made the substitution $u = x - t$ in the integral. The right-hand side can be written as

$$e^{t^2/2} \int_{-\infty}^{\infty} f_X(u) \, du = e^{t^2/2};$$

the equality here holds since the integral of a density function is always 1. That is,

$$M_X(t) = E(e^{tX}) = e^{t^2/2}.  \quad (33.7)$$

33.6 The moment generating function of a general normal variable

Let $Y$ be a normal variable with mean $\mu$ and variance $\sigma^2$, where $\sigma > 0$. Then $X = (Y - \mu)/\sigma$ is a standard normal variable. We have $Y = \sigma X + \mu$. Hence

$$M_Y(t) = E(e^{tY}) = E(e^{t(\sigma X + \mu)}) = E(e^{t\mu} e^{t\sigma X}) = e^{t\mu} E(e^{t\sigma X})$$

$$= e^{t\mu} M_X(t\sigma) = e^{t\mu} e^{(t\sigma)^2/2} = e^{t^2\sigma^2/2 + t\mu},$$

where the penultimate equality holds according to equation (33.7).
34 The characteristic function of a random variable

The characteristic function of a random variable involves the extension of the function $e^x$ to complex arguments; we will explain how to do this below. Given a real-valued random variable $X$, its characterising function is defined as the function

\begin{equation}
\phi_X(t) \overset{\text{def}}{=} E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} dF_X(x).
\end{equation}

That is, its definition is very similar to the definition of the moment generating function, but a very important difference is that every real-valued random variable has a characteristic function defined on the whole real line, while it is not unusual for such a variable to have a characteristic function only defined on a part of the real line; this is because $e^x$ is a fast increasing function when $x \to \infty$, while $|e^{ix}| = 1$ for all real $x$.

34.1 The exponential function for complex arguments

The exponential function $e^x = \exp x = \exp(x)$ can be defined by the equation

\begin{equation}
\exp x = e^x \overset{\text{def}}{=} \lim_{n \to \infty} \left(1 + \frac{x}{n}\right)^n.
\end{equation}

This equation can be used to extend the function $e^x$ for complex values of $x$; see the notes [14] for a detailed discussion. The Taylor expansion

\begin{equation}
\exp x = \sum_{n=0}^{\infty} \frac{x^n}{n!}
\end{equation}

can also be used, but with somewhat less intuitive appeal. One can show that

\begin{equation}
e^{ix} = \cos x + i \sin x;
\end{equation}

this is true for all complex $x$, but it is usually used with real $x$ to describe the exponential function for imaginary arguments. Replacing $x$ with $-x$, this equation becomes

\begin{equation}e^{-ix} = \cos x - i \sin x;
\end{equation}

using these two equations, one obtains the equations

\begin{alignat}{2}
\sin x &= \frac{e^{ix} - e^{-ix}}{2i} \quad & \text{and} & \quad \cos x &= \frac{e^{ix} + e^{-ix}}{2}.
\end{alignat}

The last four equations are called the Euler equations.

34.2 A detour in the history of logarithm

When John Napier invented logarithms in around 1590, we had a specific purpose in mind. Using the table of logarithms he created, you were able to calculate $xy$ by looking up $\log x$ and $\log y$ in the table, adding them, and noting the identity

\begin{equation}\log xy = \log x + \log y,\end{equation}

\footnote{See the website for an animation illustrating $e^{ix}$ as a limit described in formula (34.2).}
you could find \( xy \) by doing a reverse look-up in the table. This technology was incorporated also in slide rules. The use of logarithm tables was taught in highschools all over the world until about the 1960s, and slide rules were carried by many engineers in their shirt pockets until that time, when the invention of hand calculators made the technology obsolete.

It is interesting to note that Napier traveled in Europe, and observed astronomers in Amsterdam using trigonometric tables, in wide use among astronomers at that time, to simplify multiplication. The identity they relied on was using trigonometric tables, in wide use among astronomers at that time, to simplify multiplication. The invention of hand calculators made the technology obsolete.

1960s, and slide rules were carried by many engineers in their shirt pockets until that time, when slide rules. The use of logarithm tables was taught in high schools all over the world until about the

\[(34.5) \quad 2 \cos x \cos y = \cos(x + y) + \cos(x - y) \quad \text{and} \quad 2 \sin x \sin y = \cos(x + y) - \cos(x - y);\]

see [1] p. 340, pdf p. 357. The problem with this approach was that it required significantly more table lookups than the method discovered by Napier. It is interesting to note that identities \((34.4)\) and \((34.5)\) are related by Euler’s identity \((34.4)\), discovered about 150 years later.

### 34.3 The characteristic function of a standard normal variable

Let \( X \) be a standard normal variable. Then its density function is

\[f_X(x) = F'_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.\]

Hence its characteristic function equals

\[\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} dF_X(X) = \int_{-\infty}^{\infty} e^{itx} F'_X(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx;\]

the third equality here holds according to Theorem \[10.1\]. Continuing this calculation, we have

\[\phi_X(t) = \frac{1}{\sqrt{2\pi}} \lim_{A \to \infty} \int_{-A}^{A} \exp \left( -\frac{1}{2} (x^2 - 2itx + (it)^2) - \frac{1}{2} t^2 \right) dx\]

\[= \frac{e^{-t^2/2}}{\sqrt{2\pi}} \lim_{A \to \infty} \int_{-A}^{A} \exp \left( -\frac{1}{2} (x + it)^2 \right) dx = \frac{e^{-t^2/2}}{\sqrt{2\pi}} \lim_{A \to \infty} \int_{-A+it}^{A+it} e^{-u^2/2} du.\]

There are several points to explain here. After the first equality, we took the limit of the integral from \(-A\) to \(A\); this is correct, since the integral is absolutely convergent notwithstanding one should integrate from \(A\) to \(B\) with \(A \to -\infty\) and \(B \to \infty\). After the last integral, we substituted \(u = x + it\), and next we will break up the integral into three parts. To understand this properly, one needs to use line integrals in the complex plane. As for the right-hand side here, we have

\[\int_{-A+it}^{A+it} e^{-u^2/2} du = \left( \int_{-A}^{A} + \int_{A}^{A+it} + \int_{-A+it}^{A} \right) e^{-u^2/2} du.\]

Here the limit of the first and third integral tend to 0 when \(A \to \infty\), since the integrand tend to zero along the line integration, and the length \(t\) of the line of integration does not depend on \(A\); As for the second integral, it tends to \(\sqrt{2\pi}\) according to equation \((20.2)\) \[34.4\]. Hence it follows that

\[\phi_X(t) = e^{-t^2/2}.\]

\[34.2\] i.e., the integral of the absolute value of the integral exists on the interval \((-\infty, \infty)\).

\[34.3\] Further, one needs to know that complex line integrals of holomorphic (i.e., differentiable in the complex sense) functions are independent of the line of integration. If the reader is not familiar with complex function theory, he can be satisfied with following the calculation in a formal sense, or she can be motivated to take a course in complex variables. (This is my feeble attempt at humor, so as to avoid using “he or she,” or the more polite “she or he,” or using the recently reintroduced singular “they.” I am lucky that Hungarian is a genderless language.

\[34.4\] One needs to do a simple change of variable in the latter integral to see this.
34.4 The characteristic function of sums of independent random variables

If \( X \) and \( Y \) is independent, then

\[
\phi_{X+Y}(t) = E(e^{it(X+Y)}) = E(e^{itX}e^{itY}) = E(e^{itX})E(e^{itY}) = \phi_X(t)\phi_Y(t),
\]

where the third equality follows from equation (29.0). This naturally extends to the sum of any number of independent random variables, since if \( n \geq 2 \) and \( X_1, X_2, \ldots, X_n \) are independent, then \( \sum_{k=1}^{n-1} X_k \) and \( X_n \) are independent, so using the above equation, one can prove by induction on \( n \) that if \( X_1, X_2, \ldots, X_n \) are independent, then

\[
\phi_{\sum_{k=1}^{n} X_k}(t) = \prod_{k=1}^{n} \phi_{X_k}(t).
\]

35 The central limit theorem

35.1 Statement of the theorem

The Central Limit Theorem states that if \( X_1, X_2, X_2, \ldots \) is an infinite sequence of independent, identically distributed (abbreviated as i. i. d.) random variables with \( E(X_k) = \mu \) and \( V(X_k) = \sigma^2 \) (i.e., we assume that the expectation and the variance of these variables exists)\(^{35.2}\) Then the Central Limit Theorem asserts that the distribution of the average

\[
\bar{X}_n \overset{def}{=} \frac{1}{n} \sum_{k=1}^{n} X_k
\]

approximates a normal distribution. To state this precisely, note that

\[
E(\bar{X}_n) = \frac{1}{n} \sum_{k=1}^{n} E(X_k) = \frac{1}{n} \sum_{k=1}^{n} \mu = \mu
\]

and

\[
V(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} V(X_k) = \frac{1}{n^2} \sum_{i=k}^{n} \sigma^2 = \frac{1}{n}\sigma^2.
\]

according to equations (13.1), (14.3), and Theorem 14.1. Using also equation (14.2), we obtain that for \( Y_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma \) we have \( E(Y_n) = 0 \) and \( V(Y_n) = 1 \). The way \( Y_n \) results from \( \bar{X}_n \) is called standardization, and \( Y_n \) is called the standard variable corresponding to \( X_n \). The Lindeberg–Lévy Central Limit Theorem can be formulated as follows.

