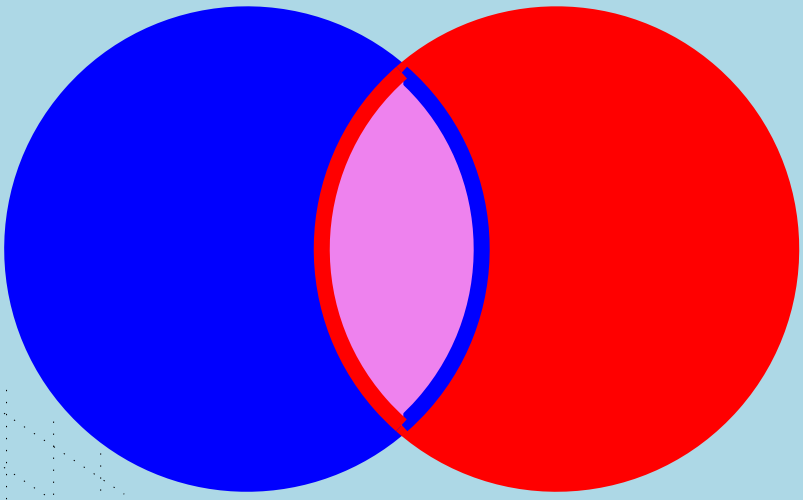
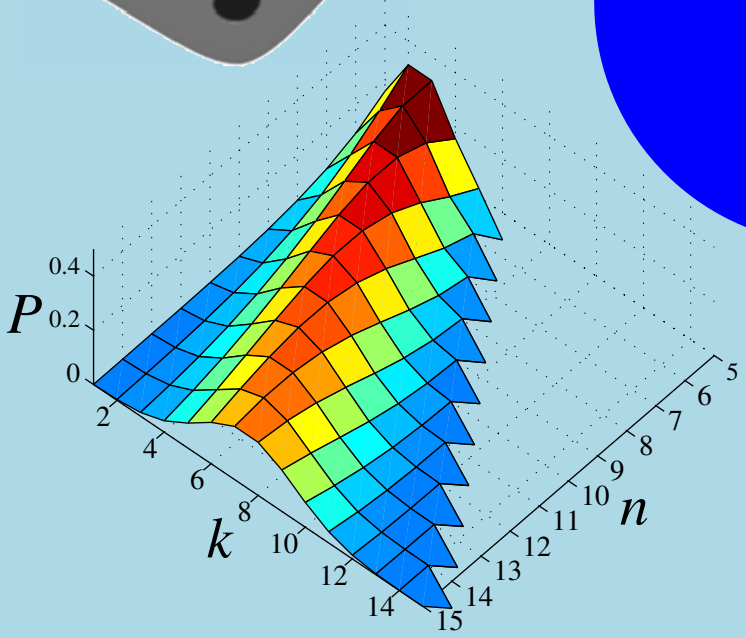
Introduction to the Probability Theory

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$





$$P(k) = C_k^n p^k (1-p)^{n-k}$$

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

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

Taha Sochi

Preface

This book originates from a collection of notes and solved problems which I created sometime ago for my personal use. However, I decided recently to improve and expand this collection and put it in the public domain for the common benefit. I also attached to the solved problems proposed exercises which mostly follow in their style and method of solution the corresponding problems or are based in their content and objective on these problems. I also added computer codes (in C++ language) to some of these problems for the purpose of calculation, test and simulation (as well as for the sake of demonstration and enhancement of the learning process).[1]

I used illustrations (such as figures and tables) when necessary or appropriate to enhance clarity and improve understanding. I also followed in many cases intuitive arguments and methods to make the notes and solutions look natural and instinctive (noting that the theory of probability at its basic level is largely based on common sense and intuition). Like my previous books, maximum clarity is one of the main objectives and criteria in determining the style of writing, presenting and structuring the book as well as selecting its items and contents.

However, the reader should also notice that the book, in most parts, does not go beyond the basic probability and hence most subjects are presented and treated at their basic level. The reader is advised to consult the table of content and the index to have a feeling about the content and substance of this book.

A rather modest mathematical background knowledge is required for digesting and understanding the book (or at least most of its contents). In fact, the book in most parts requires no more than a college or secondary school level of general mathematics. So, the intended readers of the book are primarily college (or A-level) students as well as junior undergraduate students (e.g. in mathematics or science or engineering).

The book can be used as a text or as a reference for an introductory course on this subject and may also be used for general reading in mathematics. The book may also be adopted as a source of pedagogical materials which can supplement, for instance, tutorial sessions (e.g. in undergraduate courses on mathematics or science).

An interesting feature of this book is that it is written and designed, in part, to address practical calculational issues (e.g. through sample codes and suggested methods of solution) and hence it is especially useful to those who are interested in the calculational applications of the probability theory (i.e. applied "mathematicians" working in the field of probability). Other practice-oriented features (such as the proposed exercises which we indicated above) also make the book more useful from the perspective of practice and application of the probability theory and related subjects.

Taha Sochi
London, January 2023

[1] The number of computer codes that accompany this book is 31; 7 of which are for simulation and the rest are for other purposes (mainly calculation). These codes are available from my personal website: https://tahasochi.com/blog/. They are also available from ResearchGate https://www.researchgate.net/publication/368242717_ProbabilityCodes.

Contents

Nomenclature

In the following list, we define the common symbols, notations and abbreviations which are used in the book as a quick reference for the reader.

$\forall$	for all
$\{\cdots\}$	set
$\varnothing$, $\{\}$	empty (or null) set
!	factorial
$\cap$	intersection of sets
$\cup$	union of sets
$\in$	in (or belong to)
$\notin$	not in
$\subset$	subset
$\not\subset$	not subset
$\supset$	super set
$\subseteq$	subset in general (if $\subset$ is restricted to proper subset)
$\supseteq$	super set in general (if $\supset$ is restricted to proper super set)
$A, B, C, \ldots$	sets
$\overline{A}$	complement of set A
AND	logical **AND** operator
$\mathbb{C}$	the set of complex numbers
C_m^n	number of combinations of m in n with no repetition (binomial coefficient)
$C_{n,m}$	number of combinations of m in n with repetition
$C_{n_1,\ldots,n_k}^n$	multinomial coefficient
$\mathrm{Cor}(x, y)$	correlation of two random variables x and y
$\mathrm{Cov}(x, y)$	covariance of two random variables x and y
Eq., Eqs.	Equation, Equations
erf	error function
f	function (density function)
F	cumulative distribution function
H	head (of coin)
iff	if and only if
nD	n-dimensional (e.g. 1D)
$\mathbb{N}$	the set of natural numbers $(1, 2, 3, \ldots)$
N_A	number of elements of set A (a subset of the sample space)
N_S	number of elements of the sample space
OR	logical (non-exclusive) **OR** operator
p, P	probability
$P(A)$	probability of event A
$P(A \cap B)$	probability of "A AND B"
$P(A \cup B)$	probability of "A OR B"
$P(A \,\vert\, B)$	conditional probability (i.e. probability of A given B)
P_m^n	number of permutations of m in n with no repetition
$P_{n,m}$	number of permutations of m in n with repetition
$\mathbb{Q}$	the set of rational numbers
$\mathbb{R}$	the set of real numbers
S	sample space
T	tail (of coin)
U	universal set
V	variance
x, X	variable (random variable)

$\mathbb{Z}$	the set of integers
α	parameter of the exponential distribution
$\Gamma(a)$	the (complete) gamma function
$\Gamma(a, x)$	the (incomplete) gamma function
λ	Poisson parameter
μ, μ_x	arithmetic mean (or average or expectation value)
μ_G	geometric mean
μ_H	harmonic mean
Π	product symbol
σ	standard deviation
Σ	summation symbol

Chapter 1
Preliminaries

In this chapter we present and discuss a number of subjects and issues about probability and related matters which are generally needed for understanding and appreciating the materials presented in the subsequent chapters as well as avoiding potential sources of misunderstanding and confusion.

1.1 General Remarks

In this section we present a number of general remarks (related to the conventions, terminology and commonly occurring issues in this book) outlined in the following points:

1. All numbers and variables in this book are real (i.e. not complex or imaginary).
2. It is common in the theory of probability to use the notions and notations of sets, and in this book we follow this tradition. Accordingly, the familiarity of the readers with the basics of set theory is a prerequisite. Therefore, we provide in § 2.1 some preliminary materials about set theory for completeness and reference. So, those who are familiar with the basics of set theory can skip § 2.1 (although it is useful to read it as revision and reminder). We also note that the required notation of sets is mostly given in the Nomenclature.
3. In this book we deal with only univariate probability problems and issues although there are a few exceptions.
4. We attach to the solved problems proposed exercises (which we label with **PE**). These exercises can (in most cases) be solved rather easily by following the method or content of solution in the related solved problem or consulting the given notes (i.e. in the main text). The purpose of these exercises is to reinforce understanding and provide more practicing opportunities to those who are keen to learn by practice and action (although they are used occasionally to draw the attention to related issues).
5. The calculation and simulation codes that associate this book (as mentioned in the Preface) satisfy no more than the bare minimum of requirements for achieving the intended purposes and objectives, and hence common programming measures and standard practices in coding are largely and deliberately avoided and ignored. In addition to achieving their basic functionality, these codes are meant to give an idea about how to calculate or simulate the problems in question and to verify the stated results, and hence any extra elaboration or expansion has no benefit and could also be confusing and distracting. These codes are written in C++ programming language and compiled and run successfully on Microsoft Visual C++ 6.0 and Dev-C++ version 4.9.9.2 (as well as other versions) on Windows platforms (XP and 8.1). As indicated already, these codes are generally divided (with regard to their functionality and objectives) to two main categories: calculation and simulation. Also see the Preface and the sections and parts related to simulation and calculation (e.g. § 1.6 and § 2.3.3).
6. "Random" (as well as similar words like "randomly" or "at random") which is used frequently in this book to characterize probabilistic selection (or sampling) processes means there is no bias in the selection process (i.e. the selection is fair) and hence every potential candidate has the same chance of being selected in this process.

1.2 Initial and Ultimate Probabilities

It is important to note that in the theory of probability we have two types of probability: initial (or simple) and ultimate (or composite). In fact, initial probabilities are needed by the probability theory as given conditions or inputs for calculating and obtaining ultimate probabilities. For example, in the event of throwing a fair die consecutively we have an initial probability of 1/6 for getting each one of the six faces, and we have an ultimate probability of 1/36 for getting face "1" in two consecutive trials. In this

case, the ultimate probability of 1/36 is obtained from the initial probability of 1/6 in the two trials by multiplication, i.e. $1/6 \times 1/6 = 1/36$.[2]

Now, someone may ask: how do we get these initial probabilities? However, before we answer this question we would like to remark that the probability theory is not concerned with these initial probabilities because it takes them as given initial conditions (or assumptions or hypotheses or ... etc.). In other words, provided we are given these initial probabilities (and provided these initial probabilities meet the required conditions set by the postulates of the probability theory), the probability theory determines the ultimate probabilities. In fact, we can imagine the probability theory as a machine that takes these initial probabilities as an input (or raw material) and it processes them to produce the ultimate probabilities as an output (or final product) whose nature and characteristics depend on the demand and conditions presumed in the given problem.

Returning to the question of how do we get the initial probabilities, we can say that there are different ways for getting them. Moreover, the method of getting them is case-dependent. For example, in some cases the initial probabilities are obtained experimentally (e.g. by tossing a coin many times to get the probability of getting head and tail for that coin; see for instance Problem 17 of § 3.2), while in other cases we may get them by reasoning (or guess or intuition or logic or ... etc.) where we may use rational thinking and theoretical argumentation. In fact, in some cases they may even be postulated and the results (i.e. the ultimate probabilities obtained from these postulated initial probabilities according to the probability theory) are then verified and validated experimentally or theoretically. In other words, the initial probabilities may be treated as modeling parameters and conditions that can be estimated, adjusted and tuned to get the best results.

So in brief, in this book (as well as in similar texts in the literature of probability theory) we have no interest in how initial probabilities are obtained. Our concern is on how to obtain the required ultimate probabilities from the given initial probabilities with the help of the theory of probability in conjunction with the demand and conditions of the given problem. We should finally note that in many cases the initial probabilities are not given explicitly and directly but they are given implicitly and indirectly in the form of a count or a hint or something else and hence we need to infer or guess or calculate them from the given data and information.

1.3 Subjective and Objective Probabilities

"Probability" may be used in two main meanings: subjective probability and objective probability. **Subjective probability** is about the personal judgment and feeling of an observer about a certain event. For example, when I say "Tom is probably sick today" or "Mary is probably a good friend of mine" or "John is probably a bad guy" it is essentially about my personal judgment and feeling about the health of Tom, the friendship of Mary and the characters of John. **Objective probability**, on the other hand, is about objective reality when this reality can be in one form or another regardless of the observer (or presumably so). In fact, we can distinguish between these two meanings of probability more technically and specifically and in more details by the following points:

• Subjective probability is about the personal judgment and feeling of an observer towards a *definite reality*, while objective probability is about a *dubious reality* in itself and regardless of any observer.[3]

• Objective probability is based on the notion of *outcome of a repetitive trial* (where the trial or/and its repetition could be hypothetical and where the outcome of the individual trials is not unique), while subjective probability is not since it deals with a unique and determined reality (rather than repetitive trial and non-unique outcome).

• Objective probability is strictly measurable and quantifiable according to well-known and well-defined

[2] In more technical terms (which will be discussed and explained later) the initial probabilities are the probabilities given to the points of the sample space. However, we should note that the given example is meant to demonstrate and clarify the idea of initial and ultimate probabilities rather than reflect the technicalities of this issue. In fact, whether or not 1/36 is an ultimate probability depends on the modeling of the problem and the setting and selection of its sample space.

[3] With regard to the conditional probability (and its consequences like the Bayes rule) the reality is dubious considering the hypothetical repetitive trials criterion.

rules[4] (i.e. the rules of probability theory which is the subject of this book),[5] while subjective probability is not. In fact, subjective probability is subject to psychological factors and personal considerations which are not measurable or quantifiable or predictable, and hence it varies from one person to another. For example, in my opinion "Mark is probably a good guy" but this may not be the opinion of my friend. Similarly, my (subjective) probability about the sickness of Sarah may be 50% but it could be 30% according to the (subjective) probability of my brother. On the other hand, objective probability is subject to realistic considerations and factual factors which are independent of any observer (or supposedly so). For example, the (objective) probability of getting head when we toss a fair coin is 1/2 with respect to any observer because this probability is based on realistic factors (e.g. determined by real life experiments) and hence it is independent of the observer and observation.

• As indicated earlier, objective probability is the subject of the mathematical theory of probability, while subjective probability is not. So in brief, probability theory is about objective probability not subjective probability.

In this regard it is useful to quote Feller about this issue (see the first volume of Feller book in the References): The success of the modern mathematical theory of probability is bought at a price: the theory is limited to one particular aspect of "chance." The intuitive notion of probability is connected with inductive reasoning and with judgments such as "Paul is probably a happy man," "Probably this book will be a failure," "Fermat's conjecture is probably false." Judgments of this sort are of interest to the philosopher and the logician, and they are a legitimate object of a mathematical theory. It must be understood, however, that we are concerned not with modes of inductive reasoning but with something that might be called physical or *statistical probability*. In a rough way we may characterize this concept by saying that our probabilities do not refer to judgments but to possible outcomes of a *conceptual experiment*. (End of quote noting that italicization is from Feller)

We should finally note that although we (for the sake of clarity) defined subjective probability in a way that makes it look different in subjects and instances from objective probability, subjective probability can also be recognized and attributed to subjects and instances that primarily belong to objective probability. For example, based on the factual factors the objective probability of getting a tail when tossing a fair coin is 1/2. However, someone (due to wrong conviction or poor arithmetic) may believe that this probability is 1/3 and hence his (subjective) probability in this case is 1/3. Anyway, it should be obvious that when we talk about probability in this book it should be understood to be about objective probability unless it is stated or indicated otherwise.

1.4 Relevant and Irrelevant Factors

In any particular probability problem there are certain considerations and factors that are relevant to our probability investigation and according to which this investigation should be decided and directed. Any other consideration or factor should be discarded and ignored in that investigation although it could be relevant or important to other investigations. For example, if we are interested in the gender of newborn babies then what is relevant in this investigation is being male or female regardless of being (for instance) white or black and healthy or unhealthy. Accordingly, the probabilities will be exclusively determined by the gender factor with no involvement of color or health or weight or anything else. So, if we study a sample of these babies (say the ones born in a given hospital in a given month) then we should classify them as males or females rather than (for instance) males or white females or black females. This is important especially when determining the sample space and assigning initial (or simple) probabilities to the sample points (as will be clarified more later on). So in brief, any factors and considerations that occur accidentally in the probabilistic situation under investigation should be excluded and ignored within that investigation although they could be relevant and important in another investigation.

[4] Being subject to well-known and well-defined rules does not rule out the possibility of disagreements and disputes about certain aspects and at some occasions (some of which will be indicated or investigated later).

[5] We may need to extend "the rules of probability theory" to include some rules. requirements and conditions related to initial probabilities and how they should be determined.

1.5 Factors Affecting Selection and Sampling

In this section we investigate briefly factors and considerations that affect selection and sampling and hence they have an impact on the results of counting and probability. These factors and considerations occur frequently in the investigations of probability theory (including this book) and hence the reader must be aware of them and should understand and appreciate their significance and impact on counting and probability.

We note that despite the fact that counting and probability at their basic level are generally based on (and compliant with) natural intuition and common sense, these factors and considerations (as well as other factors and considerations) can make the business of counting and probability very tricky and messy and can lead to situations and circumstances incompatible with common sense and may even be counter-intuitive. So, these factors and considerations should be investigated carefully and understood and appreciated very well to avoid confusion and mistakes.

We should also note that these factors and considerations are generally dictated and determined by the nature of the problem in question and hence they are either stated explicitly or indicated implicitly but unambiguously. However, sometimes they are not stated or indicated so and hence it is up to us (i.e. whoever tackles and solves the problem) to decide about these factors; in which case considerations like common sense, convenience, context, circumstances and objectives should be taken into account to decide about these factors and make an appropriate choice about them.

We finally note that although these factors and considerations are (rather) rarely stated explicitly (and as such) they are present in most counting and probability problems (usually in the background of these problems and within the intended context) and hence when we tackle these problems we should always take account of them in our analysis and solution.

1.5.1 Distinguishability

One of the main factors that affect counting (and hence probability) is the distinguishability and indistinguishability in the selection and sampling process.[6] For example, if we want to choose 3 cars out of 5 cars then the result of this selection (as determined for instance by how the selected 3 cars will look) will be different if the 5 cars are distinguishable from each other (e.g. by having different colors) or not distinguishable (e.g. by having the same color). More specifically, if all the 5 cars are black then we have only one possibility for this selection with regard to the color (i.e. all the selected 3 cars are black), while if the 5 cars are different in color from each other then we have more than one possibility for this selection with regard to the color (e.g. black-red-blue, blue-red-green, yellow-green-red, etc.).

In fact, we usually meet two main types of distinguishability/indistinguishability in the selection process that determine the results of counting and probability:

• **Distinguishability of objects**, i.e. whether or not the objects in the selection process can be (or are wanted to be)[7] distinguished from each other (with regard to the aspect of concern and the property of interest in that particular situation and context). An instance of this type of distinguishability is the aforementioned cars example.

• **Distinguishability of arrangements**, i.e. whether or not the arrangements and configurations of the selected objects in the selection process can be (or wanted to be) distinguished from each other (with regard to the aspect of concern and the property of interest in that particular situation and context). For example, if we have 5 cars of different colors and we want to select 3 of them (taking into account their distinguishability in color) then we either consider the yellow-green-red selection as a single arrangement (ignoring the order of the colored cars) or we consider the yellow-green-red selection as one arrangement

[6] We note that "selection and sampling process" refers to a more general meaning than what it suggests initially and primarily. So, it includes for example (beside selection and sampling) things like "expectation from a random experiment" and "assignment of states to objects" (among many other things).

[7] This is to indicate that in some cases and situations the objects can be distinguished but we deliberately ignore their distinction and treat them as identical and indistinguishable because their distinction is irrelevant or not under consideration or not of importance to us.

among other similar arrangements involving these colors such as green-red-yellow (considering the order of the colored cars).

So, in any selection process we should be aware of distinguishability/indistinguishability of the involved objects or/and the required arrangements, and accordingly we should consider and determine whether or not distinguishability is relevant (or possible or wanted or ... etc.) in that situation and problem before making any decision about how to manage the situation and tackle the problem. Otherwise, serious mistakes could be committed in solving counting and probability problems.

It is worth noting that the distinguishability of arrangements generally depends on the distinguishability of objects (as well as on other factors and considerations which affect their distinguishability such as order in space or time or rank). We should also note that when objects and arrangements are classified as indistinguishable there is no difference between them being really indistinguishable (because we are not able to distinguish between them) and being distinguishable but we do not want to distinguish between them (and hence we treat them as indistinguishable).

1.5.2 Replacement

When we make a succession of selections (by taking an object or objects from a collection of objects in each selection) then we have two possibilities: either we return the selected object(s) to the collection before we make the next selection or not. The first case is commonly called selection with "replacement". It is obvious that the result of the next selection depends on the type of the previous selection(s) and whether it is with or without replacement. This is because the collection in the case of replacement is not the same as the collection in the case of no replacement since the collection in the former case includes the object(s) selected in the previous selection(s).

For example, let have a box containing a collection of 10 balls 7 of which are black and the remaining 3 are red, and suppose that we made a first selection by taking the 3 red balls out of the box. Now, if we make next a second selection then if we replace the chosen 3 balls back into the box (i.e. before making the second selection) then our collection of choice in the second selection will be the original 10 balls (7 black and 3 red) and hence we have an opportunity to have red balls in the second selection, but if we do not replace the chosen 3 balls back into the box (i.e. before making the second selection) then our collection of choice in the second selection will be the remaining 7 black balls and hence we do not have an opportunity to have red balls in the second selection.

So, in the problems of counting and probability we should always be aware (and take account) of the type of selections that we make and if they are supposed to be with or without replacement so that the choices that we make and the probabilities that we estimate are correct and compliant with the conditions and requirements of the problem in hand. However, it should be noted that in many cases and situations the condition of "with or without replacement" may not be stated explicitly but it can be understood implicitly and inferred from the contexts and circumstances. We should also note that replacement usually apply to physical objects but not to non-physical (or abstract) objects, e.g. we can replace cards in a deck of cards or balls in a collection of balls but not letters in the alphabet or numbers in a set of numbers. However, in the case of dealing with non-physical objects repetition/no-repetition could take the role of replacement/no-replacement (see § 1.5.3).