**Theorem 35.1** (Lindeberg–Lévy Central Limit Theorem). Let \( \sigma > 0 \) and \( \mu \) be real numbers, let \( X_1, X_2, X_2, \ldots \) be an infinite sequence of independent, identically distributed random variables with \( E(X_k) = \mu \) and \( V(X_k) = \sigma^2 \) for each \( k > 0 \), and for each \( n > 0 \) write

\[
\hat{X}_n \overset{def}{=} \frac{1}{n} \sum_{k=1}^{n} X_k.
\]

\(^{35.1}\)An important abbreviation to remember, since it is used frequently.

\(^{35.2}\)The existence of the second moment of a random variable ensures that both its expectation and its variance exist, though the proof may not be appropriate for the level of this course.
Then, for all real $x$ we have

\begin{equation}
\lim_{n \to \infty} P \left( \frac{\sqrt{n} (X_n - \mu)}{\sigma} \leq x \right) = \Phi(x).
\end{equation}

There are many kinds of convergence in probability theory. The convergence described in this theorem is called convergence in distribution. That is, the standardized variable on the left-hand side of equation (35.1) is said to converge in distribution to the standard normal variable. The full proof of this theorem is beyond the level of this course, but in what follows we will outline the main steps in the proof, and explain what gaps need to be filled.

35.2 The main arguments of the proof

In the proof, we will assume that $E(X_k) = 0$ and $V(X_k) = 1$ for all $k > 0$, that is, $\mu = 0$ and $\sigma = 1$; if this were not the case, we could replace $X_k$ by the normalized variable $(X_k - \mu)/\sigma$. With these assumptions, writing $Y_n = \sqrt{n} \bar{X}_n$, formula (35.1) becomes

\begin{equation}
\lim_{n \to \infty} P(Y_n < x) = \Phi(x),
\end{equation}

where we have

$$Y_n = \sum_{k=1}^{\infty} \frac{X_k}{\sqrt{n}}.$$ 

Since the random variables on the right-hand side are independent, for the corresponding characteristic functions we have

\begin{equation}
\phi_{Y_n} = \prod_{k=1}^{n} \phi_{X_k/\sqrt{n}} = (\phi_{X_1/\sqrt{n}})^n;
\end{equation}

the second equality holds since the variables $X_k$ are identically distributed, so their characteristic functions are also identical. The proof of Theorem 35.1 now consists of two parts Writing $\phi(x)$ for the characteristic function function of the standard normal distribution, first we need to show that

\begin{equation}
\lim_{n \to \infty} \phi_{Y_n}(x) = \phi(x).
\end{equation}

In the second part, we need to show that if $Z_n$ and $Z$ are random variables such that $\lim_{n \to \infty} \phi_{Z_n}(x) = \phi_Z(x)$ for every real $x$, then $\lim_{n \to \infty} F_{Z_n}(x) = F_Z(x)$ at every $x$ where $F_Z(x)$ is continuous. Since in our case, $\phi(x)$ is continuous everywhere, from this we can conclude that formula (35.1) holds for all $x$.

Since the necessary tools to establish these results, including the rigorous definition of convergence of sequences and uniform convergence of sequences of functions are only discussed in more advanced courses, none of these steps can be discussed rigorously, but we will try to outline the steps so as to clarify the issues sufficiently at the level of the present course.

35.3 Estimating the exponential function on the unit circle

We will use Euler’s identity (34.3) and the Taylor series of $\cos x$ and $\sin x$ on the real line to approximate $e^{ix}$ for real $x$. Taylor’s formula says that, given a function that is differentiable $n + 1$ times.

---

35.3 Such descriptions may serve as motivation to students to take these more advanced courses.
times \((n \geq 0)\) in an open interval containing \(a\) and \(x\), we have
\[
(35.4) \quad f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!} (x-a)^k + \frac{f^{(n+1)}(\xi)}{(n+1)!} (x-a)^{n+1}
\]
for some \(\xi\) between \(a\) and \(x\) (that is, \(a < \xi < x\) or \(x < \xi < a\) or \(a = \xi = x\)), where the last term is the Lagrange remainder term; see [13] § 23, pp. 47–48. Using this with \(f(x) = \cos x\) or \(\sin x\), \(a = 0\), and \(n = 2\), we have
\[
(35.5) \quad \cos x = 1 - \frac{x^2}{2} - \frac{\sin \xi}{6} x^3
\]
and
\[
(35.6) \quad \sin x = x - \frac{\cos \eta}{6} x^3
\]
for some \(\xi\) and \(\eta\) between 0 and \(x\).\(^{35.4}\) Hence
\[
(35.7) \quad e^{ix} = \cos x + i \sin x = 1 - ix - \frac{x^2}{2} + \frac{\sqrt{2}}{6} \theta(x)x^3 \quad (|\theta(x)| \leq 1),
\]
for all real \(x\), where \(\theta(x)\) is a complex-valued function of \(x\):
\[
\theta(x) = -\frac{i}{\sqrt{2}} (i \sin \xi + \cos \eta):
\]
the inequality \(|\theta(x)| \leq 1\) follows since \(|\sin \xi| \leq 1\) and \(|\cos \eta| \leq 1\). Thus, given a real numbers \(M > 0\) and \(t\), for \(|x| \leq M\), we have
\[
(35.8) \quad \left| e^{ixt/\sqrt{n}} - \left(1 - \frac{ixt}{\sqrt{n}} - \frac{x^2 t^2}{2n}\right) \right| \leq \frac{\sqrt{2}}{6\sqrt{3}n} x^3 t^3 \leq \frac{\sqrt{2}}{6\sqrt{3}n} M x^2 t^3.
\]

35.4 Estimating the characteristic function of \(X_1/\sqrt{n}\)

Let \(t\) be a fixed real, and let \(n > 0\) be an integer. Let \(M > 0\) a real number, to be selected later. We will use equation (35.3) to estimate \(\phi_{X_1/\sqrt{n}}(t) = \phi(X_1)(t/\sqrt{n})\). On account of formula (35.8), first note that
\[
(35.9) \quad \int_{-\infty}^{\infty} \left(1 - \frac{ixt}{\sqrt{n}} - \frac{x^2 t^2}{2n}\right) dF_{X_1}(x)
\]
\[
= E(1) - \frac{it}{\sqrt{n}} E(X_1) - \frac{t^2}{2n} E(X_1^2) = 1 - \frac{t^2}{2n}.
\]
For the last equation, recall that \(X_1\) is a standardized variable, so \(E(X_1) = 0\) and \(E(X_1^2) = 1\). To approximate \(\phi_{X_1}\), we will use inequality (35.8) to estimate \(e^{ixt/\sqrt{n}}\) on \((-M, M)\), and use the estimate \(|e^{ixt/\sqrt{n}}| = 1\) on the rest of the real line. Thus,
\[
(35.10) \quad \left| \phi_{X_1/\sqrt{n}}(t) - \left(1 - \frac{t^2}{2n}\right) \right| \leq \int_{|x| \leq M} \frac{\sqrt{2}}{6\sqrt{3}n} M x^2 |t|^3 dF_{X}(x)
\]
\[
+ \int_{|x| > M} \left(1 + \frac{|xt|}{\sqrt{n}} + \frac{x^2 t^2}{2n}\right) dF_{X}(x) + \int_{|x| > M} dF_{X}(x);
\]
\(^{35.4}\)For \(\cos x\), the coefficient of \(x^4\) in the Taylor series is 0, so we could have taken the remainder term replacing the term involving \(x^4\), but it is better to use the remainder term involving the same power of \(x\) for both \(\sin x\) and \(\cos x\).
the first term on the right-hand side compensates for the error in the approximation given in (35.8),
the second terms corrects for taking this approximation on \((-\infty, \infty)\) rather than on \((-M, M)\), and
the third tierm estimates the integral of \(e^{ixt/\sqrt{n}}\) on the set \(|x| > M\).

We will pick an arbitrary \(\epsilon\) with \(0 < \epsilon < 1\), and make the choice \(M = \epsilon\sqrt{n}\). We are going to
estimate the integrals on the right-hand side of (35.10). The firt integral on the right-hand side is then
\[
\leq \int_{-\infty}^{\infty} \frac{\sqrt{\epsilon}}{6n} e^{x^2/\epsilon} |t|^3 \, dF_X(x) \leq \frac{\sqrt{\epsilon}}{6n} |t|^3,
\]
since \(E(X^2) = 1\). As for the third integral, we have \(1 \leq x^2/(n\epsilon^2)\) for \(|x| > M\), so this integral is
\[
\leq \int_{|x| > M} \frac{x^2}{n\epsilon^2} \, dF_X(x) \leq \frac{\epsilon}{n}
\]
provided \(M = \epsilon\sqrt{n}\) is large enough. This is because \(E(X^2) = 1\), and so, for large enough \(M\) we have
(35.11)
\[
\leq \int_{|x| > M} x^2 \, dF_X(x) \leq \epsilon^3.
\]
This inequality also handles the third term in the second integral on the right-hand side of (35.10). The first term in the second integral is the same as the integrand in the third integral. As for the
second term in the second integral, we note that \(|x| < x^2/(\epsilon\sqrt{n})\), so this integral is again
\[
\leq \int_{|x| > M} \frac{x^2}{n\epsilon^2} \, dF_X(x) \leq \frac{\epsilon^2 t}{n}.
\]
Combining all these calculations, we can find a positive constant \(C\) depending on \(t\) but not on \(\epsilon\) and
\(n\) (note that \(t\) is fixed in the present context) \([35.5]\) such that
(35.12)
\[
\left| \phi_{X_1/\sqrt{n}}(t) - \left(1 - \frac{t^2}{2n}\right) \right| < \frac{C\epsilon}{n}
\]
provided that \(n\) (and so \(M\)) is large enough.

35.5 Differentiability of the characteristic function

Noting that \(\phi_{X_1/\sqrt{n}}(t) = \phi(X_1)(t/\sqrt{n})\), inequality (35.12) can also be derived by noting that \(\phi_{X_1}(t)\)
is twice differentiable at \(t = 0\) with respect to the real variable \(t\) (but not necessarily with respect
to the complex variable \(t\)). To show this is beyond the level of this course. Here we only mention
that it is not hard to show from the assumption that \(E(X_1)\) and \(E(X_1^2)\) exist by using Lebesgue
Bounded Convergence Theorem that the interchange of differentiation and integration is justified in the equations
(35.13)
\[
\frac{d}{dt} \phi_{X_1}(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{itx} \, dF_{X_1}(x) = \int_{-\infty}^{\infty} \frac{\partial}{\partial t} e^{itx} \, dF_{X_1}(x) = \int_{-\infty}^{\infty} ix e^{itx} \, dF_{X_1}(x)
\]
and
(35.14)
\[
\frac{d^2}{dt^2} \phi_{X_1}(t) = \frac{d^2}{dt^2} \int_{-\infty}^{\infty} e^{itx} \, dF_{X_1}(x) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial t^2} e^{itx} \, dF_{X_1}(x) = -\int_{-\infty}^{\infty} x^2 e^{itx} \, dF_{X_1}(x).
\]
\[^{35.5}\]C would be easy to calculate, but it is not important do so for our purposes.
Using this, we obtain

\[ \phi_{X_1}(t) = \phi_{X_1}(0) + \int_0^t \phi'_{X_1}(u) \, du = \phi_{X_1}(0) + \int_0^t \phi'_{X_1}(0) \, du + \int_0^t \phi''_{X_1}(0)u \, du \]

\[ + \int_0^t \left( \phi'_{X_1}(u) - (\phi'_{X_1}(0) + \phi''_{X_1}(0)u) \right) \, du; \]

here, the first equation holds by the Fundamental Theorem of Calculus, given that \( \phi'_{X_1} \) is continuous, given that it is even differentiable. By the definition of the derivative, given an arbitrary \( \epsilon > 0 \), there is a \( \delta > 0 \) such that for every \( u \) with \( |u| < \delta \) we have

\[ \left| \phi'_{X_1}(u) - (\phi'_{X_1}(0) + \phi''_{X_1}(0)u) \right| < \epsilon u. \]

Hence, for any \( t \) with \( |t| < \delta \), the last integral on the right-hand side of the previous formula equals \( \epsilon \theta(t) t^2 / 2 \) for some function \( \theta(t) \) with \( |\theta(t)| \leq 1 \). Thus

\[ \phi_{X_1}(t) = \phi_{X_1}(0) + \phi'_{X_1}(0)t + \phi''_{X_1}(0) \frac{t^2}{2} + \frac{\epsilon \theta(t)t^2}{2} \]

(35.15)

\[ = 1 + E(X_1)it - E(X_1^2) \frac{t^2}{2} + \frac{\epsilon \theta(t)t^2}{2} = 1 - \frac{t^2}{2} + \frac{\epsilon \theta(t)t^2}{2}; \]

here the second equation holds since \( \phi_{X_1}(0) = 1 \) (every characteristic function is 1 at 0), and, further \( \phi'_{X_1}(0) = i E(X_1) \) and \( \phi''_{X_1}(0) = -E(X_1^2) \) according to formulas (35.13) and (35.14) with \( t = 0 \) and \( \phi'_{X_1}(0) = -E(X_1^2) \). The third equation holds since \( E(X_1) = 0 \) and \( E(X_1^2) = 1 \) by our assumptions. Hence, for fixed \( t \) and large \( n \) we have

\[ \phi_{X_1/n}(t) = \phi_{X_1} \left( \frac{t}{\sqrt{n}} \right) = 1 - \frac{t^2}{2n} + \theta \left( \frac{t}{\sqrt{n}} \right) \frac{\epsilon t^2}{2n}. \]

This equation is equivalent to inequality (35.12).

### 35.6 Lindberg’s version of the central limit theorem

While the more advanced arguments in Subsection 35.5 would greatly simplify the proof given here for for the version of the central limit theorem in Theorem 35.1 there is a more general version due to Lindberg that needs the more detailed estimates in Subsection 35.4 with minor modifications (see [18] Theorem 4.7.1, § 4.7, p. 223):

---

35.6 One needs to be a little careful here, since the equations \( \phi'_{X_1}(0) = i E(X_1) \) and hold only by the interchangeability of differentiation and integration of the quoted formulas, so the situation is not the same as in case of the moment generation function, where the formulas always hold as long as the moment generating function is defined near 0 (the existence of the moment generating function near zero guarantees the analogous interchangeability of differentiation and integration by the Lebesgue Bounded Convergence Theorem; in case of the characteristic function this is guaranteed only by the existence of the corresponding moments).

We note that equation (35.15) can also be derived different ways. One can use the Taylor formula with second order remainder term separately for the real and the imaginary parts of \( \phi_{X_1} \) (the usual formulas for the remainder term of the Taylor series are only valid for real-valued functions), but one needs the continuity of \( \phi'_{X_1} \) near zero (the continuity everywhere of this derivative can be derived from equation (35.14) with the aid of the Lebesgue Bounded Convergence Theorem).
Theorem 35.2. Let $X_1, X_2, X_2, \ldots$, be an infinite sequence of independent random variables with expectation $E(X_k) = 0$ and finite variance $\sigma_k^2 = V(X_k)$ for all $k > 0$, and write

$$D_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2}.$$

For $\epsilon > 0$ put

$$L_n(\epsilon) = \frac{1}{D_n^2} \sum_{k=1}^{n} \int_{|x| > \epsilon D_n} x^2 dF_{X_k}(x),$$

and assume that

(35.16) \[ \lim_{n \to \infty} L_n(\epsilon) = 0. \]

Then

$$\lim_{n \to \infty} P \left( \frac{1}{D_n} \sum_{k=1}^{n} X_k < x \right) = \Phi(x)$$

for all $x \in \mathbb{R}$; furthermore, the convergence in this equation is uniform on $\mathbb{R}$.\(^{35.7}\)

Equation (35.16) is called Lindberg's condition. This replaces equation (35.11) in the more general situation in the present theorem (note that $M = \epsilon \sqrt{n}$, and $D_n = \sqrt{n}$ with the assumptions of Theorem 35.1).

35.7 Finishing the first part of the proof

The fact that in inequality (35.12) $\epsilon$ can be an arbitrary number with $0 < \epsilon < 1$ immediately implies (35.3) in view of equation (34.2) by a standard argument about sequences, thus completing the first part of the proof of Theorem 35.1. For this, note that the constant $C$ does not depend on $n$ and $\epsilon$ (but it may depend on $t$; however, $t$ is fixed throughout this argument.)

However, the argument about sequences mentioned is common only in more advanced courses. While we present the argument for the sake of completeness, the reader may feel free to skip the details. According to inequality (35.12), there is a complex number $\lambda_n$ such that

(35.17) \[ \phi_{X_1/\sqrt{n}}(t) = \exp \left( -t^2 \frac{\lambda_n}{2n} \right), \]

and $|\lambda_n| < C\epsilon$ for large enough $n$. This means that there is a complex $\lambda$ with

(35.18) \[ |\lambda| \leq C\epsilon \]

such that the sequence $\{\lambda_n\}_{n=1}^{\infty}$ has a subsequence $\{\lambda_{n_k}\}_{k=1}^{\infty}$ for which $\lim_{k \to \infty} \lambda_{n_k} = \lambda$. Now, equation (35.18) holds for all values of $\epsilon$ for which (35.12) holds, i.e., for every $\epsilon$ with $0 < \epsilon < 1$; hence $\lambda = 0$. Thus $\lim_{n \to \infty} \lambda_n = 0$. Hence

$$\lim_{n \to \infty} \left( \phi_{X_1/\sqrt{n}}(t) \right)^n = \lim_{n \to \infty} \left( 1 - t^2 \frac{\lambda_n}{2n} \right)^n = \exp \left( -t^2 \frac{\lambda}{2} \right),$$

establishing (35.3).\(^{35.7}\)

The sequence of real-valued functions $f_n(s)$ is said to converge to the function $f(x)$ on the set $S$ uniformly if for every $\epsilon > 0$ there is an $N$ such that we have $|f_n(x) - f(x)| < \epsilon$ for all $n \geq N$ and for all $x \in S$. This concept of convergence is stronger than pointwise convergence. In pointwise convergence, one requires that for all $x \in S$ and for all $\epsilon > 0$ there is an $N$ such that we have $|f_n(x) - f(x)| < \epsilon$ for all $n \geq N$ and for all $x \in S$. In uniform convergence we need to choose an $N$ that works for all $x$ and $\epsilon$, while in pointwise convergence, for each $x$ we may pick a different value for $N$. See more for the discussion in Section 38 and especially in Subsection 38.1.
35.8 The second part of the proof: the Lévy continuity theorem

The second part of the proof is based on the following theorem of Paul Lévy:

**Theorem 35.3 (Lévy continuity theorem).** Given a random variables $X$ and a sequence of random variables $X_n$ for $n \geq 1$, if

$$\lim_{n \to \infty} \phi_{X_n}(t) = \phi_X(t)$$

for every $t \in \mathbb{R}$, then

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

for every $x \in \mathbb{R}$ where $F_X(x)$ is continuous.