1.5.3 Repetition

When we make a selection from a collection of objects we have two main cases with regard to the possibility of selecting identical objects, i.e. selection of identical objects is either allowed or not. The former case is commonly described as selection with repetition while the latter case is described as selection without repetition. For example, when we choose letters from the English alphabet to build words and sentences we can select any letter more than once, e.g. 's' and 'o' are used 2 times each and 'e' is used 3 times for building the sentence "these are my books" and hence this is an instance of selection with repetition. On the other hand, if we have 26 stickers labeled uniquely with the 26 English alphabet letters and we want to build a sentence from these stickers then our choice of words and sentences is limited by the condition

that our selection of letters should be without repetition because no sticker can be used more than once, and accordingly we can build the sentence "he is not bad" but not the sentence "he is good".

1.5.4 Order

Another factor that determines the type and nature of selection is the order of the selected objects. So, in one case the order of the selected objects is significant and in another case it is not. This factor affects the type of selection and the number of possibilities of the available choices. For example, let have a bag containing 9 balls numbered from 1 to 9 and we want to draw 3 balls out of this bag. **In one case**, we take 3 balls at once and in one go and hence there is no order between the selected balls. So, if the 3 selected balls are the ones numbered with $2, 5, 9$ then we have just one possibility for this selection since $2, 5, 9$ and $5, 2, 9$ are the same because order is not a factor in this selection. However, **in another case** we take the 3 balls one at a time considering their order in selection and hence the possibility $2, 5, 9$ (which means the first ball is number 2, the second ball is number 5 and the third ball is number 9) is different from the possibility $5, 2, 9$ (which means the first ball is number 5, the second ball is number 2 and the third ball is number 9). So, it is important to be aware of this factor and its impact on the type of selection and the number of available possibilities and choices.

Problems[8]

1. Let have a bag containing 10 balls numbered from 1 to 10, and assume that we randomly selected a sample of 5 balls from this bag. Identify different types of sampling the 5 balls considering some of the factors investigated in this section (i.e. § 1.5).
   Answer: For example, we may consider the factors of replacement, repetition and order (noting that "10 balls numbered from 1 to 10" indicates that the balls are distinguishable). Accordingly, we can identify (at least) 4 types of sampling in this case, that is:
   (**a**) We may draw the 5 balls one after one without returning any one of the drawn balls to the bag before drawing the next ball. In this case we obviously have order. Moreover, by assumption there is no replacement and hence there is no possibility of repetition (e.g. if ball number 3 is drawn in the first draw then there is no possibility for this ball to appear again in the remaining 4 draws). Therefore, we can label this type of sampling as **sampling with order and without replacement/repetition**.
   (**b**) We may draw the 5 balls one after one but we return the drawn ball to the bag before drawing the next ball. In this case we obviously have order. Moreover, by assumption there is replacement and hence there is a possibility of repetition (e.g. if ball number 3 is drawn in the first draw then there is a possibility for this ball to appear again in the remaining 4 draws). Therefore, we can label this type of sampling as **sampling with order and with replacement/repetition**.
   (**c**) We may draw the 5 balls at once and in one go. In this case there is no order in this type of sampling (since the 5 balls are drawn simultaneously). Moreover, there is no possibility of replacement/repetition (for the same reason). Therefore, we can label this type of sampling as **sampling without order and without replacement/repetition**.
   (**d**) We may draw the 5 balls one after one and we return each drawn ball to the bag before drawing the next ball although we do not consider the order (e.g. the selection $2, 5, 3, 9, 7$ is considered the same as $7, 2, 9, 3, 5$). In this case we obviously have no order (i.e. it is irrelevant) but we have replacement and hence there is a possibility of repetition. Therefore, we can label this type of sampling as **sampling without order and with replacement/repetition**.
   PE: Discuss the issue of distinguishability/indistinguishability of objects and arrangements in the context of this Problem.
2. Repeat Problem 1 assuming this time that the 10 balls are indistinguishable (such as by numbers or colors or anything else).
   Answer: As the balls are indistinguishable, there is no meaning of order or repetition. Therefore, we can distinguish only between sampling with replacement and sampling without replacement. However, replacement can distinguish between the two sampling processes but not between the two samples

[8] We note that these Problems belong to § 1.5 as a whole rather than to the present subsection.

obtained in these processes. So in brief, we cannot distinguish between the 5-ball samples obtained in any type of sampling. Also see Problem 39 of § 2.2.
PE: What if 5 of the 10 balls in this Problem are colored black and the other 5 are colored white and hence indistinguishability is limited to each of these groups of 5?

3. Explain how replacement affects the independence of successive trials (i.e. whether or not the results of the next trials depend on the results of the previous trials).
Answer: In general and from this perspective, the results of successive trials are independent of each other in the case of replacement and dependent in the case of no replacement. For example, if we have a pack of 10 cards 5 of which are blue and the other 5 are red and we draw several cards in succession then it is obvious that if we replace the card after each draw before making the next draw then our chance of getting a red or blue card in each draw is the same and it is independent of the results of the previous draws (because the pack is the same in each draw). However, if we do not replace the drawn cards then our chance of getting a red or blue card in each draw depends on the results of the previous draws, i.e. getting blue/red in the previous draws will increase the chance of getting red/blue in the next draws (and vice versa).
PE: Why did we say "In general and from this perspective"?
4. Give an example of a situation in which replacement has negligible effect, i.e. selection with and without replacement are virtually identical in effect.
Answer: When selecting a small sample from a large population (of identical objects or large groups of identical objects) the effect of replacement becomes negligible because the nature of the population is not affected tangibly by taking the tiny sample. For example, if we draw two balls consecutively from a collection of 10000 balls half of which (i.e. 5000) are red and the other half are blue then our chance of getting a red (or blue) ball in the second draw is practically the same regardless of replacing or not replacing the first ball (and regardless of the color of the first ball).
PE: Try to set some criteria to determine if replacement has negligible/non-negligible effect when taking a tiny sample from a large population.

1.6 Simulation of Probability Problems

Most of the probability problems can be easily simulated computationally by using random number generators to make random selections similar to the real life random selections and events. Although these generators are not really and exactly random, they are very close to be so. This is because the deterministic aspects and dependencies of these generators are entirely irrelevant to the probabilistic aspects of the virtual probabilistic experiment[9] and hence they provide a reliable way for imitating real life random occurrences and events.

In this book, we demonstrated the use of computer simulation in probability (where simulation is relevant and useful) in a few solved problems using C++ programming language. Although the codes are deliberately written and structured in a simple way using basic techniques[10] (and hence they achieve no more than the minimum of the objectives of the simulated problems), they provide a useful way for the beginners to learn how to simulate probability problems (as well as providing a simple way for testing and checking the theoretical results obtained by analytical methods). As indicated in the Preface, the codes are freely available on the internet.

We encourage those who have interest in using computational methods (whether in science or mathematics or engineering or something else, and whether in probability or something else) to inspect these codes. We also encourage them to solve the proposed exercises (i.e. the PE's) that request writing or modifying or commenting or flowcharting computer codes that simulate probability problems. In fact, computer simulation is very effective (and rather easy and motivating) method in the investigations of probability problems and hence it should be considered as one of the main tools in these investigations.

[9] We may also say: the effect of their deterministic aspect is negligible in the given context and presumed situation.
[10] In fact, one reason for the intentional simplicity of these codes and the use of rather primitive methods of simulation is to avoid complexities that usually lead to confusion and unnecessary difficulties.

Therefore, those who have serious interest in probability should consider developing the skills of simulation (as well as calculation and computation in general) using programming languages and computer codes.

1.7 Probability and Common Sense

As indicated earlier, probability in its fundamental principles and at its basic level is generally intuitive and based on common sense. However, this is not always the case especially at the high levels of probability theory and its applications where probability problems and mathematics become complicated and involve many elements and considerations some of which at least are strange and far from daily life experiences, intuition and common sense. So, we advise the "mathematicians" of the probability theory to use their common sense and intuition in tackling and solving probability problems but with caution and to certain limits and up to certain levels. Accordingly, if the well established formulations of the probability theory and its verified mathematics lead to a result that is against their common sense and intuition then they should be prepared to digest it and accept it with no hesitation. In fact, there are many problems and results in probability theory (some of which will be met in this book) where the theory fails to comply with intuition and common sense (see for example Problem 3 of § 6.1). Anyway, computer simulation (see § 1.6) can provide a simple and effective method for verifying the suspected results and getting the required confidence or certainty (although simulation is not always applicable for this purpose noting for instance that some considerations and subtleties may not lend themselves to simulation).

1.8 Controversies about Probability

There are many controversies about probability and its mathematical theory. Most of these controversies are historical but others are still going on. So, probability theory and its applications are not as universal or agreed-upon as calculus or complex analysis for instance.[11] Accordingly, we may have more than one opinion about the method of solving some probability problems, and the results that we obtain from these methods may not be identical. This is due for example to the presence of delicate considerations and subtleties in the given problem or the existence of conflicting opinions, definitions and conventions.

In the following subsections we provide a brief glimpse into some of these controversies and disputes. However, before that we should note that although these controversies (or some of them at least) are not important to our investigation of the probability theory, awareness of their existence is important (at least to avoid potential confusion and misunderstanding in some occasions and circumstances). We should also note that the impact of these controversies is mostly minimal. Moreover, they can be settled if a systematic, tidy and transparent approach is followed. Therefore, the existence of these controversies should not affect the general confidence in the probability theory and its results.

1.8.1 Controversies about the Definition of Probability

The controversies about the definition of probability are mostly of historical nature and hence they are beyond the scope of this book (which is about the theory of probability not its history). However, the reader should be aware of the following:

• Some of these controversies can have a real impact and tangible consequences on the mathematical formulation of the theory and hence their effect may not be restricted to the definition. In other words, they could have a practical impact on the results and applications of the theory and not just on its theoretical or axiomatic structure.

• Some of these (primarily-historical) controversies can creep to modern day literature and could be traced in some writings and applications (of authors and researchers of modern or relatively-modern time). Anyway, we will generally follow the mainstream and commonly-accepted axiomatic approach of modern day probability theory (and hence the readers should not worry about this issue).

[11] For example, getting the derivative of a function or its definite integral has a unique and undisputed answer in calculus.

1.8.2 Frequentism versus Bayesianism

The difference between these views (or trends or schools) is essentially about the definition and interpretation of probability where the frequentist adopts the criterion (or concept) of relative frequency (in a large number of trials) as a basis for the definition and interpretation of probability while the Bayesian adopts the criterion of relative support of evidence to proposition. In this regard we note the following:
• The probability of frequentism is essentially objective while the probability of Bayesianism is relatively subjective, i.e. it contains an element of subjectivity (see § 1.3).
• The difference between frequentism and Bayesianism is not only of philosophical or contemplative nature but it usually leads to differences in the formulation of the probability theory and the results obtained from each school in its application. In fact, this is one cause for why solving a probability problem may not have a unique and agreed-upon solution (which we indicated previously).
• Frequentism is generally the dominant trend in probability studies and applications although Bayesianism has also some staunch adherents.
• There are advantages and disadvantages in both these schools as they both have merits and weaknesses although frequentism seems to be the strongest (which may explain why it is the dominant trend).
• The more appropriate method to use (i.e. frequentism or Bayesianism) may depend on the problem in question and its nature. For example, frequentism could be the preferred method in certain branches of science or types of problems while Bayesianism could be the preferred method in other branches and types.
• The probability theory that we investigate in this book essentially represents the frequentism view (and hence we insist on adopting the concept of objective probability which is based on the idea of "outcome of repetitive trial"). This is not only because of the dominance of the frequentism school in modern times, but also because of our belief that frequentism is more objective and hence it is more appropriate to use in sciences in general and in physical sciences in particular (noting that science is the main field of application for the probability theory).

1.9 Modeling in Probability

As we will see, the mathematics of probability is generally simple (at least for the level that we deal with in this book).[12] In fact, the formal aspects of probability problems are generally trivial and are mainly achieved by elementary mathematical operations such as addition, subtraction, multiplication, summation and basic integration. What is difficult about probability and its theory, however, is the modeling of problems and casting them in a formal probabilistic style. In other words, what is difficult is to formulate the given probability problem in a formally-recognized way that makes it fit in a known and recognizable probability law or rule or pattern so that it can be tackled and solved by applying the formality of that law or rule or pattern. So, what the amateur mathematicians of probability theory should concentrate on (and be more keen to learn and acquire) is to develop the probability modeling skills, and hence they should look to the given solved problems and proposed exercises as being mainly practical instances for developing these modeling skills.[13] In fact, this objective (i.e. developing probability modeling skills) was in our sight when we selected and created most of the solved problems and proposed exercises, and this is one reason why some of these problems may look trivial while others look more difficult than they should be for the level of this book. We also considered the diversity of these problems and exercises partly because of our consideration of the necessity of developing these skills. Accordingly, I advise the readers to be more keen about acquiring these modeling skills than about solving individual problems in their formal mathematical dimensions (i.e. as if they are calculus or complex analysis problems for instance).

[12] The discussion in this section should extend to subjects related to probability such as counting (which will be investigated in § 2.2). So, "probability" here is more general than its primary meaning.
[13] Learning these modeling skills (i.e. how to pose the given problem in a manner that makes it solvable or easily solvable) is more of an art than a science.

Chapter 2
Mathematical Preliminaries

In this chapter we present some mathematical preliminaries related mostly to the set theory and the methods and rules of counting. These preliminaries are required in the development of the theory of probability and its applications which will be investigated in the later chapters.

2.1 The Basics of Set Theory

In this section we investigate briefly the basics of set theory which is commonly used (as language, conventions, symbolism, etc.) in the presentation and formulation of probability theory.

"**Set**" means a collection of objects with a common property[14] irrespective of their order (and hence the set made of the elements a, b is the same as the set made of the elements b, a). These objects are described as **members** or **elements** of the set. These elements are not repetitive, i.e. each distinct element is represented once in the set (and hence the set made of the elements a, b is the same as the set made of the elements a, a, b or a, b, b). If an object a belongs to a set A (i.e. a is a member of A) we write $a \in A$ and if it does not belong to A we write $a \notin A$. A set is specified mathematically either by **listing** its members (inside curly brackets $\{...\}$ where the objects are separated from each other by commas) or by giving its **description** (inside curly brackets $\{...\}$) by identifying the common property(s) of its members. For example, the set A of the integer numbers from 1 to 5 may be stated mathematically as:[15]

$$A = \{1, 2, 3, 4, 5\} \qquad \text{or} \qquad A = \{\text{integers from 1 to 5}\} \tag{1}$$

Two sets are equal *iff* they have the same elements (and hence the equality sign $=$ and the non-equality sign $\neq$ should be interpreted accordingly). For example, if A is the set consisting of the numbers $1, 2, 3, 4, 5$ and B is the set of positive integers < 6 then we can write $A = B$. Sets are usually (but not necessarily) labeled with uppercase letters and their elements with lowercase letters.

The **empty set** (symbolized by $\varnothing$ or $\{\}$ and may also be called the **null set**) is a set that contains no element. For example, the set of "even prime numbers > 2" is empty because no even number > 2 can be prime since it is divisible by 2. A **universal set** is a set that includes all the elements of all the related sets of concern in the particular situation and context. For example, if we are interested in the vowels of the English alphabet (irrespective of being symbolized by lowercase or uppercase) then the universal set in this case and context is $U = \{a, e, i, o, u\}$ since all vowels and groups of vowels belong to this set. As indicated, the universal set depends on the case and context (and hence if we shifted our attention to the letters of the English alphabet then the set $\{a, e, i, o, u\}$ is not universal anymore).

A set B is a **subset** of a set A if all the members of B are members of A. For example, if $A = \{1, 2, 3, 4, 5\}$ then $B = \{1, 3, 4\}$ is a subset of A. By convention, the set itself and the empty set are subsets of any set (also see Problem 2). A **proper subset** of a set is a subset that is not the same as that set (i.e. the

[14] The condition "with a common property" is to exclude collection of objects that have no common property and hence it is not sensible to treat them as a single set since set requires certain common features that characterize its elements and justify the application of its rules and associated concepts. For example, it is not sensible (in common situations) to have a set of apples and cars (i.e. as such). In fact, this condition should also be justified by the upcoming methods of specifying the set (i.e. by giving its description by identifying the properties of its members) because no such specification can be given unless the members of the set possess some common properties and features.

[15] In fact, each one of listing and description can be in several different forms. For example, we can write the above as:

$$A = \{3, 2, 4, 5, 1\} \qquad \text{or} \qquad A = \{n : \ n \in \mathbb{N}, \ 1 \leq n \leq 5\}$$

We also note that listing applies when the number of elements (of countable set) is small or can be written compactly (e.g. $1, 2, \ldots, k$).

subset has fewer elements than the set itself). For example, if $A = \{1, 2, 3, 4, 5\}$ then $B = \{1, 3, 4\}$ is a proper subset but $C = \{2, 4, 3, 1, 5\}$ is not (i.e. it is a subset but not a proper one). The common symbol for subset is $\subset$ (e.g. $C \subset D$ means C is a subset of D). Other symbols are also used.[16]

A set that contains exactly n elements is commonly described as a **set of size n. Finite set** is a set that has a finite number of elements, while **infinite set** is a set that has an infinite number of elements. **Countable set** is a set whose elements can be put in one-to-one correspondence with the set of natural numbers, while **uncountable set** cannot. Countable set can be finite or infinite (noting that some may limit the use of this term to infinite). In general, countable sets are described by discrete variables while uncountable sets are described by continuous variables.

The **intersection** of two sets, A and B, is the set of elements that belong to both A and B. The symbol used for intersection is $\cap$. For example, if $A = \{a, b, c, d, e\}$ and $B = \{d, e, f, g, h\}$ then $A \cap B = \{d, e\}$. Intersection is easily generalized to more than two sets, e.g. $A \cap B \cap C$ means the intersection of the sets A, B, C which is the set of elements that belong to all these three sets. More generally, the intersection of n sets $(A_1, \ldots, A_n)$ is $\cap_{i=1}^{n} A_i$ which is the set of elements that belong to everyone of the n sets (noting that n could be infinite, i.e. $\cap_{i=1}^{\infty} A_i$).

The **union** of two sets, A and B, is the set of elements that belong to A or B or both. The symbol used for union is $\cup$. For example, if $A = \{a, b, c, d, e\}$ and $B = \{d, e, f, g, h\}$ then $A \cup B = \{a, b, c, d, e, f, g, h\}$. Union is easily generalized to more than two sets, e.g. $A \cup B \cup C$ means the union of the sets A, B, C which is the set of elements that belong to any one of these three sets (individually or collectively). More generally, the union of n sets $(A_1, \ldots, A_n)$ is $\cup_{i=1}^{n} A_i$ which is the set of elements that belong to at least one of the n sets (noting that n could be infinite, i.e. $\cup_{i=1}^{\infty} A_i$).

The **complement** of a set A is the set that contains all the elements in the universal set that do not belong to A. The common symbol used for complement is a bar (or a line) over the symbol of the set (e.g. the complement of A is $\overline{A}$).[17] For example, if we are interested in the vowels of the English alphabet then the complement of the set $A = \{a, i\}$ is $\overline{A} = \{e, o, u\}$. It is obvious that the universal set is made of the union of any one of its subsets and its complement, i.e. if $A \subset U$ then $U = \{\text{elements of } A \text{ and elements of } \overline{A}\} = A \cup \overline{A}$.

The **difference** of a set A from a set B (which is symbolized as $A - B$ or $A \setminus B$) is the set of elements that belong to A but not to B.[18] For example, if $A = \{a, b, c, d, e\}$ and $B = \{d, e, f, g, h\}$ then $A - B = \{a, b, c\}$ while $B - A = \{f, g, h\}$. Two sets are **mutually exclusive** (or **disjoint** or **incompatible**) if their intersection is empty (i.e. they have no common element). In mathematical terms, A and B are mutually exclusive sets *iff*

$$A \cap B = \varnothing \tag{2}$$

A number of n sets $(A_1, \ldots, A_n)$ are pairwise exclusive (or disjoint) *iff* $A_i \cap A_j = \varnothing$ $(i, j = 1, \ldots, n$ and $i \neq j)$.[19]

The operations on sets are characterized by certain properties and subject to certain rules and laws which are outlined in the following list:[20]

1. **Rules of identity**:

$$\varnothing \cap A \quad = \quad \varnothing \tag{3}$$

[16] The symbol $\subseteq$ may be used for subset in general (i.e. whether proper or not) and hence $\subset$ is restricted to proper subset. The symbols $\supset$ and $\supseteq$ may also be used (corresponding to $\subset$ and $\subseteq$ respectively) to mean super set (i.e. proper and general respectively). However, in this book we generally use only $\subset$ without distinction between proper and improper (or general). Yes, if distinction is required in a certain context then we will make the distinction clear.

[17] Other common symbols for complement include prime and superscript c (i.e. A' and A^c).

[18] The operation of taking the difference of sets is described as **subtraction of sets**. The similarity between difference and complement is obvious and hence the difference of A from B may be called the **relative complement** of B with respect to A (and thus complement may be called **absolute complement**).

[19] The reader is referred to Problem 5 of § 3.3 for further details about this issue.

[20] We note that some of the following properties and rules that involve two sets can be extended and generalized to include more than two sets (and even to infinitely many sets). Some examples of these extensions and generalizations will be given later on. We should also note that although the above list is generally representative and inclusive to the main properties and rules, it is not entirely exhaustive (in fact we will meet examples of some other properties and rules in the solved Problems).

$$\begin{aligned} \varnothing \cup A &= A && (4)\\ A \cap A &= A && (5)\\ A \cup A &= A && (6)\\ U \cap A &= A && (7)\\ U \cup A &= U && (8) \end{aligned}$$

2. **Rules of complement**:

$$\begin{aligned} \overline{\varnothing} &= U && (9)\\ \overline{\overline{A}} &= A && (10)\\ \overline{A} \cap A &= \varnothing && (11)\\ \overline{A} \cup A &= U && (12)\\ \overline{U} &= \varnothing && (13) \end{aligned}$$

3. **Commutativity**:

$$\begin{aligned} A \cap B &= B \cap A && (14)\\ A \cup B &= B \cup A && (15)\\ A - B &\neq B - A && (16) \end{aligned}$$

4. **Associativity**:

$$\begin{aligned} (A \cap B) \cap C &= A \cap (B \cap C) = A \cap B \cap C && (17)\\ (A \cup B) \cup C &= A \cup (B \cup C) = A \cup B \cup C && (18) \end{aligned}$$

5. **Distributivity** (i.e. of intersection on union and union on intersection):[21]

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) && (19)\\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C) && (20) \end{aligned}$$

6. **De Morgan laws**:

$$\begin{aligned} \overline{A \cap B} &= \overline{A} \cup \overline{B} && (21)\\ \overline{A \cup B} &= \overline{A} \cap \overline{B} && (22) \end{aligned}$$

A **partition of a set** A is a division (or grouping) of A into *non-empty*, *disjoint* and *comprehensive* subsets. For example, if $A = \{0, 1, 2, 3, 4, 5, 6, 7\}$ then $\{\{0, 1, 3\}, \{2, 5, 7\}, \{4, 6\}\}$ is a partition of A because the subsets $\{0, 1, 3\}$, $\{2, 5, 7\}$ and $\{4, 6\}$ are non-empty, they have no shared elements (i.e. disjoint) and their union is A (i.e. comprehensive). Also, $\{\{0, 3\}, \{1, 4\}, \{2, 5, 6, 7\}\}$ and $\{\{5, 6, 7\}, \{0, 1, 2, 3, 4\}\}$ are two other partitions of A. The subsets in a partition of a set are commonly called **cells** (of that partition), e.g. $\{5, 6, 7\}$ is a cell of the partition $\{\{5, 6, 7\}, \{0, 1, 2, 3, 4\}\}$. It is obvious that the partition of a set is a set whose elements are sets and hence the partition is a set of sets.

Problems

1. State the following in the notation of sets (where A and B are sets representing outcomes of a trial while α, β, γ are potential outcomes):

[21] We note that there are other types of distributivity relations in the algebra of sets, e.g. distributivity relations involving difference of sets like $A \cap (B - C) = (A \cap B) - (A \cap C)$. We also note that distributivity can take place from left or from right (and hence we have left distributivity and right distributivity). However, most of these relations and properties are not needed in this book (noting that we will investigate some of them later on; see Problem 20).

(**a**) α is a member of A but not of B. (**b**) B is not a subset of A.

(**c**) The elements of B which are not in A. (**d**) All possible outcomes are in A or in B.

(**e**) β is neither in A nor in B. (**f**) γ belongs to A and B.

Answer:

(**a**) $\alpha \in (A - B)$. (**b**) $B \not\subset A$. (**c**) $B - A$ (or $B \setminus A$).

(**d**) $A \cup B = U$. (**e**) $\beta \notin (A \cup B)$. (**f**) $\gamma \in (A \cap B)$.

PE: Repeat the Problem for the following (where A, B, C and D are sets):

(**a**) The subset of a subset of A is a subset of A. (**b**) The complement of the complement of A is A.

(**c**) No proper subset can be a universal set. (**d**) α belongs to A or B or C but not to D.

(**e, f**) The complement of the intersection/union of A and B is the union/intersection of their complements.

2. Identify the status of the empty set as being proper or not proper subset.
   Answer: Following the literature, the empty set is a proper subset of every non-empty set but it is not a proper subset of itself (because it is itself). However, in our view this should be seen as a convention more than a proven fact or result.
   PE: Do you think the empty set is a real set (i.e. in the same sense as non-empty set) or it is a creation of mathematician to fill certain gaps and make some generalizations?
3. The following statements are in the notation of sets (where A and B are sets). Express them in ordinary language.

   (**a**) $\{x : \ x \in \mathbb{C}, x \notin \mathbb{R}\}$. (**b**) $\{\} \subset U$. (**c**) $\varnothing \subset A \quad (A \neq \varnothing)$.

   (**d**) $\overline{U} = \varnothing$. (**e**) $\overline{A \cap B} = \overline{A} \cup \overline{B}$. (**f**) $17 \notin \Big\{\mathbb{N} - \{2, 3, 5, 7, 11, \ldots\}\Big\}$.

   Answer:

   (**a**) The set of strictly complex numbers (i.e. non-real complex numbers).
   (**b**) The empty set is a subset of the universal set.
   (**c**) The empty set is a subset of any other set.
   (**d**) The complement of the universal set is the empty set.
   (**e**) The complement of the intersection of A and B is the same as the union of their complements.
   (**f**) The number 17 does not belong to the set of natural non-prime numbers.

   PE: Repeat the Problem for the following (where A, B and C are sets):

   (**a**) $B = \{x : \ x \in U, x \notin A\}$. (**b**) $A \cap B \cap C \neq \varnothing$. (**c**) $A \cap B \cap C = U$.

   (**d**) $A \cup B \cup C \neq U$. (**e**) $A \cap B = A \cup B$. (**f**) $A = \{x : \ x \notin U\}$.
4. Give examples of finite and infinite sets.
   Answer:
   Examples of finite set: the set of students in a school, the set of human beings (in all places and times), the set of atoms in our galaxy, the set of even numbers between 0 and 100, the set of solutions of n^{th} order polynomial.
   Examples of infinite set: the set of natural numbers $\mathbb{N}$, the set of real numbers between 0 and 1, the set of straight lines passing through a point (in 2D or 3D space), the set of points in a line segment, the set of solutions of consistent and dependent set of linear equations.
   PE: Give more examples of finite and infinite sets (six each).
5. Give examples of countable and uncountable sets.
   Answer:
   Examples of (finite) countable set: the set $\{1, 2, \ldots, 99\}$, the set of living beings on Earth, the set of stars in a galaxy, the set of solutions of consistent and independent set of linear equations.
   Examples of (infinite) countable set: the set of integers $\mathbb{Z}$, the set of prime numbers, the set of rational numbers.
   Examples of uncountable set: the set of planes parallel to the xy plane, the set of complex numbers in

the origin-centered unit disk, the set of real numbers in the interval $[1, 2]$.
PE: Give more examples of countable and uncountable sets (five each).

6. State some facts about the empty set $\varnothing$, the universal set U and their complements.
Answer:
• The complement of the empty set is the universal set, i.e. $\overline{\varnothing} = U$.
• The complement of the universal set is the empty set, i.e. $\overline{U} = \varnothing$.
• For any set A we have: $\varnothing \subset A \subset U$.
• If A is a subset of the null set then A is null, i.e. $A \subset \varnothing \ \rightarrow \ A = \varnothing$.
• $U - A = \overline{A}$.
PE: State three more facts about the empty set $\varnothing$, the universal set U and their complements.

7. State some facts about subsets.
Answer: If A, B, C are any three sets then we have:
• $A \subset A$.
• $\varnothing \subset A \subset U$.
• $(A \cap B) \subset A \subset (A \cup B)$.
• $A = B$ *iff* $A \subset B$ AND $B \subset A$.
• $A \subset B$ *iff* $A \cap B = A$.
• $A \subset B$ *iff* $A \cup B = B$.
• $A \subset B$ *iff* $\overline{B} \subset \overline{A}$.
• If $A \subset B$ AND $B \subset C$ then $A \subset C$.
• If $A \cap B = \varnothing$ then $A \subset \overline{B}$ (and $B \subset \overline{A}$).
• If $A \subset B$ then $A \cup (B - A) = B$.
PE: State four more facts about subsets.

8. Use the notation of sets to express the relation between integers $\mathbb{Z}$, complex numbers $\mathbb{C}$, natural numbers $\mathbb{N}$, real numbers $\mathbb{R}$, and rational numbers $\mathbb{Q}$.
Answer: $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.
PE: Use the notation of sets to express the relation between the sets: primates (P), animals (A), homo sapiens (H), living beings (L), and mammals (M).

9. Considering the set of Greek letters, let $A = \{\alpha, \beta, \delta, \rho\}$, $B = \{\alpha, \gamma, \mu, \psi, \omega\}$, $C = \{\{\alpha, \gamma\}, \{\rho\}\}$ and $D = \{\{\}, \{\beta, \mu\}\}$. Which of the following is true/false:

(**a**) $\alpha \subset B$. (**b**) $\delta \in A$. (**c**) $\varnothing \in D$. (**d**) $\{\alpha, \gamma\} \subset B$.

(**e**) $\{\alpha, \gamma\} \in C$. (**f**) $\omega \in \overline{A}$. (**g**) $B \cap C = \{\alpha, \gamma\}$. (**h**) $\tau \in \overline{A \cup \overline{B}}$.

Answer:
(**a**) False. (**b**) True. (**c**) True. (**d**) True.
(**e**) True. (**f**) True. (**g**) False. (**h**) False.

PE: Repeat the Problem for the following:

(**a**) $\alpha \supset B$. (**b**) $C \cap D = 0$. (**c**) $A \cup B = \{\alpha\}$. (**d**) $A - B = \{\beta, \delta, \rho\}$.

(**e**) $\{\} \subset D$. (**f**) $\alpha \in (A \cap B)$. (**g**) $A \cap C = \{\alpha, \rho\}$. (**h**) $\omega \in \left(\overline{A} \cap B\right)$.

10. Give mathematical conditions equivalent to $A \subset B$ (i.e. A is a subset of B).
Answer: For example:

$$\overline{B} \subset \overline{A} \qquad A \cap B = A \qquad A \cup B = B \qquad A \cap \overline{B} = \varnothing \qquad \overline{A} \cup B = U$$

PE: Give mathematical conditions equivalent to $\overline{A \cap B} = U$.

11. Identify the relationships between the following: $\{\}, \varnothing, \{\varnothing\}, \{0\}, 0$.
Answer: Let first determine the meaning of each one of these so that we can identify the relationship between them. $\{\}$ and $\varnothing$ mean the null set, $\{\varnothing\}$ means the set that has a single element which is the empty set (and hence it is a set of sets), $\{0\}$ means the set that has a single element which is the number 0, and 0 is the number zero and hence it is not a set. Accordingly:

$\{\} = \varnothing$ $\{\} \in \{\varnothing\}$ $\{\} \subset \{\varnothing\}$ $\{\} \subset \{0\}$ $\varnothing \in \{\varnothing\}$

$\varnothing \subset \{\varnothing\}$ $\varnothing \subset \{0\}$ $\{\varnothing\} \neq \{0\}$ $0 \in \{0\}$

PE: Identify the relationships between the following: $\varnothing, A, U$ where A is a non-empty proper subset of U.

12. Associate intersection, union and complement with logical operators.
Answer: In general, intersection is associated with AND, union is associated with (non-exclusive) OR, and complement is associated with NOT.
PE: Suggest logical operators corresponding to the following set operations (where A and B are sets and U is the universal set):
(**a**) $(A \cup B) - (A \cap B)$. (**b**) $U - A$. (**c**) $\overline{\overline{A} \cup \overline{B}}$. (**d**) $\overline{\overline{A} \cap \overline{B}}$.
13. Give a simple justification for the commutativity of the operations of intersection and union (i.e. Eqs. 14 and 15).
Answer: Intersection is like logical AND (e.g. $A \cap B$ means something that belongs to A AND to B) and union is like logical OR (e.g. $A \cup B$ means something that belongs to A OR to B). Now, since the logical operations of AND and OR are commutative, intersection and union must be commutative.[22]
PE: Give a simple justification for the non-commutativity of the operation of difference of sets (i.e. Eq. 16).
14. State some facts about complement and difference of sets.
Answer: If A, B, C are any three sets then we have:
• $A - B = A \cap \overline{B}$.
• $(A - B) \cap (B - A) = \varnothing$.
• $(A - B) \cup (B - A) = (A \cup B) - (A \cap B)$.
• $A - B = \overline{B} - \overline{A}$.
• $\overline{A \cap B} = \overline{A} \cup \overline{B}$.
• $\overline{A \cup B} = \overline{A} \cap \overline{B}$.
• $(A \cap B) - (A \cap C) = A \cap (B - C)$.
PE: Explain and justify in words the stated facts in the answer. Do you see any similarity between $A - B = \overline{B} - \overline{A}$ and the De Morgan laws (try to compare)?
15. Generalize the distributivity property (i.e. of intersection on union and union on intersection) to any number of sets.
Answer: Distributivity can be easily generalized to any number of sets by grouping (through employing associativity) with repeated application of distributivity involving three sets (followed by employing associativity to get rid of grouping). For example, if A, B, C, D are four sets then:

$$\begin{aligned} A \cap (B \cup C \cup D) &= A \cap \Big[B \cup (C \cup D)\Big] = (A \cap B) \cup \Big[A \cap (C \cup D)\Big] \\ &= (A \cap B) \cup \Big[(A \cap C) \cup (A \cap D)\Big] = (A \cap B) \cup (A \cap C) \cup (A \cap D) \\ \text{and} \quad A \cup (B \cap C \cap D) &= A \cup \Big[B \cap (C \cap D)\Big] = (A \cup B) \cap \Big[A \cup (C \cap D)\Big] \\ &= (A \cup B) \cap \Big[(A \cup C) \cap (A \cup D)\Big] = (A \cup B) \cap (A \cup C) \cap (A \cup D) \end{aligned}$$

This pattern can be easily extended by induction to any number of sets. Accordingly, we can write:

$$\begin{aligned} A \cap (\cup_i B_i) &= \cup_i (A \cap B_i) \\ \text{and} \quad A \cup (\cap_i B_i) &= \cap_i (A \cup B_i) \end{aligned}$$

PE: Justify in words each step of the above generalizations.
16. Generalize the De Morgan laws to intersections and unions involving more than two sets.
Answer: The De Morgan laws can be easily generalized to intersections and unions involving more than

[22] In fact, this is ultimately based on the fact that the relationships represented by intersection and union (as well as AND and OR) are symmetric.

two sets by grouping (through employing associativity) with repeated application of the De Morgan laws involving two sets (followed by employing associativity to get rid of grouping). For example, if A, B, C are three sets then:

$$\begin{aligned} \overline{A \cap B \cap C} &= \overline{A \cap (B \cap C)} = \overline{A} \cup (\overline{B \cap C}) = \overline{A} \cup (\overline{B} \cup \overline{C}) = \overline{A} \cup \overline{B} \cup \overline{C} \\ \text{and} \quad \overline{A \cup B \cup C} &= \overline{A \cup (B \cup C)} = \overline{A} \cap (\overline{B \cup C}) = \overline{A} \cap (\overline{B} \cap \overline{C}) = \overline{A} \cap \overline{B} \cap \overline{C} \end{aligned}$$

This pattern can be easily extended by induction to any number of sets. Accordingly, we can write:

$$\begin{aligned} \overline{\cap_i A_i} &= \cup_i \overline{A_i} \\ \text{and} \quad \overline{\cup_i A_i} &= \cap_i \overline{A_i} \end{aligned}$$

PE: Justify in words each step of the above generalizations.

17. Prove (or rather justify) the De Morgan law for union involving two sets (see Eq. 22).
Answer: Referring to Eq. 22, $A \cup B$ means the elements in the universal set that belong to A or B or both, and hence $\overline{A \cup B}$ means the elements in the universal set that belong neither to A nor to B. Similarly, $\overline{A}$ (and $\overline{B}$) means the elements in the universal set that do not belong to A (and B), and hence $\overline{A} \cap \overline{B}$ means the elements in the universal set that do not belong to A AND do not belong to B, i.e. the elements in the universal set that belong neither to A nor to B. Accordingly, the left and right hand sides of Eq. 22 have the same meaning (i.e. they have the same elements) and hence the equality is established.
PE: Repeat the Problem for the De Morgan law for intersection involving two sets (see Eq. 21).

18. In this Problem we have: $\Upsilon = \{a, b, \ldots, z, A, B, \ldots, Z\}$, $\Delta = \{a, b, \ldots, z\}$, $\Psi = \{A, B, \ldots, Z\}$, and $\Omega = \{a, e, i, o, u, A, E, I, O, U\}$.[23] Identify the following:
(**a**) $\overline{\Delta}$. (**b**) $\Delta \cap \Psi$. (**c**) $\Delta \cup \Psi$. (**d**) $\Delta - \Psi$. (**e**) $\Upsilon - \Omega$.
(**f**) $\overline{\Delta \cap \Psi}$. (**g**) $\overline{\Delta \cup \Psi}$. (**h**) $\overline{\Upsilon - \Omega}$. (**i**) $\Omega - \Upsilon$. (**j**) $\Upsilon - \overline{\Delta}$.
Answer:
(**a**) Ψ. (**b**) $\varnothing$. (**c**) Υ. (**d**) Δ. (**e**) $\{\alpha : \alpha \text{ is consonant}\}$.
(**f**) Υ. (**g**) $\varnothing$. (**h**) Ω. (**i**) $\varnothing$. (**j**) Δ.
PE: Repeat the Problem for the following:
(**a**) $\overline{\Upsilon}$. (**b**) $\Psi - \Delta$. (**c**) $\overline{\Delta - \overline{\Delta}}$. (**d**) $\overline{\Upsilon \cap \Omega}$. (**e**) $\overline{\Psi} - \overline{\Omega}$.

19. Find all the partitions of the following sets:
(**a**) $\{\alpha, \beta, \gamma\}$. (**b**) $\{\alpha, \beta, \gamma, \delta\}$.
Answer:
(**a**) We have 5 partitions:
$\{\{\alpha\}, \{\beta\}, \{\gamma\}\}$ $\{\{\alpha, \beta\}, \{\gamma\}\}$ $\{\{\alpha, \gamma\}, \{\beta\}\}$ $\{\{\alpha\}, \{\beta, \gamma\}\}$ $\{\{\alpha, \beta, \gamma\}\}$
(**b**) We have 15 partitions:
$\{\{\alpha\}, \{\beta\}, \{\gamma\}, \{\delta\}\}$ $\{\{\alpha, \beta\}, \{\gamma\}, \{\delta\}\}$ $\{\{\alpha, \gamma\}, \{\beta\}, \{\delta\}\}$ $\{\{\alpha, \delta\}, \{\beta\}, \{\gamma\}\}$
$\{\{\alpha\}, \{\beta, \gamma\}, \{\delta\}\}$ $\{\{\alpha\}, \{\beta, \delta\}, \{\gamma\}\}$ $\{\{\alpha\}, \{\beta\}, \{\gamma, \delta\}\}$ $\{\{\alpha, \beta\}, \{\gamma, \delta\}\}$
$\{\{\alpha, \gamma\}, \{\beta, \delta\}\}$ $\{\{\alpha, \delta\}, \{\beta, \gamma\}\}$ $\{\{\alpha\}, \{\beta, \gamma, \delta\}\}$ $\{\{\beta\}, \{\alpha, \gamma, \delta\}\}$
$\{\{\gamma\}, \{\alpha, \beta, \delta\}\}$ $\{\{\delta\}, \{\alpha, \beta, \gamma\}\}$ $\{\{\alpha, \beta, \gamma, \delta\}\}$
PE: What is the number of partitions of the set $\{\alpha, \beta, \gamma, \delta, \varepsilon\}$? Also find 20 of these partitions.

20. Prove (or justify) the following identities (where A, B and C are sets):
(**a**) $A - B = A \cap \overline{B}$.
(**b**) $(A - B) \cap (B - A) = \varnothing$.
(**c**) $(A - B) \cup (B - A) = (A \cup B) - (A \cap B)$.

[23] We note that in this Problem Υ is the universal set and $U \in \Omega$ is a letter (not the universal set).

(**d**) $(A \cap B) \cap (A - B) = \varnothing$.
(**e**) $(A \cap B) \cup (A - B) = A$.
(**f**) $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$.
(**g**) $A \cap (B - C) = (A \cap B) - (A \cap C)$.
(**h**) $(B - C) \cap A = (B \cap A) - (C \cap A)$.
(**i**) $A - (B \cup C) = (A - B) \cap (A - C)$.
(**j**) $(B \cup C) - A = (B - A) \cup (C - A)$.
(**k**) $A - (B \cap C) = (A - B) \cup (A - C)$.
(**l**) $(B \cap C) - A = (B - A) \cap (C - A)$.
(**m**) $(B - C) - A = (B - A) - (C - A)$.
(**n**) $A - (B - C) = (A - B) \cup (A \cap C)$.
(**o**) $U - (A \cup B) = \overline{A} \cap \overline{B}$.
(**p**) $\overline{(B - A)} \cup \overline{(A - B)} = U$.
(**q**) $A \cap (A \cup B) = A$.
(**r**) $A \cup (A \cap B) = A$.
(**s**) $A - B = \overline{B} - \overline{A}$.

Answer: These identities (or at least some of them) are intuitive. In the following we use sometimes verbal arguments and sometimes formal arguments for the purpose of diversity and to show that many set relations and identities can be proved (or justified) by simple intuitive arguments rather than by formal arguments. We also note that some of these identities can be easily extended and generalized to include more sets.

(**a**) $A - B$ represents the elements of A which do not belong to B, while $A \cap \overline{B}$ represents the elements which belong to A and to $\overline{B}$ and hence they belong to A but not to B. So, $A - B$ and $A \cap \overline{B}$ means the same thing and hence they are equal.

(**b**) $A - B$ represents the elements of A which do not belong to B, while $B - A$ represents the elements of B which do not belong to A, and hence they cannot have any common element.

(**c**) $A - B$ represents the elements of A which do not belong to B, while $B - A$ represents the elements of B which do not belong to A, and hence their union $(A - B) \cup (B - A)$ represents the elements of the union of A and B (which is $A \cup B$) excluding the common elements of A and B (which is $A \cap B$).

(**d**) The elements of A either belong to B (which are represented by $A \cap B$) or do not belong to B (which are represented by $A - B$), and hence their intersection must be empty.

(**e**) The elements of A either belong to B (which are represented by $A \cap B$) or do not belong to B (which are represented by $A - B$), and hence their union must be the entire A.

(**f**) The elements of $A \cup B$ must belong to A only (which are represented by $A - B$), or to B only (which are represented by $B - A$), or to both (which are represented by $A \cap B$), and hence their union must be the entire $A \cup B$.[24]

(**g**) The elements of B which are common to A but not to C $\big[$i.e. $A \cap (B - C)\big]$ are the same as the elements of B which are common to A (i.e. $A \cap B$) excluding the elements of A which are common to C (i.e. $A \cap C$).