This is a purely function-theoretic result, and it is not about random variables, so it is not necessary for the random variables $X$ and $X_n$ to share the same probability space. A function-theoretic reformulation would say the following:

**Theorem 35.4 (Lévy continuity theorem, function-theoretic version).** Given nondecreasing real-valued functions $F$ and $F_n$ on the real line such that $0 \leq F(x) \leq 1$ and for all $x \in \mathbb{R}$, and $0 \leq F_n(x) \leq 1$ and for all $x \in \mathbb{R}$ for all $n$ with $n \geq 1$. Assume, further, that

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} F_n(x) = 0$$

for all $n \geq 1$. If

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} e^{ixt} dF_n(x) = \int_{-\infty}^{\infty} e^{ixt} dF(x)$$

for every $t \in \mathbb{R}$, then

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

for every $x \in \mathbb{R}$ where $F(x)$ is continuous.

In the first step of the proof, one shows that

$$\lim_{n \to \infty} \int_{-\infty}^{\infty} g(x) dF_n(x) = \int_{-\infty}^{\infty} g(x) dF(x)$$

for every continuous function $g$ that is zero outside a finite interval. This can be done by the well-known Weierstrass approximation theorem formulated for trigonometric polynomials. According to this, given an $\epsilon > 0$ and a continuous function $g$ on the interval $[-L, L]$ such that $g(L) = g(-L)$, there is a trigonometric polynomial

$$T(x) = \sum_{k=-n}^{n} c_k \exp \left( \frac{\pi i k x}{L} \right)$$

($\bar{z}$ denotes the complex conjugate of the number $z$) such that $|g(x) - T(x)| < \epsilon$ for all $x \in [-L, L]$.
Assuming \( g \) is a continuous function on \( \mathbb{R} \) supported on the interval interval \((-A, A)\), approximate \( g \) with an error \(< \epsilon \) \((\epsilon > 0)\) by such a polynomial \( T(x) \) on an interval \((-L, L)\), where \( L \) is (much) larger than \( A \). By (35.20) we have

\[
\lim_{n \to \infty} \int_{-\infty}^{\infty} T(x) \, dF_{X_n}(x) = \sum_{k=-n}^{n} c_k \lim_{n \to \infty} \int_{-\infty}^{\infty} \exp \left( \frac{\pi i k x}{L} \right) \, dF_X(x) = \int_{-\infty}^{\infty} T(x) \, dF_X(x).
\]

Making \( L \to \infty \) and \( \epsilon \downarrow 0 \), formula (35.22) follows for the function \( g \) being approximated (that is, for all continuous \( g \) that is zero outside a finite interval). For this, note that the trigonometric polynomial \( T(x) \) in (35.23) has period \( 2L \), that is, the part approximating the nonzero part of \( g \) gets repeated at every period, but as \( L \to \infty \), the integrals of the repeated part can be neglected.

Next, if \( F \) is continuous at \( a \) and \( b \), approximate the function \( h(x) = 1 \) if \( a \leq x \leq b \), \( 0 \) if \( x < a \) or \( x > b \) by continuous functions \( g_\epsilon \) such that \( 0 \leq g_\epsilon(x) \leq 1 \) for all \( x \in \mathbb{R} \) and \( g_\epsilon(x) = h(x) \) for all \( x \in \mathbb{R} \setminus ((a - \epsilon, a) \cup (b, b + \epsilon)) \). Using (35.22) for \( g_\epsilon \) replacing \( g \) and making \( \epsilon \downarrow 0 \), we obtain

\[
\lim_{n \to \infty} (F_n(b) - F_n(a)) = F(b) - F(a)
\]

in view of equationi (35.19) \(35.11\). This finishes the outline of the proof of Theorem 35.4.

### 35.9 The Fourier–Stieltjes transform

Given a function \( F \) of bounded variation \(35.12\) on the real line, its Fourier–Stieltjes transform is defined as

\[
\phi(x) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i x y} \, dF(y)
\]

This is essentially the same as the definition of the characteristic function of a random variable, except by tradition, one uses a minus sign in the exponent here. Assuming \( F(x) = \frac{1}{2} (F(x + 0) + F(x - 0)) \) for all \( x \) \(35.13\) we have

\[
F(x) - F(0) = \frac{1}{\sqrt{2\pi}} \lim_{\omega \to \infty} \int_{-\omega}^{\omega} \phi(\xi) \frac{e^{i \xi x} - 1}{i \xi} \, d\xi
\]

for all real \( x \). This is called the Fourier–Stieltjes inversion formula. This formula can be used to give a more direct proof of Theorem 35.4; see [21] Theorem XVI.4.24, Vol. II, p. 262, but the technical details are at a level more demanding than in our outline.

---

\(35.10\) That is, \( g \) is zero outside this interval.

\(35.11\) For drawing this conclusion, one needs to make use of the continuity of \( F \) at \( a \) and \( b \).

\(35.12\) A function of bounded variation is a function that can be written as the difference of two bounded monotonic functions. The definition usually given is different, but it can be shown to be equivalent to this description.

\(35.13\) See footnote \(35.11\) on p. 80 for the notation.
36 Normal approximation to the binomial distribution

Given that a binomial variable $X \sim \text{Bin}(n,p)$ is the sum of $n$ i.i.d. indicator random variables, as mentioned in Subsection 16.1, it satisfies the conditions of the Central Limit Theorem 35.1. Given that $E(X) = np$ and $V(X) = npq$, the corresponding standardized variable is $(X - np)/\sqrt{np(1 - p)}$. Hence, according to the Central Limit Theorem, we have

$$P \left( \frac{X - np}{\sqrt{np(1 - p)}} \leq t \right) \approx \Phi(t) \quad (t \in \mathbb{R}).$$

With $t = (x - np)/\sqrt{np(1 - p)}$, this gives

$$P \left( \frac{X - np}{\sqrt{np(1 - p)}} \leq \frac{x - np}{\sqrt{np(1 - p)}} \right) = \Phi \left( \frac{x - np}{\sqrt{np(1 - p)}} \right) \quad (x \in \mathbb{R}).$$

That is,

$$F_X(x) = P(X \leq x) \approx \Phi \left( \frac{x - np}{\sqrt{np(1 - p)}} \right) \quad (x \in \mathbb{R});$$

this holds because the inequality on the left-hand side of the previous formula is equivalent to $X \leq x$. This approximation is of practical use when $n \geq 30$. Given that $X \in [0, n]$, one would not use this approximation for $x$ outside this interval. For $x$ close to 0, better approximations are available, based on the Poisson distribution.

36.1 Continuity correction

The binomial variable $X \sim \text{Bin}(n,p)$ can assume only integer values, whereas the normal variable approximating it is a continuous variable. Given an integer $k$ with $0 \leq k < n$, the events $X \leq k$ and $X < k + 1$ are the same. So, when one wants to approximate the probability $P(X \leq k)$ by using formula equation (36.1), one gets a more accurate result if in this equation, instead of taking $x = k$, one takes $x = k + 1/2$. Doing so is called the continuity correction.

37 Chernoff bounds

Chernoff bounds were discovered by Herman Rubin, but named after Herman Chernoff. Given a random variable $X$, if its moment generating function $E(e^{tX})$ exists, its Chernoff bounds are obtained by applying Markov’s inequality given in Lemma 31.1 to the random variable $e^{tX}$. According to this, given a real number $a$, for all $t > 0$ we have

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}} = e^{-ta}M_X(t) \quad (t > 0)$$

and

$$P(X \leq a) = P(-X \geq -a) = P(e^{-tX} \geq e^{-ta}) \leq \frac{E(e^{-tX})}{e^{-ta}} = e^{-ta}M_X(-t) \quad (t > 0).$$

The second inequality can be restated by replacing $t$ with $-t$, so as to look similar to the first inequality: for all $t < 0$ we have

$$P(X \leq a) \leq \frac{E(e^{tX})}{e^{ta}} = e^{-ta}M_X(t) \quad (t < 0).$$
37.1 Chernoff bounds for the standard normal variable

Given that the moment generation function for a standard normal variable \(X\) is \(M_X(t) = e^{t^2/2}\) according to equation (33.7), equation (37.1) gives that

\[
P(X \geq a) \leq e^{-ta}M_X(t) = e^{t^2/2 - ta}
\]

for all \(t > 0\).

The value that minimizes the exponent \(t^2/2 - ta\) is \(t = a\); since we must have \(t > 0\), we can take \(t = a\) only if \(a > 0\). This gives

\[(37.3) \quad P(X \geq a) \leq e^{-a^2/2} \quad \text{for all } a > 0.\]

A similar calculation can be performed with equation (37.2). In this case we assume \(t < 0\) so we can minimize the exponent with \(t = a\) only in case \(a < 0\). We obtain the inequality

\[
P(X \leq a) \leq e^{-a^2/2} \quad \text{for all } a < 0.
\]

Since the standard normal distribution is symmetric about the origin, this bound is equivalent to the one in (37.3).

The bound above can easily be improved by taking a different approach. For \(a \geq 0\) we have

\[
P(X \geq a) = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-(a+t)^2/2} dt = e^{-a^2/2} \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-at - t^2/2} dt.
\]

The integral on the right-hand side can be estimated in two different ways. On the one hand, we have

\[
\int_0^\infty e^{-at - t^2/2} dt \leq \int_0^\infty e^{-t^2/2} dt = \sqrt{2\pi} \int_0^\infty f_X(t) dt = \sqrt{2\pi} \frac{1}{2},
\]

giving the estimate

\[
P(X \geq a) \leq \frac{1}{2} e^{-a^2/2}
\]

for all \(a \geq 0\), improving the bound in (37.3). For large values of \(a\) a much better improvement is possible, since for \(a > 0\) we have

\[
\int_0^\infty e^{-at - t^2/2} dt \leq \int_0^\infty e^{-at} dt = \frac{1}{a},
\]

giving the estimate

\[
P(X \geq a) \leq \frac{1}{\sqrt{2\pi a}} e^{-a^2/2}
\]

for all \(a > 0\).