We may also show this formally as follows (starting from the right hand side):

$$\begin{aligned}
(A \cap B) - (A \cap C) &= (A \cap B) \cap (\overline{A \cap C}) && \text{(see part a)} \\
&= (A \cap B) \cap (\overline{A} \cup \overline{C}) && \text{(Eq. 21)} \\
&= \big[(A \cap B) \cap \overline{A}\big] \cup \big[(A \cap B) \cap \overline{C}\big] && \text{(Eq. 19)} \\
&= \big[(A \cap \overline{A}) \cap B\big] \cup \big[(A \cap B) \cap \overline{C}\big] && \text{(Eqs. 14 and 17)} \\
&= \big[\varnothing \cap B\big] \cup \big[(A \cap B) \cap \overline{C}\big] && \text{(Eq. 11)} \\
&= \varnothing \cup \big[(A \cap B) \cap \overline{C}\big] && \text{(Eq. 3)} \\
&= (A \cap B) \cap \overline{C} && \text{(Eq. 4)}
\end{aligned}$$

[24] We note that this argument also shows that $A - B$, $B - A$ and $A \cap B$ are disjoint.

$$= A \cap (B \cap \overline{C}) \qquad \text{(Eq. 17)}$$
$$= A \cap (B - C) \qquad \text{(see part a)}$$

(**h**) We use the same (verbal) argument as the argument of (g). We may also obtain this formally from the relation of (g) by using the commutativity of intersection, i.e. by shifting the order of all intersections in the relation of (g) using Eq. 14.
(**i**) $A - (B \cup C)$ are the elements of A which belong neither to B nor to C. On the other hand, $A - B$ are the elements of A which do not belong to B (although they can belong to C) and $A - C$ are the elements of A which do not belong to C (although they can belong to B), and hence by taking their intersection [i.e. $(A - B) \cap (A - C)$] we select only the elements of A that belong neither to B nor to C (by excluding the elements of C from $A - B$ and excluding the elements of B from $A - C$) which is the same as $A - (B \cup C)$.
We may also show this formally as follows:

$$A-(B\cup C) = A\cap\overline{(B\cup C)} = A\cap(\overline{B}\cap\overline{C}) = (A\cap A)\cap(\overline{B}\cap\overline{C}) = (A\cap\overline{B})\cap(A\cap\overline{C}) = (A-B)\cap(A-C)$$

where the first and last equalities are based on the identity of part (a) while the third equality is based on Eq. 5.
(**j**)

$$(B \cup C) - A = (B \cup C) \cap \overline{A} = (B \cap \overline{A}) \cup (C \cap \overline{A}) = (B - A) \cup (C - A)$$

(**k**)

$$A - (B \cap C) = A \cap \overline{(B \cap C)} = A \cap (\overline{B} \cup \overline{C}) = (A \cap \overline{B}) \cup (A \cap \overline{C}) = (A - B) \cup (A - C)$$

(**l**)

$$(B \cap C) - A = (B \cap C) \cap \overline{A} = (B \cap C) \cap (\overline{A} \cap \overline{A}) = (B \cap \overline{A}) \cap (C \cap \overline{A}) = (B - A) \cap (C - A)$$

(**m**)

$$(B - C) - A = (B - C) \cap \overline{A} = (B \cap \overline{A}) - (C \cap \overline{A}) = (B - A) - (C - A)$$

where the second equality is based on the identity of part (h).
(**n**)

$$A - (B - C) = A - (B \cap \overline{C}) = (A - B) \cup (A - \overline{C}) = (A - B) \cup (A \cap C)$$

where the second equality is based on the identity of part (k).
(**o**)

$$U - (A \cup B) = \overline{A \cup B} = \overline{A} \cap \overline{B}$$

(**p**)

$$\overline{(B - A)} \cup \overline{(A - B)} = \overline{(B - A) \cap (A - B)} = \overline{\varnothing} = U$$

(**q**) It is obvious that what is common between A and $A \cup B$ is the entire A and hence $A \cap (A \cup B) = A$.
(**r**) It is obvious that $A \cap B$ are elements of A (i.e. those elements which are elements of B as well) and A is the entirety of A and hence when we "combine" them (i.e. take their union) we get the entire A and hence $A \cup (A \cap B) = A$.
(**s**)

$$A - B = A \cap \overline{B} = \overline{B} \cap A = \overline{B} - \overline{A}$$

PE: Justify the unjustified steps of all the formal derivations given above.

21. In Problem 20 we used two methods (or types of argument) to prove or justify the identities (and relations in general) of set theory, i.e. verbal arguments and formal arguments. Can you propose another method?
Answer: Another common method is the use of Venn diagrams (see § 2.4). So, we can say we have three main methods for proving and justifying the identities of set theory:
• **Verbal argument method** which is based on demonstrating the logic or rationale behind the

identity. The advantage of this method is that it shows the rationale of the identity and hence we acquire through this method an intuitive understanding and appreciation which will be helpful in various contexts in which the identity (and its alike) is used. The disadvantage of this method is that it may not be sufficiently rigorous. Moreover, it may require excessive explanations and justifications which are hard to follow and hence it could become susceptible to errors and delusions. Furthermore, this method is only applicable to simple identities because complicated identities are generally non-intuitive and hence they are not easy (or not possible) to prove by this method because their rationale is not simple to grasp or express verbally.

• **Formal (or analytical) argument method** which is based on using formal axioms and previously-proved identities, and hence the identity in question is proved in an analytical way. The advantage of this method is being analytical and hence it is rigorous and less susceptible to errors and delusions. Moreover, it is normally neat and concise as well as being general (i.e. it is applicable in principle to any identity regardless of its complexity or its other attributes). The disadvantage of this method is its potential complexity and difficulty (e.g. it requires a well structured and ordered set of axioms and previously-proved identities, and such structures may not be available or easy to maintain or track or obtain, and this is especially true in the case of casual proving of solitary identities).

• **Venn diagrams method** where the left and right hand sides of the identity are represented (usually in stages) by Venn diagrams which, if the identity is correct, should be identical (otherwise the alleged identity is false). The advantage of this method is its simplicity (because producing Venn diagrams is usually trivial) as well as its visual nature which may help acquiring intuitive understanding and appreciation of the identity. Moreover, it should be fast and easy to do in most cases.[25] The disadvantage of this method is that it is only applicable to simple identities because complicated identities are generally difficult (or almost impossible) to express or demonstrate by Venn diagrams (i.e. alone although the Venn diagrams can be used partially in conjunction with the verbal argument method for instance).

We should finally note that it may be necessary (or convenient or helpful) to combine the above methods (and possibly other methods) to prove or justify some identities where various methods are used to build various parts of the proof. So in brief, those who do work on set theory should be aware of all these possibilities (and possibly other possibilities) when they intend to tackle a problem of this type (i.e. a problem in which a proof or justification is required). Also see Problem 4 of § 2.4.

PE: Try to prove some of the identities of Problem 20 by the method of Venn diagrams (referring for this purpose to § 2.4).

22. Are you aware of a set operation other than those investigated above?[26]

Answer: Yes. For example, the Cartesian multiplication (or Cartesian product) of sets A_i $(i = 1, 2, \ldots, n)$ is defined as the set of all ordered n-tuples $(a_1, a_2, \ldots, a_n)$ where $a_1 \in A_1$, $a_2 \in A_2$, ..., $a_n \in A_n$, .

PE: Form the Cartesian product of the sets $A = \{a, b, c, d, e, f\}$ and $B = \{\alpha, \beta, \gamma, \delta\}$.

2.2 The Rules of Counting

The rules of counting are generally based on the **fundamental principle of counting**[27] which states that if we have N things (e.g. events or actions or choices or ... etc.) labeled as $O_1, O_2, \ldots, O_N$ where O_1 can occur in m_1 different ways, O_2 can occur in m_2 different ways, ..., and O_N can occur in m_N different ways then the number of possibilities for the occurrence of these things is given by the product

[25] In fact, this is true if only manual sketches are required. However, the job of producing Venn diagrams could become lengthy and excessive if the diagrams are required to be of artistic quality (e.g. to be used in a scientific paper or a book) and this should require professional or computerized skills and resources and hence it becomes a disadvantage rather than an advantage.

[26] The purpose of this question is mainly to make the reader aware that there are set operations other than those which we investigated, and hence the investigated operations represent what are of interest to us in this book.

[27] This principle is also known by similar names such as the fundamental counting rule or the principle of counting or the rule of multiplication of choices or multiplicative rule of counting.

$m_1 \times m_2 \times \cdots \times m_N$. In fact, we have two main cases for the application of this principle; these cases will be investigated in Problem 1.

The **factorial** of a positive integer is the product of all the positive integers from 1 to that integer. The factorial of a number n is symbolized as $n!$. Accordingly:

$$n! = 1 \times 2 \times \cdots \times (n-1) \times n = n \times (n-1) \times \cdots \times 2 \times 1 \tag{23}$$

By convention, the factorial of 0 is 1, i.e. $0! = 1$.

A **permutation** of a set of objects is a particular arrangement of these objects (noting that "arrangement" means the order matters). In simple words, a permutation is a set with a given order of its members.[28] For example, ab and ba are two permutations of the set $\{a, b\}$.[29] The number of different permutations of m objects taken from a set of n different objects ($0 \leq m \leq n$) is given by:

$$P_m^n = \frac{n!}{(n-m)!} = n \times (n-1) \times \cdots \times (n-m+1) \tag{24}$$

An important special case of this formula is $P_n^n = n!$.

A **combination** is a particular grouping of objects with no consideration of their order.[30] For example, if we consider the combinations of two objects taken from the set $\{a, b, c\}$ then ab and ac are two different combinations but ab and ba represent the same combination.[31] The number of different combinations of m objects taken from a set of n different objects ($0 \leq m \leq n$) is given by:

$$C_m^n = \frac{n!}{m!(n-m)!} = \frac{n \times (n-1) \times \cdots \times (n-m+1)}{m!} \tag{25}$$

By comparing Eqs. 24 and 25 we conclude:

$$C_m^n = \frac{P_m^n}{m!} \tag{26}$$

It is worth noting that P_m^n is the number of different permutations of m objects selected from a set of n different objects assuming *no repetition* (in the selected objects) is allowed. If repetition is allowed then the number of permutations is:

$$P_{n,m} = n^m \tag{27}$$

Similarly, C_m^n is the number of different combinations of m objects selected from a set of n different objects assuming *no repetition* (in the selected objects) is allowed. If repetition is allowed then the number of combinations is:

$$C_{n,m} = C_m^{n+m-1} = \frac{(n+m-1)!}{m!\,(n-1)!} \tag{28}$$

Further clarifications about these issues will follow (see for example Problems 5 and 30).[32]

Problems

1. Give more details about the two cases for the application of the fundamental principle of counting.
   Answer: In one case the number of possibilities for the occurrence of these things are independent of each other and hence we multiply the number of possibilities as they are, while in the other case the number of possibilities (or some of them) are dependent on each other (and possibly on the sequence of

[28] In more technical terms, a permutation of m objects taken from a set of n objects ($0 \leq m \leq n$) is a tuple of size m of that set. So, if A is a set of size n then its m-permutations are the set of all its m-tuples.

[29] Or rather (a, b) and (b, a).

[30] In more technical terms, a combination of m objects taken from a set of n objects ($0 \leq m \leq n$) is a subset of size m of that set (and hence order is irrelevant). So, if A is a set of size n then its m-combinations are the set of all its subsets of size m.

[31] Or rather $\{a, b\}$ and $\{a, c\}$.

[32] It is important to note that in this book we generally use "permutation" and "combination" to refer to those without repetition (in the selected objects) unless stated otherwise.

occurrence) and hence we multiply the number of available possibilities considering this dependency. For example, if we have to choose a committee of one male and one female from a city council made of 7 males and 10 females then we have 7 choices for the selection of male and 10 choices for the selection of female (i.e. a total of $10 \times 7 = 70$ possibilities) and these selections are obviously independent of each other and of the sequence of selection (because male and female are distinct and hence the number of male/female candidates is independent of the number of female/male candidates). However, if we have to choose a president and a vice president from this council (regardless of their gender) then we have obviously 17 choices for selecting one of them and 16 (not 17) for selecting the other (i.e. a total of $17 \times 16 = 272$ possibilities) because the selection of one will affect the number of the available choices for the selection of the other (since no one can be president and vice president at the same time) and hence the number of available choices is dependent on each other.

Note 1: the significance of these cases appears in typical situations indicated by words like "repetition" and "replacement" (see § 1.5). In general, if the chosen objects can be repetitive (see for instance Problem 10) or are replaced (see for instance Problem 14) then the number of choices are independent of each other; otherwise the number of choices are not independent. However, these are not the only factors that can determine dependency (in fact there are many other factors that can determine dependency).

Note 2: in the second case (i.e. when there is dependency) the order of the occurrence is generally important, i.e. the number of possibilities could depend on the order of occurrence. For instance, in the above example of a city council (made of 7 males and 10 females) if we have to choose a president (regardless of gender) and a *female* vice president then if we start by choosing the vice president then we have $10 \times 16 = 160$ possibilities but if we start by choosing the president then we have either $17 \times 10 = 170$ possibilities (if the chosen president is male) or $17 \times 9 = 153$ possibilities (if the chosen president is female).

Note 3: to avoid unnecessary complications (e.g. related to the difference between the two cases) we generalize the fundamental principle of counting as follows: if we have N things labeled as $O_1, O_2, \ldots, O_N$ where O_1 can occur in m_1 different ways, O_2 can occur in m_2 different ways (given the choice of O_1), ..., and O_N can occur in m_N different ways (given the choice of $O_1, O_2, \ldots, O_{N-1}$) then the number of possibilities for the occurrence of these things is given by the product $m_1 \times m_2 \times \cdots \times m_N$.[33]

PE: Give a number of examples for each one of the two cases for the application of the fundamental principle of counting.

2. Use the fundamental principle of counting to justify:

(**a**) The rule of permutation (i.e. Eq. 24). (**b**) The rule of combination (i.e. Eq. 25).

Answer:

(**a**) The first object (of the m objects) can be selected in n different ways (i.e. the number of n objects), the second object can be selected in $n-1$ different ways (i.e. the number of the remaining $n-1$ objects), ... and finally the m^{th} object can be selected in $n-(m-1) = n-m+1$ different ways. So, by the fundamental principle of counting we get:

$$P_m^n = n \times (n-1) \times \cdots \times (n-m+1) = \frac{n \times (n-1) \times \cdots \times (n-m+1) \times (n-m)!}{(n-m)!} = \frac{n!}{(n-m)!}$$

(**b**) The difference between P_m^n and C_m^n is that the order (of the m objects) matters in P_m^n but not in C_m^n. This means that we can obtain C_m^n from P_m^n by dividing P_m^n by the number of permutations of m objects taken from m objects, i.e. P_m^m. Now, from part (a) we have $P_m^m = \frac{m!}{(m-m)!} = m!$ (noting that $0! = 1$) and hence:

$$C_m^n = \frac{P_m^n}{P_m^m} = \frac{P_m^n}{m!} = \frac{n!}{m!(n-m)!}$$

[33] We note that the order of selection (if there is order) is considered by the condition "given the choice of". If there is no order and the choices are not mutually independent then we should consider the dependency of the choice of each object on the choice(s) of other object(s) when we have such a dependency (which can possibly be partial as well as multiple).

PE: Use the fundamental principle of counting to find and justify the number of possible outcomes when you:

(**a**) Throw a die n times. (**b**) Throw m coins simultaneously n times.

(**c**) Throw m dice simultaneously n times. (**d**) Throw a die and a coin simultaneously n times.

3. State the **sum rule of counting** for a number of pairwise disjoint sets.
Answer: If $A_1, A_2, \ldots, A_n$ are pairwise disjoint sets then the number of elements of their union is the sum of the numbers of elements of these sets, that is:

$$N = N_1 + N_2 + \cdots + N_n \qquad (A_i \cap A_j = \varnothing)$$

where N is the number of elements of their union (i.e. $A_1 \cup A_2 \cup \ldots \cup A_n$) and $N_1, N_2, \ldots, N_n$ are the numbers of elements of these sets (with $i, j \in 1, 2, \ldots, n$ and $i \neq j$).
PE: What change should we introduce on this rule of counting if the sets are not pairwise disjoint? Consider in your answer only some simple cases (e.g. 5 pairwise disjoint sets except 2 of them which have common elements).

4. Express the factorial of n in another commonly used mathematical form.
Answer: If we use the product symbol Π then we have:

$$n! \equiv \prod_{k=1}^{n} k$$

PE: Investigate the relationship between the factorial and the gamma function (if you ever heard of the gamma function).

5. Let "permutation" mean "selection with order" and "combination" mean "selection with no order". Also, let "repetition" mean "repetition in the selected objects". State the formulae for the number of different choices that can be made when selecting k objects out of n different objects (considering the factors of order and repetition).
Answer: We have four formulae:

$$P_k^n = \frac{n!}{(n-k)!} \qquad \text{(permutation with no repetition)} \tag{29}$$

$$P_{n,k} = n^k \qquad \text{(permutation with repetition)} \tag{30}$$

$$C_k^n = \frac{n!}{k!(n-k)!} \qquad \text{(combination with no repetition)} \tag{31}$$

$$C_{n,k} \equiv C_k^{n+k-1} = \frac{(n+k-1)!}{k!(n-1)!} \qquad \text{(combination with repetition)} \tag{32}$$

Also see Problem 30.
PE: Try to identify a different type of repetition, i.e. other than "repetition in the selected objects".

6. Find the condition for:
(**a**) $P_m^n = C_m^n$. (**b**) $P_m^n = n^m$. (**c**) $C_m^n = n^m$.
Answer: We note that $m \leq n$.
(**a**)

$$\begin{aligned} P_m^n = \frac{n!}{(n-m)!} &= \frac{n!}{m!(n-m)!} = C_m^n \\ \frac{1}{1} &= \frac{1}{m!} \\ m! &= 1 \end{aligned}$$

i.e. $m = 0, 1$. So, we have $P_0^n = C_0^n = 1$ $(n \geq 0)$, and $P_1^n = C_1^n = n$ $(n \geq 1)$.
(**b**)

$$P_m^n = \frac{n!}{(n-m)!} = n^m$$

$$n \times \cdots \times (n-m+1) \quad = \quad n \times \cdots \times n$$

i.e. $m = 1$. So, $P_1^n = n^1 = n$ $(n \geq 1)$. Also noting that $P_0^n = n^0 = 1$ $(n \geq 1)$ we should also have $P_0^n = n^0$ $(n \geq 1)$. So in brief, $P_m^n = n^m$ for $n \geq 1$ and $m = 0, 1$.
(**c**)

$$\begin{aligned} C_m^n = \frac{n!}{m!(n-m)!} &= n^m \\ \frac{n}{m} \times \cdots \times \frac{(n-m+1)}{1} &= n \times \cdots \times n \end{aligned}$$

i.e. $m = 1$. So, $C_1^n = n^1 = n$ $(n \geq 1)$. Also noting that $C_0^n = n^0 = 1$ $(n \geq 1)$ we should also have $C_0^n = n^0$ $(n \geq 1)$. So in brief, $C_m^n = n^m$ for $n \geq 1$ and $m = 0, 1$.
PE: Justify all the details of the above arguments.

7. Find the following numbers of permutations: P_0^0, P_0^n, P_n^n, P_3^7, P_2^9, P_6^{13}.
Answer: We use Eq. 24:

$$P_0^0 = \frac{0!}{(0-0)!} = 1 \qquad P_0^n = \frac{n!}{(n-0)!} = 1 \qquad P_n^n = \frac{n!}{(n-n)!} = n!$$

$$P_3^7 = \frac{7!}{(7-3)!} = 210 \qquad P_2^9 = \frac{9!}{(9-2)!} = 72 \qquad P_6^{13} = \frac{13!}{(13-6)!} = 1235520$$

PE: Find the following numbers of permutations: P_{n-1}^n, P_{n-2}^{n+1}, P_1^n, P_6^{12}, P_5^{15}, P_3^{20}.

8. Find the following numbers of combinations: C_0^0, C_0^n, C_n^n, C_2^5, C_5^8, C_4^{11}.
Answer: We use Eq. 25:

$$C_0^0 = \frac{0!}{0!(0-0)!} = 1 \qquad C_0^n = \frac{n!}{0!(n-0)!} = 1 \qquad C_n^n = \frac{n!}{n!(n-n)!} = 1$$

$$C_2^5 = \frac{5!}{2!(5-2)!} = 10 \qquad C_5^8 = \frac{8!}{5!(8-5)!} = 56 \qquad C_4^{11} = \frac{11!}{4!(11-4)!} = 330$$

PE: Find the following numbers of combinations: C_{n-1}^n, C_{n-2}^{n+1}, C_1^n, C_{20}^{24}, C_2^{11}, C_3^9.

9. Plot C_m^n as a function of n and m for $1 \leq n \leq 12$ to appreciate how C_m^n varies with n and m.
Answer: See Figure 1. We note that because of the large range of C_m^n (i.e. on the vertical 'z' axis) some details are obscured.
PE: Plot P_m^n as a function of n and m for $1 \leq n \leq 7$.

10. How many 5-letter strings[34] can be made from the 26 letters of the English alphabet if (**a**) repetition is allowed (**b**) repetition is not allowed?
Answer:
(**a**) If repetition is allowed then we have 26 possibilities for each one of the 5 letters. Hence, by the fundamental principle of counting the number of possibilities of 5-letter strings is $26^5 = 11881376$. In fact, this is just $P_{26,5}$.
(**b**) If repetition is not allowed then we have 26 possibilities for the 1^{st} letter, 25 possibilities for the 2^{nd} letter, 24 possibilities for the 3^{rd} letter, 23 possibilities for the 4^{th} letter, and 22 possibilities for the 5^{th} letter. Hence, by the fundamental principle of counting the number of possibilities of 5-letter strings is $26 \times 25 \times 24 \times 23 \times 22 = 7893600$. In fact, this is just P_5^{26}.
PE: How many different choices we have in drawing 2 red balls and 3 blue balls from a bag containing 7 red (numbered) balls and 9 blue (numbered) balls considering the cases of (**a**) with replacement and (**b**) without replacement?

[34] We mean by "strings" arrangements of letters which are not necessarily sensible words.

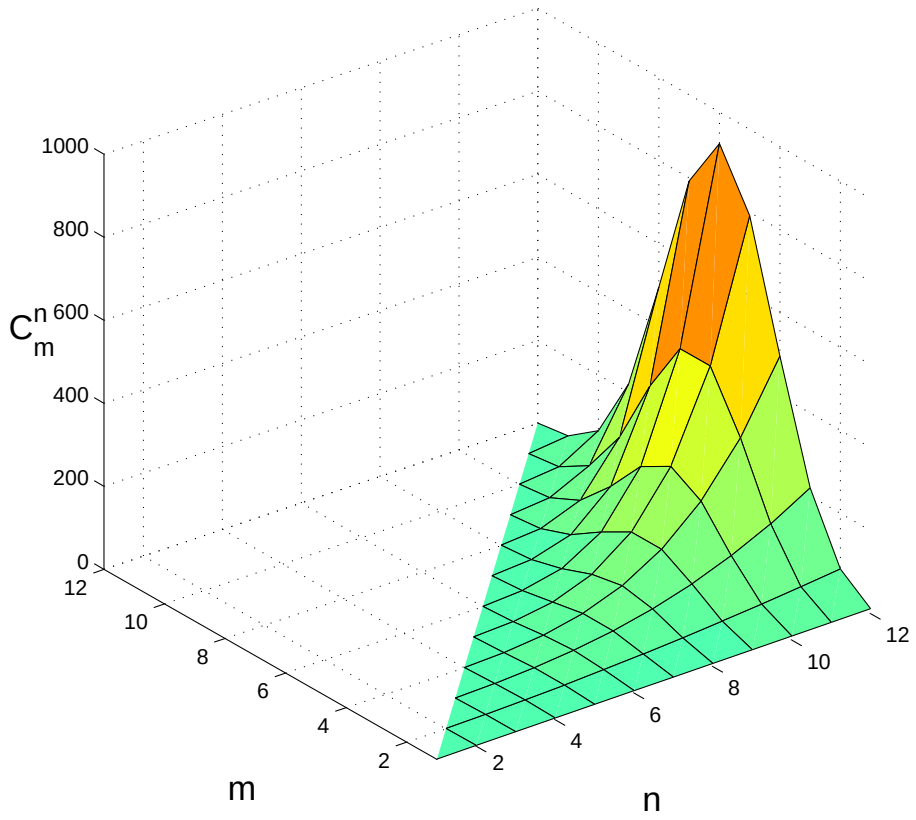


Figure 1: The plot of Problem 9 of § 2.2.