37.2 Chernoff bounds for the Poisson distribution

Let \(X\) have a Poisson distribution with parameter \(\lambda(>0)\). We have

\[
M_X(t) = e^{(e^t - 1)\lambda}
\]

If \(a \leq 0\) then to minimize the right-hand side, the best we can do is to make \(t \downarrow 0\), which gives the trivial (and so useless) inequality \(P(X \geq a) \leq e^0 = 1.\)
according to equation (33.6). Hence, by inequality (37.1) we have
\[ P(X \geq k) \leq e^{\lambda(e^t-1)-kt} \quad (t > 0). \]
The minimum of the exponent occurs when \( e^t = k/\lambda \), i.e., when \( t = \log(k/\lambda) \) when it assumes the value \((k-\lambda)\log(\lambda/k)\). This gives
\[ P(X \geq k) \leq e^{k-\lambda} \left( \frac{\lambda}{k} \right)^k \quad (37.4) \]
A comparable bound that is slightly better for large values of \( k \) can be obtained by using a different approach. In fact, we have
\[
P(X \geq k) = \sum_{n=k}^{\infty} P(X = n) = e^{-\lambda} \sum_{n=k}^{\infty} \frac{\lambda^n}{n!} = e^{-\lambda} \frac{1}{(k-1)!} \int_{0}^{\lambda} e^x (\lambda - x)^k \, dx \\
= \frac{1}{(k-1)!} \int_{0}^{\lambda} e^{x-\lambda} (\lambda - x)^k \, dx = \frac{1}{(k-1)!} \int_{0}^{\lambda} e^{-t} t^k \, dt \\
\leq \frac{1}{(k-1)!} \int_{0}^{\lambda} t^k \, dt = \frac{\lambda^k}{k!} \sim \frac{1}{\sqrt{2\pi k}} \left( \frac{e\lambda}{k} \right)^k ;
\]
we will explain the details in this derivation next. First, the the symbol \( f(k) \sim g(k) \) means that \( \lim_{k \to \infty} f(k)/g(k) = 1 \), the last relation described by \( \sim \) holds according to Stirling’s formula; \( \text{37.3} \) we included this last relation only so as to be able to compare the present bound to the bound in formula (37.4). The third equation holds by the integral form of the remainder term of the Taylor series of \( e^x \) (see [13, equation (28) on p. 50, §24]). For the fourth equation, we moved \( e^{-\lambda} \) under the integral sign, and the fifth equation resulted the substitution \( t = \lambda - x \). The inequality holds since \( \lambda > 0 \), and so \( e^{-t} \leq 1 \) in the interval of integration. For \( k > \lambda \), this estimate is better than the bound given by (37.4).

37.3 Reading
[19 §§ 8.1–8.3, part of 8.5, pp. 394–405, 413].

37.4 Homework
[19 Chapter 8, pp. 424–426], Problems 1, 3, 1, 6, 14 24.

38 The Weierstrass approximation theorem: a proof using probability theory

The [Weierstrass] approximation theorem says that function continuous on a closed interval can be approximated arbitrarily well on that interval by polynomials. Later, we will precisely formulate
\[ \log x \text{ denotes the natural logarithm of } x; \text{ in mathematical writing, } \ln x \text{ is rarely used.} \]
\[ \text{37.3 Stirling’s formula says that} \]
\[ n! \sim \sqrt{2\pi n} \left( \frac{n}{e} \right)^n ; \]
see [1] Chapter 5, § 2.5, pp. 201–206].
this result, and give the proof found by Sergei Bernstein (see §2). The proof uses Chebyshev’s inequality (31.2) applied to a random variable $X(n, x)$ having binomial distribution $\text{Bin}(n, x)$ for integers $n > 0$ and reals $x$ with $0 \leq x \leq 1$. In our discussion in Section §16 we required $0 < x < 1$ (we used $p$ instead of $x$) but adding the cases $x = 0$ and $x = 1$ causes no difficulty, even though considering these cases was not important in the context of probability theory, since $X(0, n) = 0$ and $X(1, n) = n$ with probability 1. The important relations $\mathbb{E}(X(n, x)) = nx$, $\mathbb{V}(X(n, x)) = nx(1-x)$, and $\mathbb{P}(X(n, x) = k) = \binom{n}{k}x^k(1-x)^{n-k}$ for $0 \leq k \leq n$, remain valid also in cases $x = 0$ and $x = 1$, the last one with the special stipulation that we take $x^0 = 1$ in case $x = 0$ and $(1-x)^0 = 1$ in case $x = 1$. We will adopt this stipulation throughout this section.

### 38.1 Requirements from elementary analysis

The topic of elementary analysis is the same as basic calculus, but with rigorous proofs. We will use two important results in elementary analysis. The first one is asserts that continuous functions on a closed interval are bounded:

**Theorem 38.1.** Given a function $f$ that is continuous on a closed interval $[a, b]$, there is an $M$ such that $|f(x)| \leq M$ for all $x \in [a, b]$.

Such an $M$ is usually called a bound for the absolute value of $f$ on $[a, b]$, and $f$ itself can be described as bounded. Another important result we need is the following.

**Theorem 38.2.** Given a function $f$ that is continuous on a closed interval $[a, b]$, for every $\epsilon > 0$ there is a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ whenever $x, y \in [a, b]$ are such that $|x - y| < \delta$.

The property of $f$ described in this theorem is called uniform continuity, and $f$ having this property is called uniformly continuous. One can concisely formulate this result by saying that a function that is continuous on a closed interval is also uniformly continuous. The difference of the concept of continuity and uniform continuity may be a little hard to appreciate, so we make a detour to explain the difference in some detail.

#### 38.1.1 Continuity versus uniform continuity

Since the property of uniform continuity mentioned in Theorem 38.2 may not be familiar to some of the readers of these notes, we will make some effort to explain the difference between continuity and uniform continuity. Let $S$ be a set of real numbers,

**Definition 38.1.** The function $f : S \to \mathbb{R}$ is said to be continuous on $S$ if for every $x \in S$ and for every $\epsilon > 0$ there is a $\delta > 0$ such that for every $y \in S$, if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$.

**Note.** The above definition says that $f$ is continuous in $S$ at $x$ for every $x \in S$. Using logic notation, the function $f : S \to \mathbb{R}$ is said to be continuous on $S$ if

$$\forall x \in S)(\forall \epsilon > 0)(\exists \delta > 0)(\forall y \in S)(|x - y| < \delta \to |f(x) - f(y)| < \epsilon).$$

The first two quantifiers are interchangeable here, since two quantifiers of the same type (i.e., two universal quantifiers, or two existential quantifiers) are interchangeable, so we can write this also as

$$\forall \epsilon > 0)(\forall x \in S)(\exists \delta > 0)(\forall y \in S)(|x - y| < \delta \to |f(x) - f(y)| < \epsilon).$$

\(^{(38.1)}\) $0^0$ is usually undefined, for good reason, but such a stipulation is always adopted when discussing power series, for example.

\(^{(38.2)}\) The college course discussing elementary analysis is usually called Advanced Calculus. While the students reading these notes may not have taken Advanced Calculus, the explanations we give here should suffice.
**Definition 38.2.** The function $f : S \to \mathbb{R}$ is said to be uniformly continuous on $S$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that for every $x \in S$ and for every $y \in S$, if $|x - y| < \delta$ then $|f(x) - f(y)| < \epsilon$.

**Note.** Using logic notation, the function $f : S \to \mathbb{R}$ is said to be uniformly continuous on $S$ if

$$\forall \epsilon > 0 \exists \delta > 0 \forall x \in S \forall y \in S (|x - y| < \delta \to |f(x) - f(y)| < \epsilon).$$

The formal difference between continuity and uniform continuity is indicated by the different order of the second and third quantifiers in formulas (38.1) and (38.2); aside from this difference, the two formulas are identical. However, these quantifiers are of different type (one is a universal quantifier, the other is an existential quantifier), and so they are not interchangeable. Hence the meanings of these two formulas are different.

To explain the difference in a less formal way, in case of continuity, $\delta$ depends on the choice of $x$ as well as on $\epsilon$ (and, of course, on the function $f$ itself), whereas in case of uniform continuity, $\delta$ depends only on $\epsilon$ but not on $x$ (but it does depend on the function $f$ itself).

### 38.2 Statement and proof of the Weierstrass approximation theorem

We will formulate the result specifically for the interval $[0, 1]$.

**Theorem 38.3** (Weierstrass approximation theorem). Let $f$ be a function that is continuous on the interval $[0, 1]$, and let $\epsilon > 0$ be a real number. Then there is a polynomial $P$ such that $|f(x) - P(x)| < \epsilon$ for all $x \in [0, 1]$.

More specifically: writing

$$P_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1 - x)^{n-k} \quad (n > 0),$$

there is an integer $N > 0$ such that $|f(x) - P_n(x)| < \epsilon$ for all $x \in [0, 1]$ and for all $n \geq N$.

The statement in the second paragraph is due to Sergey Bernstein.

**Proof.** Let $M$ be a bound for $|f|$ in the interval $[0, 1]$ (see Theorem 38.1), that is, $M$ is such that $|f(x)| \leq M$ for all $x \in [0, 1]$. Further, let $\delta > 0$ be such that $|f(x) - f(y)| < \epsilon/2$ whenever $x, y \in [0, 1]$ and $|x - y| < \delta$.