11. List (in a compact way) all the possibilities of the 5-letter strings in parts (a) and (b) of Problem 10.
Answer: If W symbolizes the set of these possibilities then we have:
(**a**) $W = \Big\{\alpha\beta\gamma\delta\varepsilon :\ \alpha,\beta,\gamma,\delta,\varepsilon \in \{\text{English alphabet}\}\Big\}$.
(**b**) $W = \Big\{\alpha\beta\gamma\delta\varepsilon :\ \alpha,\beta,\gamma,\delta,\varepsilon \in \{\text{English alphabet}\},\ \alpha \neq \beta \neq \gamma \neq \delta \neq \varepsilon\Big\}$ where $\neq$ means "different" or "not the same".
PE: List (in a compact way) all the possibilities of the choices of the PE of Problem 10.
12. List all the 3-letter strings that can be made from the letters {a,b,c} if (**a**) repetition is allowed (**b**) repetition is not allowed.
Answer:
(**a**) We have $3^3 = 27$ distinct strings which are:

aaa	bbb	ccc	aab	aac	aba	aca	baa	caa
bba	bbc	bab	bcb	abb	cbb	cca	ccb	cac
cbc	acc	bcc	abc	acb	bac	bca	cab	cba

(**b**) We have $P_3^3 = 6$ distinct strings which are:

abc	acb	bac	bca	cab	cba

PE: List all the 3-digit numbers that can be made from the digits {1,2,3} if (**a**) repetition is allowed (**b**) repetition is not allowed.
13. List all the 3-digit (**a**) permutations (**b**) combinations of the digits {1,2,3,4}.
Answer:

(**a**) We have $P_3^4 = 24$ permutations which are:

123	124	132	134	142	143	213	214	231	234	241	243
312	314	321	324	341	342	412	413	421	423	431	432

(**b**) We have $C_3^4 = 4$ combinations which are:

{1,2,3} {1,2,4} {1,3,4} {2,3,4}

PE: List all the 3-letter (**a**) permutations (**b**) combinations of the letters {a,b,c,d}.

14. In a gambling game, 5 balls are drawn randomly from an urn containing 9 balls (numbered from 1 to 9). What is the number of possibilities for this game if:
(**a**) The balls are drawn at once.
(**b**) The balls are drawn in sequence with replacement (i.e. each drawn ball is returned to the urn before drawing the next ball).
(**c**) The balls are drawn in sequence with no replacement.
Answer:
(**a**) Because the balls are drawn at once there is no order and hence this is a problem of combination. So, what is required is to find the number of combinations of 5 objects taken from a set of 9 objects, i.e. $C_5^9 = 126$ possibilities.
(**b**) We have order (as suggested by "sequence") with replacement, and hence we have 9 possibilities for each one of the 5 balls. So, by the fundamental principle of counting we have $9^5 = 59049$ possibilities.
(**c**) We have order (as suggested by "sequence") with no replacement, and hence this is a problem of permutation.[35] So, what is required is to find the number of permutations of 5 objects taken from a set of 9 objects, i.e. $P_5^9 = 15120$ possibilities.
PE: How the results of this Problem will be affected if the balls are colored: 3 blue, 3 green and 3 red? Justify your answer eloquently.
15. Classify the following as permutation or combination problems:
(**a**) Using three (different) medicines: one at morning, one at midday and one at night.
(**b**) Selecting five members of parliament for a parliamentary committee.
(**c**) Selecting a president, a prime minister and a secretary of state (from a ruling party).
Answer:
(**a**) This is a permutation problem because the (chronological) order is important.
(**b**) This is a combination problem because there is no indication of significance of order (i.e. the members of committee are selected and treated equally).
(**c**) This is a permutation problem because the (ranking) order is important.
PE: Repeat the Problem for the following:
(**a**) Selecting a football team for a match from the players of a football club (assuming any player in the club can take any role in the match).
(**b**) Choosing 5 cars of different models from 20 available models.
(**c**) Aligning the students of a class in a queue.
16. Two women and three men are to be selected from a list of candidates made of 13 women and 17 men. How many possibilities we have for this selection?
Answer: The order in the selection of women and in the selection of men is irrelevant and hence we have C_2^{13} possibilities for the selection of women and C_3^{17} possibilities for the selection of men. So, by the fundamental principle of counting the number of possibilities for the selection of 2 women and 3 men is the product of C_2^{13} and C_3^{17}, i.e. $C_2^{13} \times C_3^{17} = 78 \times 680 = 53040$.
PE: How the result of this Problem will change if:
(**a**) The gender of the selected 5 is irrelevant.
(**b**) We have to remove 4 men candidates before selecting the 2 women and 3 men.
17. A passport identification number is made of an uppercase English letter followed by a 7-digit number. How many distinct passports can be issued if:

[35] In fact, being a problem of permutation or combination depends on the rules of the game (and this should apply even to part b). However, we treated this as a problem of permutation (due to the strong suggestion of "sequence").

(**a**) The letter is consonant and the digits are not repetitive.
(**b**) There is no restriction on the letter and the digits but the 7-digit number must start (from left) with a non-zero digit.
Answer:
(**a**) We have 21 possibilities for the letter (because we have 21 consonant letters) and P_7^{10} possibilities for the 7-digit number (because we have 10 possibilities for the 1^{st} digit, 9 possibilities for the 2^{nd} digit, ... , and 4 possibilities for the 7^{th} digit). Hence, by the fundamental principle of counting the number of distinct passports is:

$$21 \times P_7^{10} = 21 \times \frac{10!}{(10-7)!} = 12700800$$

(**b**) We have 26 possibilities for the letter (because we have 26 letters in the English alphabet), 9 possibilities for the 1^{st} digit (because it cannot be 0) and 10 possibilities for each one of the remaining 6 digits of the 7-digit number (because we have no restriction on them). Hence, by the fundamental principle of counting the number of distinct passports is:

$$26 \times 9 \times 10^6 = 234000000$$

PE: Repeat the Problem for the following cases:
(**a**) There is no restriction on the letter and the digits but the 7-digit number cannot be less than a million.
(**b**) The letter cannot be O or I (to avoid confusion with 0 and 1) and the digits cannot be 0.

18. The constitution of a (liberal democratic) country states that at least 1/3 of the members of parliament must be women. If the number of members of parliament is 60, how many possibilities we have for the representation of women in this parliament?
Answer: 1/3 of 60 is 20. Hence, the number of women must be between 20 and 60 (inclusive). This means that we have 41 possibilities, i.e. $60 - 19 = 41$.
PE: Repeat the Problem if the constitution requires that both genders must be represented and the representation of women must not be less than 1/4.
19. A football club consists of 2 goalkeepers and 15 players. How many possibilities we have for selecting a team (of 1 goalkeeper and 10 players) for a match?
Answer: There are two possibilities for choosing the goalkeeper and C_{10}^{15} for choosing the 10 players (noting that the order is irrelevant) and hence by the fundamental principle of counting the number of possibilities is $2 \times C_{10}^{15} = 2 \times 3003 = 6006$.
PE: What if the club consists of 2 goalkeepers, 4 defenders, 6 midfielders and 5 strikers and we want a team of 1 goalkeeper, 3 defenders, 4 midfielders and 3 strikers?
20. Mention a well-known use of C_m^n in algebra.
Answer: C_m^n appears in the expansion of algebraic expressions like $(x+y)^n$ by the **binomial theorem**, i.e.

$$(x+y)^n = \sum_{k=0}^{n} C_k^n x^{n-k} y^k = C_0^n x^n y^0 + C_1^n x^{n-1} y^1 + \cdots + C_k^n x^{n-k} y^k + \cdots + C_{n-1}^n x^1 y^{n-1} + C_n^n x^0 y^n \quad (33)$$

For this reason, C_m^n is also known as the **binomial coefficient**.
Note: a simple mathematical device for calculating the binomial coefficients is the Pascal triangle[36] (see Figure 2) which is characterized by (and constructed according to) the following:
• The two edges of the triangle are 1's.
• The n^{th} row of the triangle has $n+1$ entries (where $n = 0, 1, 2, \ldots$).

[36] In fact, the Pascal triangle has uses and benefits other than (and more important than) calculating the binomial coefficients such as demonstrating their relations and patterns (like symmetries) and deriving mathematical relations and identities involving these coefficients. We also note that calculating the binomial coefficients by the Pascal triangle is practical only for small coefficients (and it is generally trivial).

$n=0$						1						
$n=1$					1		1					
$n=2$				1		2		1				
$n=3$			1		3		3		1			
$n=4$		1		4		6		4		1		
$n=5$	1		5		10		10		5		1	
$n=6$	1		6		15		20		15		6	1

Figure 2: The first 7 rows of the Pascal triangle. See Problem 20 of § 2.2.

• Each internal entry is the sum of the nearest two entries (on its two sides) in the row above it.
• The entries in the n^{th} row of the triangle are the binomial coefficients in the expansion of binomial expressions like $(x+y)^n$.
• The triangle is symmetric (by mirror reflection) in the vertical line passing through its top vertex.
PE: Investigate the potential use of permutations in the mathematics of group theory, number theory, genetics, computer science, cryptography and fractals (as well as other branches and fields of pure and applied mathematics and related disciplines).

21. Show that for any $n \in \mathbb{N}$ we have (with a_k being constants):
(**a**) $3^n = \sum_{k=0}^{n} a_k\, 2^k$. (**b**) $5^n = \sum_{k=0}^{n} a_k\, 2^{2k}$.
Answer: We use the binomial theorem (Eq. 33).
(**a**)

$$3^n = (1+2)^n = \sum_{k=0}^{n} C_k^n\ 1^{n-k}\ 2^k = \sum_{k=0}^{n} C_k^n\, 2^k = \sum_{k=0}^{n} a_k\, 2^k \qquad (a_k = C_k^n)$$

(**b**)

$$5^n = (1+4)^n = \sum_{k=0}^{n} C_k^n\ 1^{n-k}\ 4^k = \sum_{k=0}^{n} C_k^n\, 4^k = \sum_{k=0}^{n} C_k^n\, 2^{2k} = \sum_{k=0}^{n} a_k\, 2^{2k} \qquad (a_k = C_k^n)$$

PE: Show that for any $n \in \mathbb{N}$ we have $9^n = \sum_{k=0}^{n} a_k 2^{3k}$. Also give a general formula representing the pattern seen in the examples of this Problem.

22. Rewrite the Pascal triangle of Figure 2 in terms of the combination symbols (i.e. binomial coefficients).
Answer: See Figure 3.
PE: Describe the pattern of the combination symbols in Figure 3. Also construct the next 3 rows (corresponding to $n = 7, 8, 9$) of this triangle.

23. Prove the following permutation identities:
(**a**) $P_n^n = n!$. (**b**) $P_{n-1}^n = n!$ $(n \geq 1)$. (**c**) $P_{n-1}^n = P_n^n$.
(**d**) $P_m^n = nP_{m-1}^{n-1}$. (**e**) $P_k^n / P_{k-1}^n = n-k+1$. (**f**) $P_k^n + kP_{k-1}^n = P_k^{n+1}$.
Answer: We use Eq. 24.
(**a**) We have:

$$P_n^n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$$

$$
\begin{array}{lccccccccccc}
n=0 & & & & & & C_0^0 & & & & & \\
n=1 & & & & & C_0^1 & & C_1^1 & & & & \\
n=2 & & & & C_0^2 & & C_1^2 & & C_2^2 & & & \\
n=3 & & & C_0^3 & & C_1^3 & & C_2^3 & & C_3^3 & & \\
n=4 & & C_0^4 & & C_1^4 & & C_2^4 & & C_3^4 & & C_4^4 & \\
n=5 & C_0^5 & & C_1^5 & & C_2^5 & & C_3^5 & & C_4^5 & & C_5^5 \\
n=6 & C_0^6 & & C_1^6 & & C_2^6 & & C_3^6 & & C_4^6 & & C_5^6 \quad C_6^6
\end{array}
$$

Figure 3: The first 7 rows of the Pascal triangle in terms of the combination symbols. See Problem 22 of § 2.2.

(**b**) We have:

$$P_{n-1}^n = \frac{n!}{(n-n+1)!} = \frac{n!}{1!} = \frac{n!}{1} = n!$$

(**c**) This can be obtained by combining the results of parts (a) and (b).
(**d**) We have:

$$nP_{m-1}^{n-1} = n \times \frac{(n-1)!}{(n-1-m+1)!} = \frac{n!}{(n-m)!} = P_m^n$$

(**e**) We have:

$$\frac{P_k^n}{P_{k-1}^n} = \frac{n!}{(n-k)!} \times \frac{(n-k+1)!}{n!} = \frac{(n-k+1)!}{(n-k)!} = n-k+1$$

(**f**) We have:

$$
\begin{aligned}
P_k^n + kP_{k-1}^n &\stackrel{?}{=} P_k^{n+1} \\
\frac{n!}{(n-k)!} + k\,\frac{n!}{(n-k+1)!} &\stackrel{?}{=} \frac{(n+1)!}{(n+1-k)!} && \text{(Eq. 24)} \\
\frac{n!}{(n-k+1)!}\left[\frac{(n-k+1)!}{(n-k)!} + k\right] &\stackrel{?}{=} \frac{(n+1)!}{(n+1-k)!} && \text{(factorizing)} \\
\frac{(n-k+1)!}{(n-k)!} + k &\stackrel{?}{=} n+1 && \text{(canceling)} \\
(n-k+1) + k &\stackrel{?}{=} n+1 && \text{(simplifying)} \\
n+1 &= n+1
\end{aligned}
$$

PE: Prove part (f) by another method.

24. Find the unknown x in the following (where x is a positive or non-negative integer):
(**a**) $P_x^n = x!$. (**b**) $P_{x-1}^n = n!$. (**c**) $6P_{x-3}^x = n!$. (**d**) $3P_5^x = P_6^x$.

Answer:
(**a**) We have:

$$
\begin{aligned}
P_x^n &= x! \\
n \times (n-1) \times \cdots \times (n-x+1) &= x \times (x-1) \times \cdots \times 1
\end{aligned}
$$

So, by comparison we get $x = n$.
(**b**) We have:

$$\begin{aligned} P^n_{x-1} &= n! \\ \frac{n!}{(n-x+1)!} &= \frac{n!}{1} \end{aligned}$$

So, by comparing the denominators we have either $n-x+1=0$ and hence $x=n+1$ or $n-x+1=1$ and hence $x=n$.
(**c**) We have:

$$\begin{aligned} 6P^x_{x-3} &= n! \\ P^x_{x-3} &= \frac{n!}{3!} \\ x\times(x-1)\times\cdots\times 4 &= n\times(n-1)\times\cdots\times 4 \end{aligned}$$

So, by comparison we get $x = n$ $(n \geq 3)$.
(**d**) We have:

$$\begin{aligned} 3P^x_5 &= P^x_6 \\ 3\big[x(x-1)(x-2)(x-3)(x-4)\big] &= x(x-1)(x-2)(x-3)(x-4)(x-5) \\ 3 &= (x-5) \qquad (x \neq 0,1,2,3,4 \text{ since } x \geq 6) \\ x &= 8 \end{aligned}$$

PE: Form and solve two other relations (involving permutations with an unknown) similar to the relations given in this Problem.

25. Prove the following combination identities:

(**a**) $C^n_{n-m} = C^n_m$. (**b**) $C^n_{m-1} + C^n_m = C^{n+1}_m$. (**c**) $C^{n-1}_{m-1} + C^{n-1}_m = C^n_m$.
(**d**) $C^n_m + C^n_{m+1} = C^{n+1}_{m+1}$. (**e**) $C^n_m = \frac{n}{m} C^{n-1}_{m-1}$. (**f**) $C^n_m C^m_k = C^n_k C^{n-k}_{m-k}$.
(**g**) $C^{m+n}_q = \sum_{p=0}^{q} C^m_p\, C^n_{q-p}$. (**h**) $\sum_{i=m}^{n} C^i_m = C^{n+1}_{m+1}$. (**i**) $C^{2n}_n = \sum_{p=0}^{n} \left(C^n_p\right)^2$.
(**j**) $\sum_{k=0}^{n} C^n_k = 2^n$. (**k**) $\sum_{p=0}^{m} C^{n+p}_p = C^{n+m+1}_m$. (**l**) $C^{2n}_2 = 2C^n_2 + n^2$.
(**m**) $C^{3n}_2 = 3C^n_2 + 3n^2$. (**n**) $C^{-n}_m = (-1)^m\, C^{n+m-1}_m$. (**o**) $\sum_{r=0}^{n-1} C^{k+r}_k = C^{k+n}_{k+1}$.

Answer:
(**a**) From Eq. 25 we have:[37]

$$C^n_{n-m} = \frac{n!}{(n-m)!\,(n-[n-m])!} = \frac{n!}{(n-m)!\,m!} = \frac{n!}{m!(n-m)!} = C^n_m$$

(**b**) This is the **Pascal identity** which produces the interior entries of the $(n+1)^{th}$ row of the Pascal triangle (see Problems 20 and 22) from the entries of the n^{th} row of the triangle.[38] From Eq. 25 we have:

$$\begin{aligned} C^n_{m-1} + C^n_m &= \frac{n!}{(m-1)!(n-m+1)!} + \frac{n!}{m!(n-m)!} \\ &= \frac{m\times n!}{m!(n-m+1)!} + \frac{(n-m+1)\times n!}{m!(n-m+1)!} \end{aligned}$$

[37] This equation is intuitive because for each combination of m objects (chosen from n objects) there is a corresponding combination of $n-m$ objects (left out of the combination of m objects) and hence the number of the two combinations must be equal, i.e. $C^n_{n-m} = C^n_m$.
[38] In fact, this identity can be inferred from Figures 2 and 3 (noting that the triangle is actually constructed using this identity).

$$\begin{aligned} &= \frac{\big[m \times n!\big] + \big[(n-m+1) \times n!\big]}{m!(n-m+1)!} \\ &= \frac{n!\big[m + (n-m+1)\big]}{m!(n-m+1)!} \\ &= \frac{n!(n+1)}{m!(n-m+1)!} \\ &= \frac{(n+1)!}{m!(n+1-m)!} = C_m^{n+1} \end{aligned}$$

(**c**) This is the same as the identity of part b (i.e. Pascal identity) with $n-1$ replacing n.
(**d**) This is the same as the identity of part b (i.e. Pascal identity) with $m+1$ replacing m.
(**e**) We have:

$$\frac{n}{m} C_{m-1}^{n-1} = \frac{n}{m} \times \frac{(n-1)!}{(m-1)!\,(n-1-m+1)!} = \frac{n!}{m!\,(n-m)!} = C_m^n$$

(**f**) We have:

$$\begin{aligned} C_m^n C_k^m &= \frac{n!}{m!(n-m)!} \times \frac{m!}{k!(m-k)!} = \frac{n!}{(n-m)!} \times \frac{1}{k!(m-k)!} = \frac{n!}{k!} \times \frac{1}{(m-k)!(n-m)!} \\ &= \frac{n!}{k!(n-k)!} \times \frac{(n-k)!}{(m-k)!(n-m)!} = \frac{n!}{k!(n-k)!} \times \frac{(n-k)!}{(m-k)!(n-k-m+k)!} = C_k^n C_{m-k}^{n-k} \end{aligned}$$

(**g**) This is the **Vandermonde identity**. Let assume that we have m members of a given club and n non-members and we want to form a team of q individuals from these $m+n$ individuals. It is obvious that we have C_q^{m+n} ways for the formation of this team (which is the left hand side of this identity). However, we can look to the formation of this team differently by taking p individuals from the members (which can be done in C_p^m different ways) and $q-p$ individuals from the non-members (which can be done in C_{q-p}^n different ways) and hence we have $C_p^m\, C_{q-p}^n$ ways for the formation of this team. Now, since p can vary between 0 and q (inclusive) then the total number of ways for the formation of this team is $\sum_{p=0}^{q} C_p^m\, C_{q-p}^n$ (which is the right hand side of this identity). So, the left hand side and the right hand side of this identity are equal (as required).
(**h**) This is the **hockey-stick identity**. We have:

$$\begin{aligned} \sum_{i=m}^{n} C_m^i &= C_m^m + C_m^{m+1} + C_m^{m+2} + \cdots + C_m^{m+k} && (m+k=n) \\ &= 1 + C_m^{m+1} + C_m^{m+2} + \cdots + C_m^{m+k} && (C_m^m = 1) \\ &= C_{m+1}^{m+1} + C_m^{m+1} + C_m^{m+2} + \cdots + C_m^{m+k} && (C_{m+1}^{m+1} = 1) \\ &= C_{m+1}^{m+2} + C_m^{m+2} + \cdots + C_m^{m+k} && \text{(Pascal identity)} \\ &= C_{m+1}^{m+3} + \cdots + C_m^{m+k} && \text{(Pascal identity)} \\ &\vdots \\ &= C_{m+1}^{m+k} + C_m^{m+k} \\ &= C_{m+1}^{m+k+1} && \text{(Pascal identity)} \\ &= C_{m+1}^{n+1} && (m+k=n) \end{aligned}$$

(**i**) We have:

$$\begin{aligned} C_n^{2n} &= C_n^{n+n} \\ &= \sum_{p=0}^{n} C_p^n\, C_{n-p}^n && \text{(Vandermonde identity with } m=q=n) \end{aligned}$$

$$= \sum_{p=0}^{n} C_p^n \, C_p^n \qquad \text{(see part a)}$$
$$= \sum_{p=0}^{n} \left(C_p^n\right)^2$$

So, this is a special case of the Vandermonde identity with $m = q = n$.
(**j**) From Eq. 33 with $x = y = 1$ we have:

$$2^n = (1+1)^n = \sum_{k=0}^{n} C_k^n \; 1^{n-k} \; 1^k = \sum_{k=0}^{n} C_k^n$$

(**k**) Let us prove this by induction. For $m = 1$ we have:

$$\sum_{p=0}^{1} C_p^{n+p} = C_0^n + C_1^{n+1} = 1 + (n+1) = n + 1 + 1 = C_1^{n+1+1} = C_m^{n+m+1}$$

So, it is true for $m = 1$. Now, let assume that it is true for m and hence we must check if it is true for $m+1$, that is:

$$\begin{aligned}
\sum_{p=0}^{m+1} C_p^{n+p} &= \left(\sum_{p=0}^{m} C_p^{n+p}\right) + C_{m+1}^{n+m+1} \\
&= C_m^{n+m+1} + C_{m+1}^{n+m+1} && \text{(because it is presumably true for } m\text{)} \\
&= C_{m+1}^{n+m+2} && \text{(Pascal identity)} \\
&= C_{m+1}^{n+(m+1)+1}
\end{aligned}$$

So, it is true for $m+1$ and hence by mathematical induction it is true in general.
(**l**) Let have a collection of $2n$ balls: n of which are black and n of which are white. Now, the number of all subsets of size 2 balls in this $2n$-ball collection is given by C_2^{2n} (which is the left hand side of this identity). The number of all subsets of size 2 balls is also given by the sum of all 2-black combinations (which is C_2^n), all 2-white combinations (which is C_2^n), and all black-white combinations (which is n^2), and hence this sum is given by $2C_2^n + n^2$ (which is the right hand side of this identity). Accordingly, $C_2^{2n} = 2C_2^n + n^2$.
(**m**) Let have a collection of $3n$ balls: n of which are black, n of which are white, and n of which are red. Now, the number of all subsets of size 2 balls in this $3n$-ball collection is given by C_2^{3n} (which is the left hand side of this identity). The number of all subsets of size 2 balls is also given by the sum of all 2-black combinations (which is C_2^n), all 2-white combinations (which is C_2^n), all 2-red combinations (which is C_2^n), all black-white combinations (which is n^2), all black-red combinations (which is n^2), and all white-red combinations (which is n^2), and hence this sum is given by $3C_2^n + 3n^2$ (which is the right hand side of this identity). Accordingly, $C_2^{3n} = 3C_2^n + 3n^2$.
(**n**) We have (noting that n and m are integers with $n > 0$ and $m \geq 0$ and we use in the first line the definition of the binomial coefficient for negative powers):

$$\begin{aligned}
C_m^{-n} &= \frac{(-n)(-n-1)(-n-2)\cdots(-n-m+2)(-n-m+1)}{m!} && \text{(binomial coefficient)} \\
&= (-1)^m \, \frac{n(n+1)(n+2)\cdots(n+m-2)(n+m-1)}{m!} && \text{(taking out } m \text{ factors of } -1\text{)} \\
&= (-1)^m \, \frac{(n+m-1)(n+m-2)\cdots(n+2)(n+1)n}{m!} && \text{(reversing the order)} \\
&= (-1)^m \, \frac{(n+m-1)(n+m-2)\cdots(n+2)(n+1)n \times (n-1)!}{m!\,(n-1)!} && \left[\times (n-1)!\right]
\end{aligned}$$

$$= (-1)^m \frac{(n+m-1)!}{m!\,(n-1)!} \qquad \text{(Eq. 23)}$$
$$= (-1)^m\, C_m^{n+m-1} \qquad \text{(Eq. 25)}$$

(**o**) We have:

$$\begin{aligned}
\sum_{r=0}^{n-1} C_k^{k+r} &= C_k^k + C_k^{k+1} + C_k^{k+2} + \cdots + C_k^{k+n-1} \\
&= 1 + C_k^{k+1} + C_k^{k+2} + \cdots + C_k^{k+n-1} && (C_k^k = 1) \\
&= C_{k+1}^{k+1} + C_k^{k+1} + C_k^{k+2} + \cdots + C_k^{k+n-1} && (C_{k+1}^{k+1} = 1) \\
&= C_{k+1}^{k+2} + C_k^{k+2} + \cdots + C_k^{k+n-1} && \text{(Pascal identity)} \\
&= C_{k+1}^{k+3} + \cdots + C_k^{k+n-1} && \text{(Pascal identity)} \\
&\vdots \\
&= C_{k+1}^{k+n-1} + C_k^{k+n-1} \\
&= C_{k+1}^{k+n} && \text{(Pascal identity)}
\end{aligned}$$

PE: Do the following:
(**a**) Mark the Pascal triangle (see Figures 2 and 3) in a way that indicates the Pascal identity and demonstrates its validity.
(**b**) Mark the Pascal triangle (see Figures 2 and 3) in a way that indicates the hockey-stick identity and demonstrates its validity.
(**c**) Prove the Vandermonde identity algebraically.
(**d**) Prove the hockey-stick identity by induction.
(**e**) Can we conclude from parts (l) and (m) that $C_2^{kn} = k\,C_2^n + C_2^k\,n^2$ $(k = 2, 3, \ldots)$?

26. Show that the total number of combinations of n objects taken $1, 2, \ldots, n$ at a time is $(2^n - 1)$.
Answer: From part (j) of Problem 25 we have $C_0^n + C_1^n + \cdots + C_n^n = 2^n$. Now, if we note that $C_0^n = 1$ then we get $C_1^n + \cdots + C_n^n = 2^n - 1$, i.e. the total number of combinations of n objects taken $1, 2, \ldots, n$ at a time is $(2^n - 1)$ as required.
PE: Verify this result computationally (using a programming language or a spreadsheet) for the cases $n = 1, 2, \ldots, 10$.

27. We have 16 (numbered) balls and three bags where the capacity of these bags are 3, 5 and 8 balls. How many possibilities we have for filling the three bags with the 16 balls?
Answer: In a sense, this is a combination problem (since the balls in each bag are not ordered sets). Accordingly:
• We can fill the 3-ball bag in C_3^{16} ways.
• We can fill the 5-ball bag in C_5^{13} ways (noting that 3 balls are already put in the 3-ball bag and hence we have only 13 balls for filling the 5-ball bag).
• We can fill the 8-ball bag in C_8^8 ways (noting that only 8 balls are available for this filling).
Thus, by the fundamental principle of counting the number of possibilities for filling the three bags is:

$$C_3^{16} \times C_5^{13} \times C_8^8 = \frac{16!}{3!\,13!} \times \frac{13!}{5!\,8!} \times \frac{8!}{8!\,0!} = \frac{16!}{3!\,5!\,8!} = 720720$$

We may also argue differently as follows: we have 16! permutations for filling the bags (i.e. $16 \times 15 \times 14$ for filling the 3-ball bag, $13 \times 12 \times \cdots \times 9$ for filling the 5-ball bag, and $8 \times 7 \times \cdots \times 1$ for filling the 8-ball bag). However, because the order in each bag is irrelevant, then we divide $16 \times 15 \times 14$ by 3!, divide $13 \times 12 \times \cdots \times 9$ by 5!, and divide $8 \times 7 \times \cdots \times 1$ by 8!, and hence we get the same result. This argument can be presented formally as follows:

$$\frac{P_3^{16}}{3!} \times \frac{P_5^{13}}{5!} \times \frac{P_8^8}{8!} = \frac{16 \times 15 \times 14}{3!} \times \frac{13 \times 12 \times \cdots \times 9}{5!} \times \frac{8 \times 7 \times \cdots \times 1}{8!} = \frac{16!}{3!\,5!\,8!}$$

Also see Problem 30.

Note: it may be suspected that the order of filling the bags may affect the number of possibilities. However, this is not the case (partly because of what we got in part a of Problem 25, i.e. $C^n_{n-m} = C^n_m$). For example, if we reverse the order then we get:

$$C^{16}_8 \times C^8_5 \times C^3_3 = \frac{16!}{8!\,8!} \times \frac{8!}{5!\,3!} \times \frac{3!}{3!\,0!} = \frac{16!}{8!\,5!\,3!} = 720720$$

PE: Repeat the Problem for 29 cards to fill 4 bags of capacities: 9, 6, 4 and 10.

28. Show that the number of possibilities for filling k slots with n objects where the slots have capacities $n_1, n_2, \ldots, n_k$ (with $\sum_{i=1}^k n_i = n$) is:

$$C^n_{n_1,n_2,\ldots,n_k} = \frac{n!}{n_1!\,n_2!\ldots n_k!}$$

Answer: If we follow the method of Problem 27 then we may argue as follows:
• We can fill the n_1-object slot in $C^n_{n_1}$ ways.
• We can fill the n_2-object slot in $C^{n-n_1}_{n_2}$ ways (noting that n_1 objects are already put in the n_1-object slot and hence we have only $n - n_1$ objects for filling the n_2-object slot).
⋮
• We can fill the n_k-object slot in $C^{n-(n_1+\cdots+n_{k-1})}_{n_k} = C^{n_k}_{n_k}$ ways.
Thus, by the fundamental principle of counting the number of possibilities for filling the k slots is:

$$\begin{aligned} C^n_{n_1} \times C^{n-n_1}_{n_2} \times \cdots \times C^{n-(n_1+\cdots+n_{k-1})}_{n_k} &= \\ \frac{n!}{n_1!\cancel{(n-n_1)!}} \times \frac{\cancel{(n-n_1)!}}{n_2!\cancel{(n-n_1-n_2)!}} \times \cdots \times \frac{\cancel{\left[n-(n_1+\cdots+n_{k-1})\right]!}}{n_k!\,0!} &= \\ \frac{n!}{n_1!\,n_2!\ldots n_k!} &= C^n_{n_1,n_2,\ldots,n_k} \end{aligned}$$

We may also argue like our second argument in Problem 27 and we get the same result (also see Problem 30).

Note: $C^n_{n_1,n_2,\ldots,n_k}$ is a generalization of the binomial coefficient and hence it is called **multinomial coefficient** (noting that the *binomial* coefficient corresponds to *two* slots one of capacity m and one of capacity $n-m$). As a generalization to the binomial theorem (see Problem 20), we have the following **multinomial theorem**:

$$(x_1 + x_2 + \cdots + x_k)^n = \sum C^n_{n_1,n_2,\ldots,n_k}\, x_1^{n_1}\, x_2^{n_2} \ldots x_k^{n_k} \tag{34}$$

where the sum is taken for all $n_1 + n_2 + \cdots + n_k = n$ (for non-negative $n_1, n_2, \cdots, n_k$). Also see Problems 29 and 30.

PE: Do the following:
(**a**) Give a number of examples of specific "objects" and "slots" to which this type of "combination" applies (noting that "objects" and "slots" are generic labels that can apply to many things and concepts).
(**b**) Calculate $C^{20}_{2,3,4.5.6}$ and $C^{33}_{9,11,13}$.
(**c**) Expand $(x+y+z)^3$ algebraically and hence verify the multinomial theorem for this case.

29. Show that the binomial coefficient is a special case of the multinomial coefficient.

Answer: For $k = 2$, the multinomial coefficient is (noting that $n_1 + n_2 = n$):

$$C^n_{n_1,n_2} = \frac{n!}{n_1!\,n_2!} = \frac{n!}{n_1!\,(n-n_1)!} = C^n_{n_1} = C^n_{n_2}$$

PE: Noting that $C^n_{n_1} = C^n_{n_1,(n-n_1)} = C^n_{n_1,n_2}$, what can you conclude? Make a comment.

30. Show that $C^n_{n_1,n_2,...,n_k}$ represents the number of permutations of n objects such that: n_1 of which are identical (or indistinguishable), n_2 are identical, ..., and n_k are identical (with $\sum_{i=1}^{k} n_i = n$).
Answer: We may argue that being identical is like being stored in a slot where the objects in that slot are not distinguished from each other by order or anything else, and hence we can use the rationale and derivation of Problem 28. We may also argue (more appropriately for this purpose) that if the n objects are different then the number of their permutations (i.e. P^n_n) is $n!$. However, because n_1 of these n objects are identical then the permutations of the n_1 objects (whose number is $n_1!$) within the n-object permutations are identical and hence we should divide $n!$ by $n_1!$ to get the number of distinct permutations of the n objects. This similarly applies to the $n_2, \ldots, n_k$ identical objects within the n objects. Therefore, the number of distinct permutations of such n objects is:

$$\frac{n!}{n_1!\, n_2! \ldots n_k!} = C^n_{n_1,n_2,...,n_k}$$

Accordingly, $C^n_{n_1,n_2,...,n_k}$ represents the number of **permutations with repetitions** (i.e. repetition of n_1 objects, repetition of n_2 objects, ..., repetition of n_k objects).
It is important to note that if the n objects consist of some repetitive (or identical) objects and some non-repetitive objects then each non-repetitive object contribute a factor of 1! in the denominator and hence these factors can be ignored. For example, if we have n_1 identical objects and n_2 other identical objects as well as 3 distinct objects (and hence $n = n_1 + n_2 + 3$) then from the formula of $C^n_{n_1,n_2,...,n_k}$ we get:

$$C^n_{n_1,n_2,...,n_k} = C^n_{n_1,n_2,1,1,1} = \frac{n!}{n_1!\, n_2!\, 1!\, 1!\, 1!} = \frac{n!}{n_1!\, n_2!} = C^n_{n_1,n_2}$$

So, based on this understanding $C^n_{n_1,n_2,...,n_k}$ can apply even for $\sum_{i=1}^{k} n_i < n$. In fact, we can even consider $P^n_n = n!$ (for n distinct objects) as a special case of $C^n_{n_1,n_2,...,n_k}$ because when we have n distinct objects then $n_1 = n_2 = \cdots = n_k = 1$ and hence:

$$C^n_{n_1,n_2,...,n_k} = C^n_{1,1,...,1} = \frac{n!}{1!\, 1! \ldots 1!} = n! = P^n_n$$

Accordingly, we can consider $C^n_{n_1,n_2,...,n_k}$ as the more general symbol (and mathematical entity) for representing the numbers of combinations and permutations involving and considering n objects.
We also note that $C^n_{n_1,n_2,...,n_k}$ (which represents the number of permutations with repetitions regardless of the condition $n_1 + n_2 + \ldots + n_k = n$, i.e. possibly $\sum_{i=1}^{k} n_i < n$) reduces to the multinomial coefficient (see Problem 28) when $n_1 + n_2 + \ldots + n_k = n$.
Note: the reader should be aware that the term "permutations with repetitions" (and its alike) may be used differently where "repetition" means the repetition of the individual objects in selections of multiple objects (rather than repetition in the n objects of the set from which the selections are made), and hence the number of "permutations with repetitions" (of n objects taken k at a time) in this sense is $P_{n,k} = n^k$ (see Eq. 30) as opposite to the number of "permutations with no repetitions" which is given by $P^n_k = n!/(n-k)!$ according to Eq. 29 (and as compared to the number of "permutations with repetitions" in the above sense which is given by $C^n_{n_1,n_2,...,n_k}$). So in brief, the number of "permutations with repetitions in the objects of the *selection set* which is made of n objects" is $C^n_{n_1,n_2,...,n_k}$, and the number of "permutations with repetitions in the objects of the *selected set* which is made of k objects" is $P_{n,k}$.[39] In fact, it is better to adopt two different terms for these two types of "permutations with repetitions" (e.g. one is called "permutations with repetitions" and the other is called "permutations with duplication"). However, we do not want to go against the literature in this issue. Also see Problem 5.
PE: Calculate the following [noting that if $n > (n_1 + \cdots + n_k)$ then the remaining $n - (n_1 + \cdots + n_k)$

[39] It should also be noted that in $C^n_{n_1,n_2,...,n_k}$ the entire set of n objects are taken in the permutation (and hence we have "n-permutations" out of n objects), while in $P_{n,k}$ a set of size k objects (out of a set of size n objects) are taken in the permutation (and hence we have "k-permutations" out of n objects).

are distinct or non-repetitive]:

(**a**) $C^9_{4,5}$. (**b**) $C^9_{2,5}$. (**c**) $C^{13}_{3,6}$. (**d**) $C^{13}_{3,4,6}$.

31. Find the number of distinct strings that can be formed from the letters of the word (**a**) "attachment" (**b**) "subsequently" (**c**) "incomprehensibility".
Answer: Referring to Problem 30:
(**a**) The word "attachment" contains 10 letters where "a" is repeated 2 times and "t" is repeated 3 times. Therefore, the number of distinct strings that can be formed from the letters of "attachment" is $\frac{10!}{2!\,3!} = 302400$.
(**b**) Following the method of part (a) we have $\frac{12!}{2!\,2!\,2!} = 59875200$.
(**c**) Following the method of part (a) we have $\frac{19!}{4!\,2!\,2!} = 1267136462592000$.
PE: Repeat the Problem for the words (**a**) "Sussex" (**b**) "explicit" (**c**) "indistinguishable".

32. Let have 7 drawers and 5 shirts and we want to store these shirts in (5 of) these drawers restricted by the condition that no more than 1 shirt can be stored in any drawer. How many ways we have for storing these shirts?
Answer: We have C^7_5ways for selecting 5 drawers (out of 7) to be used for storage, and we have 5! ways for distributing the 5 shirts on the selected 5 drawers (i.e. the first selected drawer can be used for storing any one of the 5 shirts, the second selected drawer can be used for storing any one of the remaining 4 shirts, and so on). Hence we have:

$$C^7_5 \times 5! = \frac{7!}{5!(7-5)!} \times 5! = \frac{7!}{(7-5)!} = P^7_5 = 2520$$

ways for storing these shirts.
We may also argue (differently) that we have 7 drawers to be assigned to 5 shirts where we have 7 ways for assigning a drawer to the first shirt, 6 ways for assigning a drawer to the second shirt, ..., and 3 ways for assigning a drawer to the fifth shirt and hence we have:

$$7 \times 6 \times 5 \times 4 \times 3 = \frac{7!}{(7-5)!} = P^7_5 = 2520$$

ways for storing these shirts.. Also see Problem 37 of § 3.2.
PE: Repeat this Problem assuming we have 6 cars to be stored in 9 garages where each garage can accommodate no more than 1 car.

33. Let have 7 shirts and we want to store 5 of them in a drawer and the remaining 2 in a second drawer. In how many ways this can be done?
Answer: This is obviously a combination problem because we want to select 5 (non-ordered) shirts out of 7 shirts for storage in the first drawer (noting that the remaining 2 shirts will inevitably go to the second drawer). Hence, we have $C^7_5 = 21$ ways. We may also argue (differently) that we want to select 2 (non-ordered) shirts out of 7 shirts for storage in the second drawer (noting that the remaining 5 shirts will inevitably go to the first drawer), and hence, we have $C^7_2 = 21$ ways as before (noting that $C^7_2 = C^7_5$; see part a of Problem 25).
PE: Solve this Problem as a permutation with repetition problem (see Problem 30).

34. In statistical mechanics we have three main types of statistics (or distribution):
(**a**) Maxwell-Boltzmann statistics for (classical) distinguishable particles.
(**b**) Fermi-Dirac statistics for (quantum) indistinguishable particles that cannot occupy the same state (e.g. electrons).
(**c**) Bose-Einstein statistics for (quantum) indistinguishable particles that can occupy the same state (e.g. photons).
If we have n particles and k available states (where these n particles are supposed to occupy some or all of these k states), find the number of possibilities of occupancy for each one of these statistics.
Answer:
(**a**) For **Maxwell-Boltzmann** statistics each one of the n particles can occupy any one of the k states

and hence we have n of k possibilities, i.e. k^n possibilities.[40] This is because the assignment of each particle to any state is independent of the assignment of the other particles. For example, if we have 5 particles and 3 states then the first particle can occupy state 1 or 2 or 3 and .... and the fifth particle can occupy state 1 or 2 or 3 and hence we have 3^5 possibilities for occupancy.

(**b**) For **Fermi-Dirac** statistics we have C_n^k possibilities for occupancy (with the restriction that $k \geq n$ because each available state cannot be occupied by more than one particle and hence the available states must be at least as many as the number of particles). This is because we have C_n^k ways for selecting n states (out of k states) to be assigned to the n particles noting that the particles are indistinguishable.[41]

(**c**) For **Bose-Einstein** statistics we have C_n^{n+k-1} possibilities for occupancy. To explain and justify this let us use a simple example where we have 5 balls to be stored in this way (i.e. by the Bose-Einstein statistics) on a shelf that is divided to 8 compartments by 7 (i.e. $8-1$) interior (movable) separators (and contained between two walls). There are many distinguishable configurations (or arrangements) for the occupancy of these balls in the 8 compartments. To demonstrate this, we present in Figure 4 four of these arrangements.

As we see, the balls (on one hand) and the separators (on the other hand) are not distinguished in this type of storage (because all we are interested in is the number of balls in each particular compartment), and hence each one of these arrangements is distinguished by the positions of the separators between the balls (or similarly by the positions of the balls between the separators). In other words, we are interested in the combination of the $8-1=7$ positions of separators in $5+7$ positions of "balls and separators" (i.e. selecting 7 positions out of 12 positions).[42] So in brief, we are essentially selecting 7 distinct positions (i.e. the positions of the separators) out of the $5+7$ distinct positions (i.e. the positions of the "balls and separators") noting that the selection of these positions is not subject to order and hence it is like selecting 7 distinct balls (at once) out of 12 distinct balls. Accordingly, the number of possible arrangements is $C_{8-1}^{5+8-1} = C_5^{5+8-1}$ (see part a of Problem 25).

Now, if the n particles in our Problem correspond to the 5 balls in this example, and the k states in our Problem correspond to the 8 compartments in this example (noting that the separators, whose number is the number of compartments minus 1, correspond to $k-1$ states), then we can conclude from this example (by generalizing its pattern) that the number of possible configurations for occupancy in the Bose-Einstein statistics is $C_{k-1}^{n+k-1} = C_n^{n+k-1}$. Also see Problem 4 of § 7.2.

Note: the above three types of distribution represent the main (i.e. famous and physically significant) distributions but they obviously do not exhaust all the possibilities. For example, we can imagine a statistics of distinguishable particles (as in Maxwell-Boltzmann statistics) that cannot occupy the same state (as in Fermi-Dirac statistics).[43] In fact, there are other types of statistics (some can be found in the literature and some can be proposed and synthesized).

PE: Give several examples from daily life for each one of the above distributions.[44] Also consider and

---

[40] In other words, each one of the n particles can assume any one of the k states and hence by the fundamental principle of counting the number of possibilities is:

$$\overbrace{k \times k \times \cdots \times k}^{n \text{ factors}} = k^n$$

[41] In fact, the situation here is similar in part to the situation in Problem 32 (noting that the particles here, unlike the shirts of Problem 32, are indistinguishable and hence the 5! factor of Problem 32 does not arise here).

[42] We may equally say: we are interested in the combination of the 5 positions of balls in $5+7$ positions of "balls and separators" (i.e. selecting 5 positions out of 12 positions). We may also consider this as a permutation with repetition problem (see Problem 30) where we have 12 objects 5 of which are identical and 7 of which are identical and hence the number of their distinguishable permutations is $C_{5,7}^{12} = \frac{12!}{5!\,7!}$.

[43] The number of possible configurations for occupancy (of n particles and k available states where $k \geq n$) in such statistics is P_n^k. This is because we can select any one of the k available states for the occupancy of the 1^{st} particle, any one of the remaining $k-1$ states for the occupancy of the 2^{nd} particle, ... , and any one of the remaining $k-n+1$ states for the occupancy of the n^{th} particle, and hence by the fundamental principle of counting we have $k \times (k-1) \times \cdots \times (k-n+1) = P_n^k$ possible configurations.