Let $n > 0$ be an integer. Pick a fixed $x \in [0, 1]$, and let $k$ be an integer with $0 \leq k \leq n$. Then we have

$$\left| f(x) - f\left(\frac{k}{n}\right) \right| \leq 2M$$

and

$$\left| f(x) - f\left(\frac{k}{n}\right) \right| \leq \frac{\epsilon}{2} \quad \text{if} \quad |k - nx| < n\delta$$

according to the above estimates. Writing $X = X(n, x)$ for a random variable with $\text{Bin}(n, x)$ distribution, we have

$$P_n(x) = \sum_{k=0}^{n} f\left(\frac{n}{k}\right) P(X = k) \quad \text{and} \quad f(x) = \sum_{k=0}^{n} f(x) P(X = k).$$
Hence
\[|f(x) - P_n(x)| \leq \sum_{k=0}^{n} |f(x) - f\left(\frac{k}{n}\right)| P(X = k)\]
\[\leq 2M \sum_{k:0 \leq k \leq n, |k-nx| \geq n\delta} P(X = k) + \frac{\epsilon}{2} \sum_{k:0 \leq k \leq n, |k-nx| < n\delta} P(X = k)\]
\[= 2M P(|X - nx| \geq n\delta) + \frac{\epsilon}{2} P(|X - nx| < n\delta),\]
where the inequality follows from the above estimates for the change of \(f\). We have \(E(X) = nx\) and \(V(X) = nx(1-x) < n\). Hence, by Chebyshev’s inequality (31.2) we have
\[P(|X - nx| \geq n\delta) \leq \frac{1}{n\delta^2}.\]
This estimate together with the trivial estimate \(P(|X - nx| < n\delta) \leq 1\), gives
\[|f(x) - P_n(x)| \leq \frac{2M}{n\delta^2} + \frac{\epsilon}{2} \leq \frac{2M}{N\delta^2} + \frac{\epsilon}{2}\]
for \(n \geq N\). Choosing \(N > 4M/(\delta^2\epsilon)\), the right-hand side will be less than \(\epsilon\), completing the proof.

39 The strong law of large numbers

39.1 Kolmogorov’s inequality

Lemma 39.1 (Kolmogorov’s inequality). Let \(n > 0\) be an integer and let \(X_i\) for \(i = 1, 2, \ldots, n\) be independent random variables such that \(E(X_i) = 0\) and \(E(X_i^2) = V(X_i)\) exists. Let \(\epsilon > 0\) and let \(A\) be the event
\[A = \left\{ \omega \in \Omega : \sum_{i=1}^{k} X_i(\omega) \geq \epsilon \text{ for some } k \text{ with } 1 \leq k \leq n \right\}.\]
Then
\[P(A) \leq \frac{1}{\epsilon^2} \sum_{i=1}^{n} V(X_i).\]

Proof. Let \(A_k\) be the subset of \(A\) for which the inequality in the definition of \(A\) holds for the least \(k\). That is,
\[A_k = \left\{ \omega \in \Omega : \sum_{i=1}^{k} X_i(\omega) \geq \epsilon \text{ and } \sum_{i=1}^{l} X_i(\omega) < \epsilon \text{ for all } l \text{ with } 1 \leq l < k \right\}.\]
Then the sets \(A_k\) are disjoint and \(A = \bigcup A_k\).
Let \(I_k\) be the indicator variable of \(A_k\); that is,
\[I_k(\omega) = \begin{cases} 1 & \text{if } \omega \in A_k, \\ 0 & \text{if } \omega \notin A_k. \end{cases}\]
Given that $I_k$ depends only on the random variables $X_i$ for $i$ with $1 \leq i \leq k$, the variables $I_k$, $X_{k+1}, \ldots, X_n$ are independent. Further, note that if $1 \leq i \leq k$ and $k < j \leq n$ then the random variables $I_k X_i$ and $X_j$ are independent. Hence

\[ E\left( I_k \left( \sum_{i=1}^{n} X_i \right)^2 \right) = E\left( I_k \left( \sum_{i=1}^{k} X_i \right)^2 \right) + 2 \sum_{i=1}^{k} \sum_{j=k+1}^{n} E(I_k X_i X_j) \]

in the second inequality, the last term in the middle member was dropped, since it is nonnegative. Given that $I_k X_i$ and $X_j$ in the sum on the right-hand side are independent, we have $E(I_k X_i X_j) = E(I_k X_i) E(X_j) = 0$, since $E(X_j) = 0$. Further, by the definition of $I_k$ and $A_k$ we have

\[ I_k \left( \sum_{i=1}^{n} X_i \right)^2 \geq \epsilon^2 I_k, \]

the above inequality becomes

\[ E\left( I_k \left( \sum_{i=1}^{n} X_i \right)^2 \right) \geq \epsilon^2 E(I_k) = \epsilon^2 P(A_k). \]

Noting that $I_k$ for $k$ with $1 \leq k \leq n$ are indicator variables of pairwise disjoint events, we have

\[ 0 \leq \sum_{k=1}^{n} I_k \leq 1. \]

As $E(X_i) = 0$, we have $E(X_i)$; an analogous equation holds for any sum involving $X_i$. As $X_i$ are independent, the sum of the variances equal the variance of the sums. Hence, we can conclude that

\[ \sum_{i=1}^{n} V(X_i) = V\left( \sum_{i=1}^{n} X_i \right) = E\left( \left( \sum_{i=1}^{n} X_i \right)^2 \right) \]

\[ \geq \sum_{k=1}^{n} E\left( Y_k \left( \sum_{i=1}^{n} X_i \right)^2 \right) \geq \sum_{k=1}^{n} \epsilon^2 P(A_k) = \epsilon^2 P(A); \]

here the first inequality follows from (39.2), and second inequality follows from (39.1). The assertion of the lemma is equivalent to the inequality we obtained, completing the proof.

---

39.1 We are relying on the intuitive notion of independence saying that random variables are independent if they have no influence on one-another. According to this intuitive view, one would expect that if $n > 0$ is an integer and $X_i$ are independent random variables for $i$ with $1 \leq i \leq n$, $k$ is integer, and $f(x_1, x_2, \ldots, x_k)$ is a real-valued “nice” function of the real variables $x_1, x_2, \ldots, x_k$, then the random variables $f(X_1, X_2, \ldots, X_k)$ and $X_{k+1}, \ldots, X_n$ are independent. If one replaces “nice” with “Borel measurable,” then one obtains a rigorous result. See § 45, p. 194. We are not in a position to establish this result rigorously, nor are we in a position to explain the concept of Borel measurability beyond saying that it is a much less stringent requirement than, e.g., continuity.
39.2 The strong law of large numbers

Theorem 39.1 (The strong law of large numbers). Let \( m \) and \( \sigma \geq 0 \) be real numbers, and let \( X_1, X_2, X_3, \ldots \), be independent random variables with \( \text{E}(X_i) = m \) and \( \text{V}(X_i) = \sigma^2 \) for \( i \) with \( 1 \leq i < \infty \). Writing
\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i,
\]
we have
\[
P(\lim_{n \to \infty} \bar{X}_n = m) = 1.
\]

The only difference in the assumptions between the weak law of large numbers (Theorem 31.1) is that here we assume that the random variables \( X_i \) are independent, rather than pairwise independent, while the conclusion of the present result is much stronger, as we will explain. This is, however, not the real difference, since the strong law of large number can also be proved under the assumption of pairwise independence if we also assume that the random variables \( X_i \) are identically distributed; see [6]. We will return to a discussion of this difference below in Subsection 39.6.

39.3 The precise definition of convergence

In order to discuss the what convergence with probability 1 in equation (39.3) means, we need to review the precise definition of convergence of sequences.

In an informal discussion, one says that \( L \) is the limit of the sequence \( \{a_n\} \), in symbols \( L = \lim_{n \to \infty} a_n \), if for large \( n \) the number \( a_n \) is close to \( L \). The problem with this description is that the words “large” and “close” do not have clear mathematical meanings. To correct this deficiency, we will call the number \( n \) \( N \)-large if \( n \geq N \). Here \( N \) is usually thought of as a large (whatever that means) integer, but in fact no restriction needs to be put on \( N \), so \( N \) need not be large, nor an integer (but it must be a real number) for \( N \)-large to make sense. Similarly, for \( \epsilon > 0 \), we will say that the number \( L \) is \( \epsilon \)-close to the number \( x \) if \( |L - x| < \epsilon \). Here one usually thinks of \( \epsilon \) being small (whatever that means), so that \( \epsilon \)-close really means close, but this again not a requirement for \( \epsilon \)-close to make sense, and no restrictions other than \( \epsilon > 0 \) are put on \( \epsilon \).

This gives a clue as to how to make the definition of limit precise. What we want is to ensure that \( a_n \) is as close to \( L \) as we want by making sure that \( n \) is large enough. That is, \( L \) is called the limit of the sequence \( \{a_n\} \) if for every \( \epsilon > 0 \) we can find an \( N \) such that if \( n \) is \( N \)-large then \( a_n \) is \( \epsilon \)-close to \( L \).

This is almost the final form, though it assumes that a sequence can only have one limit (this is true, but it is better left as a statement to be proved than to include it in the definition), and then we need to write the meaning of \( N \)-large and \( \epsilon \)-close directly into the definition. That is,

Definition 39.1. We say that \( L \) is a limit of the sequence \( \{a_n\} \) if for every \( \epsilon > 0 \) there is an \( N \) such that \( |L - a_n| < \epsilon \) if \( n > N \).