[44] For example, storing 10 distinct balls (e.g. by having different colors or size or weight) in 3 large boxes (where each box can accommodate all the 10 balls) is similar to the Maxwell-Boltzmann statistics, storing 6 identical balls in 9 small

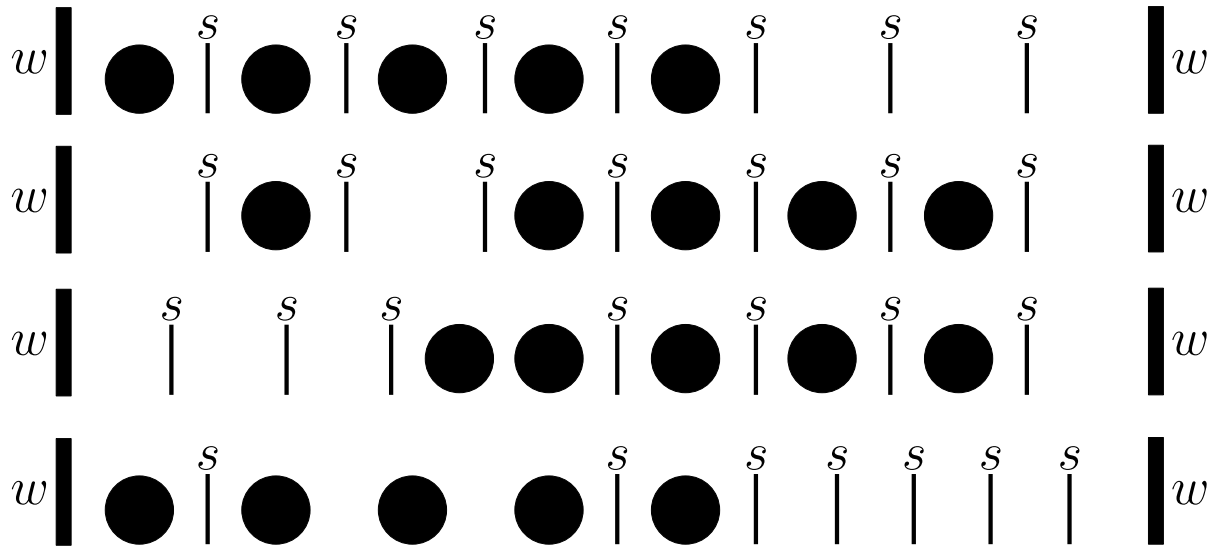


Figure 4: Four specific arrangements of the possible $C_{8-1}^{5+8-1} = C_5^{5+8-1}$ configurations of the occupancy of 5 balls in 8 compartments according to the Bose-Einstein statistics. We note that the black circles represent balls, w stands for wall and s for separator. See Problem 34 of § 2.2.

investigate other types of distribution, e.g. distributions similar to the Maxwell-Boltzmann statistics or the Bose-Einstein statistics but (some or all of) the available states have specific capacities.

35. Referring to Problem 34, let have 3 particles and 4 states. Find all the possibilities of occupancy according to:

(**a**) Maxwell-Boltzmann statistics. (**b**) Fermi-Dirac statistics. (**c**) Bose-Einstein statistics.

Answer: Let us use 3 English letters (i.e. different or identical depending on the case) to represent the 3 particles and use 3 separators to represent the 4 states (ignoring the walls on the two sides).

(**a**) In the Maxwell-Boltzmann statistics the particles are distinguishable and therefore we use 3 different letters (a, b and c) to represent the particles. Moreover, there is no restriction on the number of particles that can occupy any one of the states. In this statistics we have $k^n = 4^3 = 64$ possibilities which are:

abc| | |	**|abc| |**	**| |abc|**	**| | |abc**	**a|bc| |**	**a| |bc|**	**a| | |bc**	**b|ac| |**
b| |ac|	**b| | |ac**	**c|ab| |**	**c| |ab|**	**c| | |ab**	**bc|a| |**	**|a|bc|**	**|a| |bc**
ac|b| |	**|b|ac|**	**|b| |ac**	**ab|c| |**	**|c|ab|**	**|c| |ab**	**bc| |a|**	**|bc|a|**
| |a|bc	**ac| |b|**	**|ac|b|**	**| |b|ac**	**ab| |c|**	**|ab|c|**	**| |c|ab**	**bc| | |a**
|bc| |a	**| |bc|a**	**ac| | |b**	**|ac| |b**	**| |ac|b**	**ab| | |c**	**|ab| |c**	**| |ab|c**
a|b|c|	**a|c|b|**	**a|b| |c**	**a|c| |b**	**a| |b|c**	**a| |c|b**	**b|a|c|**	**b|c|a|**
b|a| |c	**b|c| |a**	**b| |a|c**	**b| |c|a**	**c|a|b|**	**c|b|a|**	**c|a| |b**	**c|b| |a**
c| |a|b	**c| |b|a**	**|a|b|c**	**|a|c|b**	**|b|a|c**	**|b|c|a**	**|c|a|b**	**|c|b|a**

(**b**) In the Fermi-Dirac statistics the particles are indistinguishable and therefore we use a single letter (i.e. a) to represent the particles. Moreover, we have the restriction that no more than one particle can occupy any state. In this statistics we have $C_n^k = C_3^4 = 4$ possibilities which are:

a|a|a|	**a|a| |a**	**a| |a|a**	**|a|a|a**

boxes (where each box can accommodate only one ball) is similar to the Fermi-Dirac statistics, and storing 10 identical balls in 3 large boxes (where each box can accommodate all the 10 balls) is similar to the Bose-Einstein statistics.

(**c**) In the Bose-Einstein statistics the particles are indistinguishable and therefore we use a single letter (i.e. a) to represent the particles. Moreover, there is no restriction on the number of particles that can occupy any one of the states. In this statistics we have $C_n^{n+k-1} = C_3^6 = 20$ possibilities which are:

aaa| | |	**|aaa| |**	**| |aaa|**	**| | |aaa**	**a|aa| |**
a| |aa|	**a| | |aa**	**aa|a| |**	**|a|aa|**	**|a| |aa**
aa| |a|	**|aa|a|**	**| |a|aa**	**aa| | |a**	**|aa| |a**
| |aa|a	**a|a|a|**	**a|a| |a**	**a| |a|a**	**|a|a|a**

PE: Repeat this Problem for 3 particles and 3 states.

36. Repeat Problem 35 assuming a statistics of distinguishable particles that cannot occupy the same state.

Answer: In this statistics the particles are distinguishable and therefore we use 3 different letters (a, b and c) to represent the particles. Moreover, we have the restriction that no more than one particle can occupy any state. As explained earlier (see footnote [43]), we have $P_n^k = P_3^4 = 24$ possibilities which are:

a|b|c|	**a|c|b|**	**a|b| |c**	**a|c| |b**	**a| |b|c**	**a| |c|b**	**b|a|c|**	**b|c|a|**
b|a| |c	**b|c| |a**	**b| |a|c**	**b| |c|a**	**c|a|b|**	**c|b|a|**	**c|a| |b**	**c|b| |a**
c| |a|b	**c| |b|a**	**|a|b|c**	**|a|c|b**	**|b|a|c**	**|b|c|a**	**|c|a|b**	**|c|b|a**

Note: some readers may have noticed that the answer of this Problem is identical to the last three rows of the answer of part (a) of Problem 35. This should be no surprise since the answer of the present Problem can be obtained from the answer of part (a) of Problem 35 by excluding all the possibilities of part (a) of Problem 35 that have multiple occupancy to any single state.

PE: Repeat Problem 35 assuming a statistics of indistinguishable particles where the capacity of state 1 is one particle (i.e. it cannot accommodate more than one particle), the capacity of state 2 is two particles, the capacity of state 3 is three particles, and the capacity of state 4 is four particles.

37. A digital byte is an ordered arrangement of 8 bits where each bit can be either 0 or 1. How many different bytes that contain exactly four 1's we have?

Answer: If we consider the positions of the bits within the byte as states and consider the four 1's as indistinguishable particles (noting that no one of these 1's can occupy more than one position) then this is like a Fermi-Dirac statistics and hence we have $C_4^8 = 70$ different bytes containing exactly four 1's.

We may also argue (differently and more simply) that we can assign the first 1 to any of the 8 positions, the second 1 to any of the remaining 7 positions, the third 1 to any of the remaining 6 positions, and the fourth 1 to any of the remaining 5 positions, and hence we have $8 \times 7 \times 6 \times 5 = P_4^8$ possibilities. However, because the 1's are indistinguishable we should divide this by 4! (i.e. the number of their permutations) and hence we should have $P_4^8/4! = C_4^8 = 70$. In fact, this can be understood (more simply) as choosing 4 positions (out of the 8 available positions) for the assignment of the "1" bits.

PE: Repeat the Problem for the different bytes that contain exactly two 1's and list all these bytes.

38. Referring to Problem 1 of § 1.5.4, find the number of possibilities for the random selection of 5 balls out of 10 in each one of the 4 cases considered in that Problem.

Answer:

(**a**) For sampling with order and without replacement we have $P_5^{10} = 30240$ possibilities (because the 1^{st} drawn ball can be any one of the 10 balls, the 2^{nd} drawn ball can be any one of the remaining 9 balls, ..., and the 5^{th} drawn ball can be any one of the remaining 6 balls and hence by the fundamental principle of counting we have $10 \times 9 \times 8 \times 7 \times 6 = P_5^{10}$ possibilities). In fact, this is an application of Eq. 29.

(**b**) For sampling with order and with replacement we have 10^5 possibilities (because each one of the 5 drawn balls can be any one of the 10 balls and hence by the fundamental principle of counting we have $10 \times 10 \times 10 \times 10 \times 10 = 10^5$ possibilities). In fact, this is an application of Eq. 30.

(**c**) For sampling without order and without replacement we have $C_5^{10} = 252$ possibilities (because this

is the same as the number of distinct sets of size 5 in a set of size 10). In fact, this is an application of Eq. 31.
(**d**) For sampling without order and with replacement we have $C_{10,5} = 2002$ possibilities. In fact, this is an application of Eq. 32.
PE: Considering the answer of this Problem, can we make the following generalization:

$$N_{nn} \leq N_{nr} \leq N_{on} \leq N_{or}$$

where N_{nn} is the number of possibilities without order and without replacement, N_{nr} is the number of possibilities without order and with replacement, N_{on} is the number of possibilities with order and without replacement, and N_{or} is the number of possibilities with order and with replacement? Justify your answer.

39. Referring to Problem 2 of § 1.5.4, find the number of possibilities for the random selection of 5 balls out of 10 in each one of the 2 cases considered in that Problem (i.e. with and without replacement).
Answer: As the balls are indistinguishable we have only one possibility in each one of the 2 cases.
PE: Find the number of possibilities for the random selection of 5 balls out of 10 identical balls if 5 of the 10 balls are black and the other 5 are white. Assume the selection is (**a**) with replacement and (**b**) without replacement. Also consider order.
40. We have 40 passengers and 5 buses each of 60-passenger capacity. How many ways these passengers can be distributed on these buses if:
(**a**) The assignment of passengers to buses is considered, i.e. it is important which passenger travels on which bus.
(**b**) The assignment of passengers to buses is not considered, i.e. the only important thing is the number of passengers traveling on which of these buses.
Answer:
(**a**) We have 40 distinguishable passengers to be assigned to 5 distinct buses. Referring to part (a) of Problem 34, this is like a Maxwell-Boltzmann statistics with $n = 40$ and $k = 5$ and hence we have $k^n = 5^{40}$ ways.
(**b**) We have 40 indistinguishable passengers to be assigned to 5 distinct buses. Referring to part (c) of Problem 34, this is like a Bose-Einstein statistics with $n = 40$ and $k = 5$ and hence we have $C_n^{n+k-1} = C_{40}^{44} = 135751$ ways.
PE: Repeat the Problem assuming we have 40 passengers and 10 taxis each of 4-passenger capacity.
41. In **circular permutation** n distinct objects ($n \geq 3$) are arranged in a circle (i.e. the order of the objects in the circle is considered but with no consideration of start/end and hence nothing presumably changes if we rotate the circle due to the circular symmetry). Find the number of circular permutations of n objects arranged in a circle.
Answer: If the permutation is straight then we have $n!$ permutations. So, let us bend this straight permutation to make a circle. Accordingly, we still have $n!$ permutations but with the condition that if we rotate any one of these circular permutations by an angle of size $2\pi/n$ we get the same permutation. Now, in a circle we have n angles of size $2\pi/n$. This means that we need to divide $n!$ by n to get the number of distinct circular permutations. Hence, the number of circular permutations of n objects is $(n-1)!$.
We may also argue (in another way) that we can assign one object (say the first object) to any one of the n positions in the circle (noting that nothing changes if we rotate the circle and hence the assignment to any one of the n positions is indistinguishable from the assignment to any other position which means that we essentially have only one way for this assignment). The remaining $(n-1)$ objects can then be arranged in $(n-1)!$ distinct ways (noting their positions relative to the first object considering a specific sense of order, e.g. clockwise). Hence, by the fundamental principle of counting we must have $1 \times (n-1)! = (n-1)!$ circular permutations.
PE: Give some examples of circular permutation in real life. What is the number of circular permutations of 6 objects? What is the number if the 6 objects are identical?
42. Give all the circular permutations of (**a**) the letters {a, b, c} (**b**) the letters {a, b, c, d} (**c**) the letters {a, b, c, d, e}.

Answer:
(**a**) We have $(3-1)!=2$ circular permutations which are: abc, acb.
(**b**) We have $(4-1)!=6$ circular permutations which are: abcd, abdc, acbd, acdb, adbc, adcb.
(**c**) We have $(5-1)!=24$ circular permutations which are:

abcde abced abdce abdec abecd abedc acbde acbed acdbe acdeb acebd acedb
adbce adbec adcbe adceb adebc adecb aebcd aebdc aecbd aecdb aedbc aedcb

PE: Plot the above circular permutations on circles.[45]

43. How many 5-letter strings can be made from the letters {s, d, i, e, f, j, o} if no vowel is allowed to be at the beginning or the end and with no repetition?
Answer: We have 4 consonants and 3 vowels. Now, the first letter can be any one of the 4 consonants and the last letter can be any one of the remaining 3 consonants while the 3 middle letters can be filled with the 3-letter permutations of the remaining 5 letters. Hence, the number of 5-letter strings is:

$$4\times P_3^5\times 3=4\times\frac{5!}{(5-3)!}\times 3=720$$

PE: Repeat the Problem assuming repetition is allowed and the last letter can be a vowel.

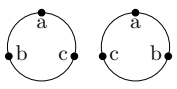
44. We have 7 numbered balls (4 black and 3 white) to be arranged in a row. In how many ways this can be done if:
(**a**) The balls are grouped by color. (**b**) Only the black balls are grouped.
Answer:
(**a**) In this case we have 2 main possibilities: black first and white first. Now, the blacks can be arranged in 4! ways and the whites can be arranged in 3! ways. Therefore, we have $2\times 4!\times 3!=288$ ways.
(**b**) In this case we have 4 main possibilities: black first, black second, black third, and black last. Again, the blacks can be arranged in 4! ways and the whites can be arranged in 3! ways. Therefore, we have $4\times 4!\times 3!=576$ ways.
PE: Repeat the Problem (i.e. both parts) assuming:
(**a**) The balls of each color should also be ordered according to their numbers increasingly.
(**b**) The balls are not numbered (and hence they are indistinguishable except by color).

45. We have 9 numbered balls (4 black, 3 white and 2 red). In how many ways they can be arranged:
(**a**) In a row grouped by color. (**b**) In a circle grouped by color.
Answer:
(**a**) The 3 colors can be arranged in a row in 3! ways, the 4 blacks in 4! ways, the 3 whites in 3! ways and the 2 reds in 2! ways. Hence, there are $3!\times 4!\times 3!\times 2!=1728$ ways.
(**b**) The 3 colors can be arranged in a circle in 2! ways (see Problem 41), the 4 blacks in 4! ways, the 3 whites in 3! ways and the 2 reds in 2! ways. Hence, there are $2!\times 4!\times 3!\times 2!=576$ ways.
PE: Repeat the Problem assuming the balls are not numbered and no grouping by color is imposed.

46. Define and explain C_k^{-n} where n is a positive integer and $k=0,1,\ldots,\infty$.
Answer: C_k^{-n} is the "binomial coefficient" appearing in the binomial theorem expansion for negative integer powers which is given by:

$$(x+y)^{-n}=\sum_{k=0}^{\infty}C_k^{-n}\,x^k\,y^{-n-k}=\sum_{k=0}^{\infty}(-1)^k\,C_k^{n+k-1}\,x^k\,y^{-n-k}\qquad(|x/y|<1)\qquad(35)$$

[45] For example, the circular permutations abc and acb can be plotted as:



For example, the well-known expansion of $(x+1)^{-n}$ for $|x|<1$ is given by:

$$(x+1)^{-n} = 1 - nx + \frac{n(n+1)}{2}x^2 - \frac{n(n+1)(n+2)}{6}x^3 + \frac{n(n+1)(n+2)(n+3)}{24}x^4 - \cdots \qquad (36)$$

If we compare the coefficients in Eqs. 35 and 36 we see:

$C_0^{n+0-1} = 1$ $C_1^{n+1-1} = n$ $C_2^{n+2-1} = \frac{n(n+1)}{2}$

$C_3^{n+3-1} = \frac{n(n+1)(n+2)}{6}$ $C_4^{n+4-1} = \frac{n(n+1)(n+2)(n+3)}{24}$

And this general pattern continues.

PE: By implementing the binomial theorem expansion for negative integer powers (according to the above-given form) in a spreadsheet (or in a computer code or in a script of a mathematical software package) show that the expansions of the following binomial expressions converge to the indicated values (where convergence is symbolized by $\to$).

(**a**) $(0.3+1)^{-4} \to 1.3^{-4}$. (**b**) $(-0.15+1)^{-10} \to 0.85^{-10}$. (**c**) $(3.1+4.6)^{-7} \to 7.7^{-7}$.

(**d**) $(-1.4+2.1)^{-11} \to 0.7^{-11}$. (**e**) $(4.1-6.8)^{-6} \to (-2.7)^{-6}$. (**f**) $(-5.9-8.4)^{-9} \to (-14.3)^{-9}$.

2.3 Dealing with Very Large and Very Small Numbers

It is common in the probability theory (and related subjects like counting methods and combinatorics in general) to face problems involving very large or/and very small numbers which only few calculators (or compilers or mathematical software packages) can manage.[46] For example, the calculation of the factorial of 2369 or the value of 0.5^{3475} cannot be done by ordinary means like calculating the factorial of 15 or the value of 0.5^8. In fact, in some cases even if a calculator that can manage this is available it could be difficult to make use of it due to practical and procedural issues. So, being able to deal with this sort of extreme calculations independently and by our simple means (and hence be less dependent on external help such as specialized software packages which may not be available or not usable) is a big advantage and can even be a necessity in some situations.

It is worth noting that this sort of problems and difficulties have no easy solution or fix in general, and hence we need to invent and tailor innovative methods that can deal with these problems case by case. In the following subsections we will outline some elementary methods that can be used to overcome (or at least mitigate the severity of) these difficulties. However, we should insist that these problems and difficulties could require (in some cases) exceptional and non-elementary methods if they are solvable at all.

2.3.1 Use of Calculational Tricks

There are many calculational tricks (which are usually case specific) that can be used for overcoming the difficulties of extreme calculations. A particularly useful trick in the calculations of probability is what we call "splitting method" which can be used when we have calculations involving products or/and quotients of too many factors some of which are small and some are large. In such cases we split these factors in small groups each of which involves some small factors and some large factors and the results obtained from calculating these groups are then processed easily (e.g. by multiplying or/and dividing them) to obtain the final result. For example, it is common in the probability theory to face probabilities given by

[46] We should note that dealing with very large and very small numbers (which is the subject and title of this section) should be seen as a typical (and more common) example for calculational difficulties encountered in the field of probability (and related subjects). So, the proposed methods in the following subsections should be more general in their benefits for easing and overcoming the difficulties encountered in this field. Moreover, other methods for easing and overcoming these difficulties could be proposed in this regard. In fact, it may be more appropriate to title this section with something like "Dealing with Calculational Difficulties". However, we preferred to title it according to the most common calculational difficulties encountered in this field (considering our limited needs in this book and avoiding the obligation to go beyond our scope if we choose a more general title).

this type of formula:

$$C_{469}^{1321}0.5^{1321}$$

As we see, neither C_{469}^{1321} nor 0.5^{1321} is easy to calculate because the first is too big (and hence it is seen as infinity by the ordinary calculators) while the second is too small (and hence it is rounded to zero by the ordinary calculators) and this makes it impossible to calculate this formula as a product of C_{469}^{1321} and 0.5^{1321}. However, we can put this formula in the following form which can be easily calculated (e.g. by using a spreadsheet):

$$\begin{aligned}
C_{469}^{1321}0.5^{1321} &= \frac{1321!}{469!(1321-469)!}\times 0.5^{1321}\\
&= \frac{1321\times 1320\times\cdots\times 853}{469!}\times 0.5^{1321}\\
&= \frac{1321\times 1320\times\cdots\times 853}{469!}\times\left(0.5^3\right)^{440}\times 0.5\\
&= \frac{1321\times 1320\times\cdots\times 853}{469!}\times 0.125^{440}\times 0.5\\
&= \frac{1321\times 0.125}{469}\times\frac{1320\times 0.125}{468}\times\cdots\times\frac{883\times 0.125}{31}\times\frac{882\times 0.125}{30}\times\\
&\quad \frac{881}{29}\times\frac{880}{28}\times\cdots\times\frac{853}{1}\times 0.5\\
&\simeq 0.35208\times 0.35256\times\cdots\times 3.56048\times 3.675\times 30.37931\times 31.42857\times\cdots\times 853\times 0.5\\
&\simeq 7.9\times 10^{-27}
\end{aligned}$$

Problems

1. Find the probability of a given event which is given by $P=C_{1407}^{2375}\,0.6^{1407}\,0.4^{968}$.
   Answer: Using the splitting method we have:

$$\begin{aligned}
P &= C_{1407}^{2375}\,0.6^{1407}\,0.4^{968}\\
&= \frac{2375!}{1407!(2375-1407)!}\,0.6^{1407}\,0.4^{968}\\
&= \frac{2375\times 2374\times\cdots\times 969}{1407!}\,0.6^{1407}\,0.4^{968}\\
&= \frac{2375\times 2374\times\cdots\times 969}{1407!}\,0.6^{968}\,0.6^{439}\,0.4^{968}\\
&= \frac{2375\times 2374\times\cdots\times 969}{1407!}\,0.24^{968}\,0.6^{439}\\
&= \frac{2375\times 0.24}{1407}\times\frac{2374\times 0.24}{1406}\times\cdots\times\frac{1409\times 0.24}{441}\times\frac{1408\times 0.24}{440}\times\\
&\quad \frac{1407\times 0.6}{439}\times\frac{1406\times 0.6}{438}\times\cdots\times\frac{970\times 0.6}{2}\times\frac{969\times 0.6}{1}\\
&\simeq 0.40512\times 0.40523\times\cdots\times 0.76680\times 0.768\times 1.92301\times 1.92603\times\cdots\times 291\times 581.4\\
&\simeq 0.012544
\end{aligned}$$

   PE: Find the probability of a given event which is given by $P=C_{1407}^{2375}\,0.61^{1407}\,0.39^{968}$.
2. Calculate the following factorials:
   (**a**) 235!. (**b**) 569!. (**c**) 873!. (**d**) 1382!. (**e**) 1948!. (**f**) 2634!.
   Answer: For only very few practical purposes the exact integer value is required. This is because there are too many digits in the number which makes it almost useless in its exact integer format. So, in the overwhelming majority of applications (especially in science) what is required is an approximate value that gives the right order of magnitude of the factorial in a fractional scientific format,[47] and

[47] In fact, this format is what we usually need in the calculations of probabilities.

this is what we will do here. Using the laws of logarithms, we have:

$$\log(n!) = \log(1) + \log(2) + \cdots \log(n-1) + \log(n) = \sum_{i=1}^{n} \log(i)$$

This sum can be easily calculated with a humble calculator (e.g. a spreadsheet) and the result is obtained straightforwardly. On using logarithm to the base 10 we get:

(**a**) $\sum_{i=1}^{235} \log(i) \simeq 456.726522256951.$ (**b**) $\sum_{i=1}^{569} \log(i) \simeq 1322.32202904815.$

(**c**) $\sum_{i=1}^{873} \log(i) \simeq 2190.23599056559.$ (**d**) $\sum_{i=1}^{1382} \log(i) \simeq 3741.95651163172.$

(**e**) $\sum_{i=1}^{1948} \log(i) \simeq 5564.15753179453.$ (**f**) $\sum_{i=1}^{2634} \log(i) \simeq 7868.07968605461.$

Therefore:

(**a**) $235! = 10^{\log(235!)} \simeq 10^{456.726522256951} = 10^{0.72652225695} \times 10^{456} \simeq 5.32748526199126 \times 10^{456}$

(**b**) $569! = 10^{\log(569!)} \simeq 10^{1322.32202904815} = 10^{0.32202904815} \times 10^{1322} \simeq 2.09908027765868 \times 10^{1322}$

(**c**) $873! = 10^{\log(873!)} \simeq 10^{2190.23599056559} = 10^{0.23599056559} \times 10^{2190} \simeq 1.72183117031237 \times 10^{2190}$

(**d**) $1382! = 10^{\log(1382!)} \simeq 10^{3741.95651163172} = 10^{0.95651163172} \times 10^{3741} \simeq 9.04714668414713 \times 10^{3741}$

(**e**) $1948! = 10^{\log(1948!)} \simeq 10^{5564.15753179453} = 10^{0.15753179453} \times 10^{5564} \simeq 1.43724826990598 \times 10^{5564}$

(**f**) $2634! = 10^{\log(2634!)} \simeq 10^{7868.07968605461} = 10^{0.07968605461} \times 10^{7868} \simeq 1.20139564857308 \times 10^{7868}$

PE: Calculate the following factorials (approximately):

(**a**) 297!. (**b**) 473!. (**c**) 701!. (**d**) 1391!. (**e**) 1883!. (**f**) 2564!.