39.4 Almost sure convergence, and witnessing its failure

The phrase “almost surely” in probability means an event that happens with probability 1; the phrase is abbreviated as a.s. That is, given a sequence of random variables \( Y_1, Y_2, Y_3, \ldots \) and a real number \( L \), to say that
\[
\lim_{n \to \infty} Y_n = L \quad \text{a. s.}
\]
means that \( P(\lim_{n \to \infty} Y_n = L) = 1 \). The best way to understand what this means is to understand how it can fail. Given a series \( \{a_n\} \) of numbers, for the convergence \( \lim_{n \to \infty} a_n = L \) to fail, there must be an \( \epsilon > 0 \) such that for each integer \( N > 0 \) there is an \( n > N \) such that \( |L - a_n| > \epsilon \). In this case, we say that \( \epsilon \) witnesses that failure of \( \lim_{n \to \infty} a_n = L \). Analogously to this, we will say:

**Definition 39.2.** Let \( \epsilon > 0 \), let \( Y_n \) be random variables for \( n > 0 \), and let \( L \) be a real number. The event

for every integer \( N > 0 \) there is an \( n > N \) such that \( |L - Y_n| \geq \epsilon \)

will be described as the event of \( \epsilon \) witnessing the failure of \( \lim_{n \to \infty} Y_n = L \). If this event has positive probability, we say that \( \epsilon \) witnesses the failure of \( \lim_{n \to \infty} Y_n = L \) with positive probability.

We have the following

**Lemma 39.2.** Let \( Y_n \) be random variables for \( n > 0 \), and let \( L \) be a real number. Assume that no \( \epsilon > 0 \) witnesses the failure of \( \lim_{n \to \infty} Y_n = L \) with positive probability. Then

\[
\lim_{n \to \infty} Y_n = L \quad \text{a. s.}
\]

**Proof.** For an integer \( k > 0 \), let \( E_k \) be the event that \( 1/k \) witnesses the failure of \( \lim_{n \to \infty} Y_n = L \). We have \( P(E_k) = 0 \) according to our assumption. Note that \( E_k \subset E_{k+1} \); hence the events \( E_k \setminus E_{k+1} \) are pairwise disjoint. Hence putting

\[
E \overset{def}{=} \bigcup_{k=1}^{\infty} E_k = \bigcup_{k=1}^{\infty} (E_k \setminus E_{k+1})
\]

has probability

\[
P(E) = \sum_{k=1}^{\infty} P(E_k \setminus E_{k+1}) \leq \sum_{k=1}^{\infty} P(E_k) = 0.
\]

Clearly, \( E \) is the event \( \lim_{n \to \infty} Y_n \) either does not exist or it exists but it is not equal to \( L \). This completes the proof. \( \square \)

**39.5 Proof of the strong law of large numbers**

With this preparation, we are ready to present the proof of Theorem 39.1.

**Proof of Theorem 39.1.** Without loss of generality, we may assume that \( E(X_i) = 0 \) for all \( i \). Indeed, if this is not the case we can replace \( X_i \) with \( X_i - E(X_i) \). That is, we will assume \( m = E(X_i) = 0 \) in what follows.

Let \( \epsilon > 0 \) be fixed; we are going to estimate the probability of \( \epsilon \) witnessing the failure of \( \lim_{n \to \infty} \bar{X}_n = 0 \). Let \( m > 0 \) be an integer, and let \( B_m \) be the event that

\[
|\bar{X}_n| \geq \epsilon \quad \text{for some } n \text{ with } 2^m < n \leq 2^{m+1}.
\]

and, further, let \( C_m \) be the event that

\[
\left| \sum_{i=1}^{k} X_i \right| \geq 2^m \epsilon \quad \text{for some } k \text{ with } 1 \leq k \leq 2^{m+1}.
\]


It is clear that $B_m \subset C_m$. Hence, using Kolmogorov’s inequality (Lemma 39.1), we have

$$P(B_m) \leq P(C_m) \leq \frac{1}{2^m \epsilon^2} \sum_{i=1}^{2^m+1} \sigma^2 = 2^{-m+1} \epsilon^{-2} \sigma^2.$$ 

Hence, given $N > 0$, noting that the events $B_m$ for $m \geq N$ are pairwise disjoint,

$$P\left(\bigcup_{m=N}^{\infty} B_m\right) = \epsilon^{-2} \sigma^2 \sum_{m=N}^{\infty} 2^{-m+1} = \epsilon^{-2} \sigma^2 2^{-N+1} \sum_{i=0}^{\infty} 2^{-i} = \epsilon^{-2} \sigma^2 2^{-N+2};$$

the event $E_N \overset{\text{def}}{=} \bigcup_{m=N}^{\infty} B_m$ described on the left-hand side is the event that

$$|\bar{X}_n| \geq \epsilon \quad \text{for some } n \text{ with } n > 2^N.$$ 

The event $\epsilon$ witnessing the failure of $\lim_{n \to \infty} \bar{X}_n = 0$ can be described as $\bigcap_{N=1}^{\infty} E_N$. This intersection must have probability $\leq P(E_N)$ for all $N > 0$, and so it must have probability 0. This completes the proof of the theorem. 

\[\square\]

### 39.6 Comparing the weak and strong laws

While the conclusion of the strong law of large numbers (Theorem 39.1) is much stronger than that of the weak law (Theorem 31.1), the assumptions about the random variables $X_i$, in the way we formulated these results, are the same in both theorems; so, what is the point of the weak law?

To explain these results, we simplified the assumptions we establish them, and the assumptions can be weakened in a way that the different applicability of these results is brought out. What can be changed is that we do not need to assume that the variances of the random variables $X_i$ are all the same. \[39.2\] Writing $\sigma_i^2$ for the variance of $X_i$, in order to derive the weak law, it is enough to assume that

\[\lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = 0.\]

The assumption classically used to derive the strong law requires more:

\[\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} \text{ converges}\]

(see \[5\] Theorems A and D, pp. 202 and 204)). \[39.3\]

\[39.2\] The assumption about their expectation is unimportant as long as the limit of $E(\bar{X}_i)$ exist, since it is harmless to assume that $E(X_i) = 0$, since $X_i$ can be replaced by $X_i - E(X_i)$, as we remarked at the beginning of the proof of Theorem 39.1 in Subsection 39.5.

\[39.3\] We are definitely not asserting that these conditions are necessary necessary to derive the conclusions of the weak and strong laws, respectively. In fact, the question of necessary conditions for these laws to hold does not seem reasonable without stating more precisely in what sense one wants necessary conditions. It may happen that $\bar{X}_n$ converges (in some sense) for some very specific reason, unexplainable by general assumptions.

107
39.7 An example where the weak law is applicable but the strong law is not

An example satisfying condition (39.4) but not condition (39.5) is the sequence for which $\sigma_1 = 0$ and $\sigma_n^2 = n / \log n$ for $n > 1$ (note that $\log 1 = 0$, and we want to avoid having 0 in the denominator). To see this, writing $\lfloor x \rfloor$ for the integer part of $x$, that is, for the largest integer $n$ such that $n \leq x$, as for condition (39.4) with this sequence, we have

$$\sum_{i=1}^{n} \frac{i}{\log i} \leq \sum_{i=1}^{\lfloor \sqrt{n} \rfloor} \frac{i}{\log \sqrt{n}} + \frac{2}{\log n} \sum_{i=1}^{n} \frac{i}{\log \sqrt{n}} \leq \frac{(1 + \lfloor \sqrt{n} \rfloor)\lfloor \sqrt{n} \rfloor}{2} + \frac{n(n+1)}{2 \log n} = o(n^2)$$

as $n \to \infty$. Here, after the first inequality symbol, the sum was broken up into parts. In the first sum, we omitted $\log i$ from the denominator; this increases the terms for $i \geq 3$, but not for $i = 2$. To compensate for this, we added an extra term for $i = 1$ (note that $2/\log 2 < 1 + 2$). In the second sum, $1/\log i$ was changed to the larger $1/\log \sqrt{n}$ in the given range of summation. After the second inequality, the sum formula of the arithmetic progression was used for the first sum, and in the second sum the range of the summation was extended. This shows that condition (39.4) is satisfied for the given sequence.

As for condition (39.5), we have

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} = \sum_{n=2}^{\infty} \frac{1}{n \log n},$$

and this series is divergent according to the integral test. Indeed, for $A > 2$, using the substitution $t = \log x$ we have

$$\int_{2}^{A} \frac{1}{x \log x} \, dx = \int_{\log 2}^{\log A} \frac{dt}{t} = \log t \bigg|_{\log 2}^{\log A} = \log \log A - \log \log 2 \to \infty$$

as $A \to \infty$. This shows that condition (39.5) is not satisfied for the given sequence.

40 Maxima examples

Consider the following problem: A pair of fair coins are tossed 1200 times. Let $X$ denote the number of times both coins come up head. Find the probability that $290 \leq X \leq 310$. In the solution of the problem, first note that $X$ has binomial distribution. That is,

$$P(X = k) = \binom{1200}{k} \left( \frac{1}{4} \right)^k \left( \frac{3}{4} \right)^{1200-k}.$$
Hence we have

\[
P(290 \leq X \leq 310) = \sum_{k=290}^{310} \binom{1200}{k} \left( \frac{1}{4} \right)^k \left( \frac{3}{4} \right)^{1200-k}.
\]

Before modern computers, this sum used to be approximated by taking the normal approximation to the binomial distribution. Today, it is easy to evaluate it with high precision on computers. For example, using the Maxima programming language, the following program evaluates this sum:

```maxima
1 kill(all);
2 linel : 60;
3 binprob(n, p, k) := float(binomial(n,k)*p^k*(1-p)^(n-k));
4 bindist(n, p, k, l) :=
5 do(
6   sum : 0,
7   for i : k step 1 thru l do(
8     term : binprob(n, p, i),
9     sum : float(sum + term)),
10   return(sum)
11 );
12 bindist(1200,1/4,290,310);
```

We will explain this program line by line. First, the numbers at the beginning of the line are not part of the program; they added so that we can explain what is being said in each line. First of all, line breaks in the program are just separators; they have the same meaning as spaces. Each command needs end with a semicolon.

The command `kill(all)` in the first line destroys the values of all the temporary variables in the memory of maxima, so as not to interfere with the current program. This is unnecessary if the file is run from the command line, but it is useful if the program is run as part of an open maxima session. The command on line 2 gives the value 60 to the variable `linel`, which stands for the maximum line length in the output. The colon : on the line indicates assignment. The purpose of limiting the line length of the output is to allow the inclusion of the output in the present manuscript. On line 3, the function `binprob(n,p,k)` with variables `n`, `p`, and `k` is defined. This calculates the probability of `X = k` for the variable `X` with `Bin(n,p)` distribution. The symbol `:=` indicates assignment to functions. The function `float` on the right-hand side of this assignment indicates that the calculation should use floating point numbers. Maxima is a computer algebra system usually doing exact calculations, with algebraic expressions. Such calculations can be very resource intensive, and they would be completely wasteful if we are only interested in numerical answers; using the function `float` prevents such waste. The function `binomial` calculates the binomial coefficient, the symbol asterisk * represents multiplication, and the symbol caret ^ stands for exponentiation. On line 4, the function `bindist` with four arguments are calculated. This sums up the probabilities of `P(X = i)` from `i = k` to `i = l` for the random variable `X` with `Bin(n,p)` distribution. The symbol := indicates assignment to functions. The function `float` on the right-hand side of this assignment indicates that the calculation should use floating point numbers. Maxima is a computer algebra system usually doing exact calculations, with algebraic expressions. Such calculations can be very resource intensive, and they would be completely wasteful if we are only interested in numerical answers; using the function `float` prevents such waste. The function `return` on line 11, the sum in formula (40.1) is evaluated.

---

40.1 See more about the Maxima programming language in [16, Section 1].

40.2 One can use a dollar sign $ instead, with a slightly different meaning. When a command line ends in a semicolon, in the output file an explanation is added as to what the command does; when the command end with a dollar sign, this explanation is suppressed. The output of the above program will be discussed below.

109
The output of this program

```
(%i1) batch("binomdist.max")
(%i2) kill(all)
(%o0) done
(%i1) linel:60
(%o1) 60
(%i2) binprob(n,p,k):=float(binomial(n,k)*p^k*(1-p)^(n-k))
   k
(%o2) binprob(n, p, k) := float(binomial(n, k) p
   n - k
   (1 - p) )
(%i3) bindist(n,p,k,l):=do
   (sum:0,
   for i from k thru l do
   (term:binprob(n,p,i),
   sum:float(sum+term)),return(sum))
(%o3) bindist(n, p, k, l) := do (sum : 0,
   for i from k thru l do (term : binprob(n, p, i),
   sum : float(sum + term)), return(sum))
(%i4) bindist(1200,1/4,290,310)
(%o4) 0.5160583906382258
(%o4) binomdist.max
```

Lines 1–5 describe the Maxima program. After this, the lines are labeled with (%i[number]) for input lines, and (%o[number]) for output lines, where [number] is a line number. The input line repeat the lines of the input, the output lines explain the meaning or the result of these input lines. Much of this is self-explanatory; on line 26, the value of the sum is given. On lines 6 and 27, the name of the input file is mentioned.

When one uses the normal approximation to the binomial distribution, one proceeds as follows. A Bin(n, p) distribution has expectation np and standard deviation $\sqrt{np(1-p)}$. Accordingly, writing $D(X)$ for the standard deviation of $X$, we have

$$E(X) = 1200 \cdot \frac{1}{4} = 300 \quad \text{and} \quad D(X) = \sqrt{1200 \cdot \frac{1}{4} \cdot \frac{3}{4}} = 15.$$ 

Thus, the distribution of $X$ can be approximated by a that of a normal variable $Y \sim \mathcal{N}(300, 15^2)$, i.e., by a normal variable $Y$ having expectation 300 and standard deviation 15. Writing

$$Z = \frac{Y - 300}{15},$$
Y is a standard normal variable. Hence, using continuity correction,

\[
P(290 \leq X \leq 310) \approx P(289.5 \leq Y \leq 310.5)
\]

\[
= P \left( \frac{289.5 - 300}{15} \leq \frac{Y - 300}{15} \leq \frac{310.5 - 300}{15} \right)
= P \left( -0.7 \leq Z \leq 0.7 \right)
= P(Z \leq 0.7) - P(Z < -0.7) = P(Z \leq 0.7) - P(Z > 0.7)
= P(Z \leq 0.7) - (1 - P(Z \leq 0.7)) = 2P(Z \leq 0.7) - 1 = 2\Phi(0.7) - 1;
\]

here, the \( \Phi \) denotes the distribution function of the standard normal distribution. The first equality in the last line follows since the distribution of \( Z \) is symmetric about 0. We can use the normal distribution table to find that \( \Phi(0.7) \approx 0.7580 \). Thus, we gave

\[
P(290 \leq X \leq 310) \approx 0.5160.
\]

Instead of using the table of the normal distribution, one can use Maxima to find this value. The following program accomplishes this:

```maxima
1 kill ( all );
2 load ( distrib );
3 normval : float ( cdf_normal (0.7, 0, 1));
4 prob : 2*normval -1;
```

On line 2, one loads the package `distrib` containing various probability distributions. On line 3, the cumulative distribution function \( \text{cdf}_\text{normal}(x,m,s) \) of a normal variable \( Y \) with mean \( m \) is used with \( m = 0 \) and \( s = 1 \) is used; this shows that we do not need to standardize the variable to use Maxima. On line 4, the value \( 2\Phi(0.7) - 1 \) is given. Here is the output of the program.

```
(%i1) batch ("normal.max")

10 (%i2) kill(all)
11 (%i0) done
12 (%i1) load(distrib)
13 (%i1) /usr/share/maxima/5.38.1/share/distrib/distrib.mac
14 (%i2) normval:float(cdf_normal(0.7,0,1))
15 (%o2) 0.758036347776927
16 (%i3) prob:2*normval-1
17 (%o3) 0.5160726955538539
18 (%o3) normal.max
```

The result on line 18 agrees up to four decimals with the more accurate result obtained above by summing the probabilities of the binomial distribution above.

Instead of using the distribution package in Maxima, one can use numerical integration to calculate the value of the normal distribution:
Here lines 3-5 use a numerical integration program to calculate the value of the standard normal distribution. The method is based on the Gauss–Kronrod quadrature formula implemented by the Quadpack package in Maxima. On line 4 the function, the variable of integration \( t \), the limits 0 and \( x \), and the parameter key is used. The value 6 specified for the key selects the specific Gauss–Kronrod integration rule to be used. The variables \( \text{epsrel} \) and \( \text{epsabs} \) given on line 4 are the desirable limits for the relative and absolute errors for the approximation, and the \( \text{limit} \) on line 5 is the maximum number of subintervals used in the integration. The output of the of the program is a list of several members. On line 5 the argument \([1]\) asks only for the first member of this list, which is the value of the integral. The output of this program is

```
10 (%%2) kill(all)
11 (%00)         done
12 (%%1) func:1/sqrt(2*%pi)*exp((-t^2)/2)
13                2
14                t
15                -- 
16                2
17 (%%1) "------------------
18            sqrt(2) * sqrt(%pi)
19 (%%2) normdist(x):=1/2
20       +quad_qag(func,t,0,x,6,'epsrel =
21         1.0E-13,'epsabs = 1.0E-15,
22            'limit = 1300)[1]
23         1
24       6,
25 (%%2) normdist(x) := - + subscript(quad_qag(func, t, 0, x,
26           2)
27         'epsrel = 1.0E-13, 'epsabs =
28         1.0E-15, 'limit = 1300), 1)
29 (%%3) normval: normdist(0.7)
30 (%%3) 0.758036347776927
31 (%%4) prob:2*normval-1
32 (%%4) 0.5160726955538539
33 (%%4) norm_integr.max
```

Maxima 5.38.1 http://maxima.sourceforge.net
using Lisp GNU Common Lisp (GCL) GCL 2.6.12
Distributed under the GNU Public License. See the file COPYING.
Dedicated to the memory of William Schelter.
The function bug_report() provides bug reporting information.
(%i1) batch("norm_integr.max")
(%o0) done
(%i1) func:1/sqrt(2*%pi)*exp((-t^2)/2)
(%o1) 2
(%i2) normdist(x):=1/2+quad_qag(func,t,0,x,6,'epsrel =
epsabs = 1e-15,
'limit =1300)[1];
(%o2) normdist(x) := - + subscript(quad_qag(func, t, 0, x,
2) 'epsrel = 1.0E-13, 'epsabs =
1.0E-15, 'limit = 1300), 1)
(%i3) normval : normdist(.7);
(%o3) 0.758036347776927
(%i4) prob : 2*normval-1;
(%o4) 0.5160726955538539
(%o4) norm_integr.max
Note that the numerical integration program gives exactly the same value given by the distrib package.

References