3. Calculate the following permutations:[48]

(**a**) P_{52}^{236}. (**b**) P_{159}^{668}. (**c**) P_{982}^{1420}. (**d**) P_{1204}^{1875}.

Answer: Again, for only very few practical purposes the exact integer value is required. So, we repeat our argument and approach in Problem 2 and hence calculate an approximate value that gives the order of magnitude in a scientific format. Using the laws of logarithms, we have:

$$\log\left(P_k^n\right) = \log\left[\frac{n!}{(n-k)!}\right] = \log\Big[n \times (n-1) \times \cdots \times (n-k+1)\Big] = \sum_{i=n-k+1}^{n} \log(i)$$

This sum can be easily calculated (by a spreadsheet or a computer code for instance). On using logarithm to the base 10 we get:

(**a**) $P_{52}^{236} = 10^{\log\left(P_{52}^{236}\right)} \simeq 10^{120.751443959533} = 10^{0.751443959533} \times 10^{120} \simeq 5.64214131616555 \times 10^{120}$

(**b**) $P_{159}^{668} = 10^{\log\left(P_{159}^{668}\right)} \simeq 10^{440.237515517598} = 10^{0.237515517598} \times 10^{440} \simeq 1.72788771785137 \times 10^{440}$

(**c**) $P_{982}^{1420} = 10^{\log\left(P_{982}^{1420}\right)} \simeq 10^{2893.061957369540} = 10^{0.061957369540} \times 10^{2893} \simeq 1.15334004011238 \times 10^{2893}$

(**d**) $P_{1204}^{1875} = 10^{\log\left(P_{1204}^{1875}\right)} \simeq 10^{3717.479122616300} = 10^{0.479122616300} \times 10^{3717} \simeq 3.01385681969706 \times 10^{3717}$

PE: Calculate the following permutations (approximately):

(**a**) P_{66}^{391}. (**b**) P_{503}^{782}. (**c**) P_{701}^{1647}. (**d**) P_{1305}^{1958}.

4. Calculate the following binomial coefficients:

(**a**) C_{493}^{1072}. (**b**) C_{769}^{1492}. (**c**) C_{1438}^{1937}. (**d**) C_{1204}^{2784}.

Answer: If we repeat the argument and method of the previous Problems then we have:

$$\log\left(C_k^n\right) = \log\left[\frac{n!}{k!(n-k)!}\right] = \log\left[\frac{n \times (n-1) \times \cdots \times (n-k+1)}{k!}\right] = \sum_{i=n-k+1}^{n} \log(i) - \sum_{j=1}^{k} \log(j)$$

[48] "Permutations" and its alike (e.g. combinations) in such context should mean "number of permutations".

These sums can be easily calculated (by a spreadsheet or a computer code for instance) and their difference is obtained. On using logarithm to the base 10 we get:

(**a**) $C_{493}^{1072} = 10^{\log\left(C_{493}^{1072}\right)} \simeq 10^{319.592532861485} = 10^{0.592532861485} \times 10^{319} \simeq 3.91320735824353 \times 10^{319}$

(**b**) $C_{769}^{1492} = 10^{\log\left(C_{769}^{1492}\right)} \simeq 10^{447.143929164639} = 10^{0.143929164639} \times 10^{447} \simeq 1.39292959140671 \times 10^{447}$

(**c**) $C_{1438}^{1937} = 10^{\log\left(C_{1438}^{1937}\right)} \simeq 10^{478.277091080755} = 10^{0.277091080755} \times 10^{478} \simeq 1.89274052481793 \times 10^{478}$

(**d**) $C_{1204}^{2784} = 10^{\log\left(C_{1204}^{2784}\right)} \simeq 10^{825.190214220116} = 10^{0.190214220116} \times 10^{825} \simeq 1.54958077671787 \times 10^{825}$

PE: Calculate the following binomial coefficients (approximately):
(**a**) C_{526}^{1169}. (**b**) C_{839}^{1613}. (**c**) C_{1992}^{2184}. (**d**) C_{1333}^{2901}.

5. Calculate the following multinomial coefficients:
(**a**) $C_{4,5,6,19}^{34}$. (**b**) $C_{3,12,16,27,29}^{87}$. (**c**) $C_{11,15,22,36,44,65}^{193}$. (**d**) $C_{4,54,77,97,103,137}^{472}$.

Answer: If we follow the argument and method of the previous Problems then we have:

$$\begin{aligned}
\log\left(C_{n_1,n_2,\ldots,n_k}^{n}\right) &= \log\left[\frac{n!}{n_1!\,n_2!\ldots n_k!}\right] \\
&= \log(n!) - \log(n_1!) - \log(n_2!) - \cdots - \log(n_k!) \\
&= \sum_{i=1}^{n}\log(i) - \sum_{i=1}^{n_1}\log(i) - \sum_{i=1}^{n_2}\log(i) - \cdots - \sum_{i=1}^{n_k}\log(i) \\
&= \sum_{i=1}^{n}\log(i) - \sum_{j=1}^{k}\left(\sum_{i=1}^{n_j}\log(i)\right)
\end{aligned}$$

These sums can be easily calculated (e.g. by a spreadsheet or a computer code) and their algebraic sum is obtained. On using logarithm to the base 10 we get:

(**a**) $C_{4,5,6,19}^{34} = 10^{\log\left(C_{4,5,6,19}^{34}\right)} \simeq 10^{0.0683449986} \times 10^{15} \simeq 1.17042879714 \times 10^{15}$

(**b**) $C_{3,12,16,27,29}^{87} = 10^{\log\left(C_{3,12,16,27,29}^{87}\right)} \simeq 10^{0.5611911824} \times 10^{50} \simeq 3.64075271761 \times 10^{50}$

(**c**) $C_{11,15,22,36,44,65}^{193} = 10^{\log\left(C_{11,15,22,36,44,65}^{193}\right)} \simeq 10^{0.155895067} \times 10^{131} \simeq 1.43184189946 \times 10^{131}$

(**d**) $C_{4,54,77,97,103,137}^{472} = 10^{\log\left(C_{4,54,77,97,103,137}^{472}\right)} \simeq 10^{0.265327326} \times 10^{322} \simeq 1.84215990731 \times 10^{322}$

PE: Calculate the following multinomial coefficients (approximately):
(**a**) $C_{2,6,6,7}^{21}$. (**b**) $C_{4,7,18,22,31}^{82}$. (**c**) $C_{24,33,34,56,61,77}^{285}$. (**d**) $C_{2,45,71,84,91,99}^{392}$.

6. Calculate the following powers:
(**a**) 0.45^{1693}. (**b**) 0.71^{3712}. (**c**) 1236^{2019}. (**d**) 6391^{3361}.

Answer: We have:

$$\log\left(x^{y}\right) = y\log(x)$$

So, on using logarithm to the base 10 we get:

(**a**) $0.45^{1693} = 10^{1693\log(0.45)} \simeq 10^{-587.111214178343} \simeq 7.74079955396 \times 10^{-588}$

(**b**) $0.71^{3712} = 10^{3712\log(0.71)} \simeq 10^{-552.129009554793} \simeq 7.43002791118 \times 10^{-553}$

(**c**) $1236^{2019} = 10^{2019\log(1236)} \simeq 10^{6242.785292449900} \simeq 6.09947491993 \times 10^{6242}$

(**d**) $6391^{3361} = 10^{3361\log(6391)} \simeq 10^{12790.5167957807} \simeq 3.28697030560 \times 10^{12790}$

PE: Calculate the following powers (approximately):
(**a**) 0.32^{981}. (**b**) 0.85^{2395}. (**c**) 2516^{1705}. (**d**) 5928^{4920}.

2.3.2 Use of Analytical Approximations

There are many analytical approximation rules and formulae (found in the literature of calculus for instance) that can be used for overcoming the difficulties of extreme calculations. The well-known example of these rules and formulae (which is widely used in the probability theory) is the Stirling approximation for factorials of large numbers which is given by:[49]

$$n! \simeq \sqrt{2\pi n}\, n^n\, e^{-n} \qquad \text{(large } n\text{)} \tag{37}$$

Although this may not be of great help for calculating the factorial itself, it can be useful for simplifications (e.g. by cancellation) when the factorial is involved in a formula.[50] We also note that $n^n\, e^{-n}$ which can be written and calculated as $(n/e) \times (n/e) \times \cdots \times (n/e)$ may also help in easing the calculations (i.e. by keeping the size of the involved numbers manageable and under control, as we did earlier in some of our calculational tricks by pairing large factors with small factors).

Problems

1. Plot a graph for the ratio of the Stirling approximation to the corresponding factorial and comment.
   Answer: See Figure 5.
   Comment: We note the following:
   • The approximation improves with increasing n. In fact, the ratio approaches 1 asymptotically from below.
   • The approximation is good even for low n, and hence the "large n" restriction is to ensure greater accuracy and to indicate that the approximation improves with increasing n.
   • The Stirling approximation is always lower in value than the factorial, and this should be taken into account when assessing results obtained by the Stirling approximation (e.g. when assessing the type of errors introduced by this approximation).
   PE: Investigate the use of Stirling's approximation in the probability theory.

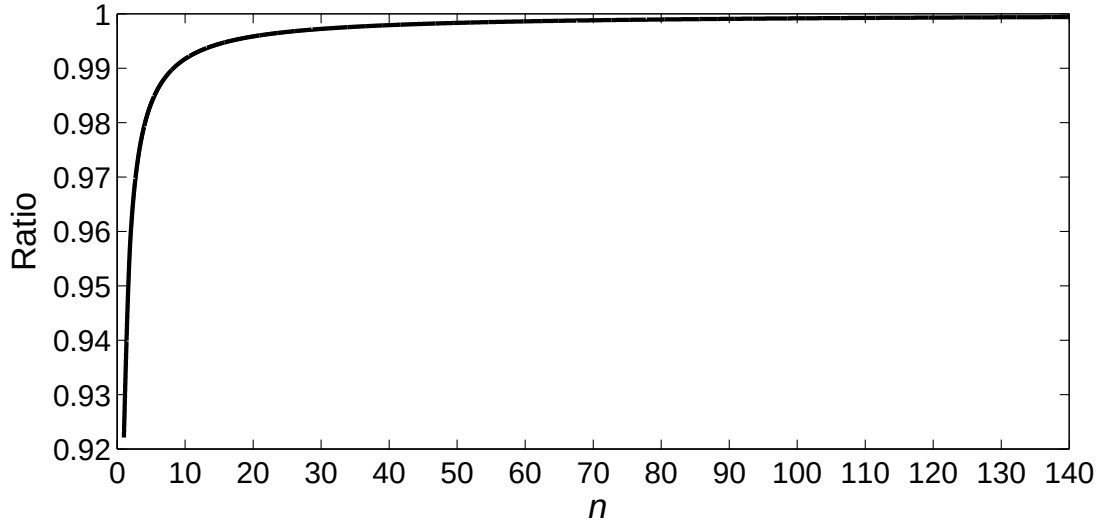


Figure 5: The plot of the ratio of Stirling's approximation to the corresponding factorial. See Problem 1 of § 2.3.2.

[49] For more details about Stirling's formula, the reader is referred to the textbooks of calculus.

[50] For example, Stirling's approximation (as given by the above formula) is widely used in analytical derivations and theoretical arguments (possibly much more than its use in practical calculations especially these days where computers usually overcome most practical difficulties).

2.3.3 Use of Computing and Programming

Some problems (in probability and related subjects) require using computational methods and programming languages to solve (if they are solvable). This is when the use of normal calculators (or similar tools like spreadsheets) is impossible or impracticable. For example, calculating cumulative probability distributions involving factorials or binomial coefficients of large numbers may require designing and writing special codes or routines that employ relatively complicated computational methods and mathematical operations (noting that direct application of elementary operations like multiplication cannot do the job due to their limitations in application). We will see in the Problems of this subsection some simple examples of the use of coding and programming to do some probability-related calculations. We will also meet in the future more such examples (see for instance Problem 12 of § 4.1.2, Problem 7 of § 4.1.3, Problem 11 of § 4.1.4 and Problem 9 of § 4.2.2).

Problems

1. Write a simple program that can calculate the factorial of large numbers.
   Answer: See the Factorial.cpp code.
   PE: Plot a flowchart representing the algorithm of the Factorial.cpp code.
2. Write a simple program that can calculate permutations involving large numbers.
   Answer: See the Permutation.cpp code.
   PE: Plot a flowchart representing the algorithm of the Permutation.cpp code.
3. Write a simple program that can calculate binomial coefficients involving large numbers.
   Answer: See the BinomialCoefficient.cpp code.
   PE: Plot a flowchart representing the algorithm of the BinomialCoefficient.cpp code.
4. Write a simple program that can calculate multinomial coefficients involving large numbers.
   Answer: See the MultinomialCoefficient.cpp code.
   PE: Plot a flowchart representing the algorithm of the MultinomialCoefficient.cpp code.
5. Write a simple program that can calculate (non-negative integer) powers of positive numbers (small and large).
   Answer: See the Power.cpp code.
   PE: Plot a flowchart representing the algorithm of the Power.cpp code. Also explain why in the final stage (i.e. output stage) the code distinguishes between the case of negative logarithm and the case of non-negative logarithm.

2.3.4 Use of Other Functions and Distributions

Some functions and distributions can be approximated by other functions and distributions which are easier to calculate or estimate and may be the only possible or viable choice for solving the given problem. For example, the Poisson distribution (see § 4.1.4) can approximate the binomial distribution (see § 4.1.2) under certain conditions and hence in some circumstances we may use the Poisson distribution to solve a binomial distribution problem. Similarly, the normal distribution (see § 4.2.2) can approximate the binomial distribution under certain conditions and hence in some circumstances we may use the normal distribution to solve a binomial distribution problem. We will meet in the future more clarifications and examples about the use of functions and distributions to approximate other functions and distributions.

Problems

1. Give some examples (related to the probability theory) of approximations of functions and distributions made by using other functions and distributions.
   Answer: For example:
   • The Poisson probability distribution is used (under certain conditions) to approximate the binomial probability distribution (see § 4.1.4).
   • The binomial probability distribution is used (under certain conditions) to approximate the hypergeometric probability distribution (see Problem 5 of 4.1.6).
   • The normal distribution is used (under certain conditions) to approximate a number of other probability distributions such as the binomial and Poisson distributions (see for example § 4.2.2).

• We may even consider the use of the Stirling formula to approximate factorials as another example (see § 2.3.2 as well as Problem 39 of § 3.2).
The reader is referred to the subsections of § 6.2 for more details and examples.
PE: Give more examples (related to the probability theory) of functions and distributions used to approximate other functions and distributions.

2.4 Venn and Tree Diagrams

Venn diagrams and tree diagrams are two visual abstract devices which are commonly used in the literature of probability theory for the purpose of demonstration, illustration, simplification, and so on.[51] **Venn diagram** is strongly linked to the subject of sets (see § 2.1) where it employs graphic encirclement to show the relation between sets as collections of elements.[52] On the other hand, **tree diagram** is strongly linked to the subject of counting (see § 2.2) and probability where it is used to demonstrate the branching of various possibilities for the outcome of probabilistic trials and experiments. It is also used to demonstrate and outline hierarchical structures (regardless of their link to probability) and other similar purposes.

In the following Problems we give some examples of these devices in the context of sets and counting (which were investigated in § 2.1 and § 2.2). Further investigation of these devices (as well as other devices) in the context of probability will be given in § 3.4.

Problems

1. Describe tree diagram in some detail.
Answer: Tree diagram is a graphic device used to list all the possibilities for the outcome of a series of trials (or observations) where each trial (or observation) can take place in a finite number of ways.[53] Tree diagram branches from left to right (or from top to bottom) starting usually in a single point (or node) and branching consecutively to multiple points where from each point (except the end points) a number of branches (equal to the number of possibilities that can emerge from that point) appear. The end points of tree diagram represent the ultimate outcomes of that series of trials (or observations) where each end point is identified distinctively by the route that connects it to the start point.
It is noteworthy that tree diagram is commonly used to facilitate the enumeration and listing in a systematic way and hence reducing confusion and avoiding error. For example, if we want to list all the 3-digit permutations of the set $\{1, 2, 3, 4\}$ in an improvised and spontaneous way it may be confusing and prone to errors and mistakes, but if we do it through the use of a tree diagram (as we will do in Problem 7) then it will be easy and robust.
PE: Describe Venn diagram in detail.
2. A set U of (lower case) English letters contains the elements: $c, k, t, r, o, p, j, s, d, w, z$. A and B are two subsets of U where A contains the elements: c, s, p, z, k and B contains the elements: c, j, s, d, w.
(**a**) Draw a Venn diagram representing these sets.
(**b**) Use your Venn diagram to identify the following: $\overline{A}$, $\overline{B}$, $A \cap B$ and $A \cup B$.
Answer:
(**a**) The Venn diagram is given in Figure 6.
(**b**)
$\overline{A} = \{t, r, o, j, d, w\}$ $\overline{B} = \{k, t, r, o, p, z\}$ $A \cap B = \{c, s\}$ $A \cup B = \{c, k, p, j, s, d, w, z\}$
PE: Do the following:
(**a**) Draw a Venn diagram representing the letters of "PROBABILITY THEORY" where the sets Ω

[51] We already met an important use of Venn diagrams in proving identities of set theory (see Problem 21 of § 2.1).
[52] Venn diagram is usually made of a rectangle (representing the universal set, or the sample space in the case of probability) inside which one or more closed curves (representing sets, or events in the case of probability) are sketched. These closed curves encircle the members (symbolized as points) of the sets which these curves represent and hence they identify theses sets and demonstrate the relationships between them (e.g. if they are disjoint or not).
[53] In fact, this description is appropriate in the context of probability (which is our prime interest in this book). As indicated earlier, tree diagram has more general uses and purposes than this (e.g. in demonstrating and outlining hierarchical structures).

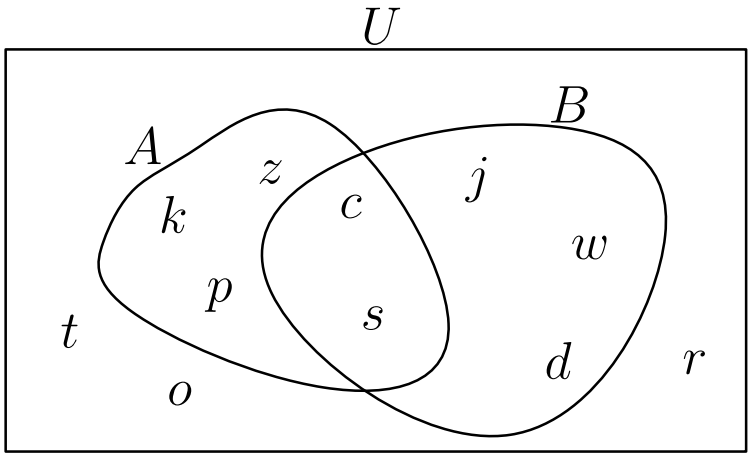


Figure 6: Venn diagram of Problem 2 of § 2.4.

and Δ represent the letters of "PROBABILITY" and "THEORY" respectively.
(**b**) Use your Venn diagram to identify the following: $\overline{\Omega}$, $\overline{\Delta}$, $\Omega \cap \Delta$, $\Omega \cup \Delta$, $\overline{\Omega} \cap \Delta$ and $\overline{\Omega} \cup \overline{\Delta}$.

3. Use Venn diagrams to demonstrate intersection, union and complement of two sets (A and B). Also demonstrate disjoint sets.
Answer: See Figure 7.

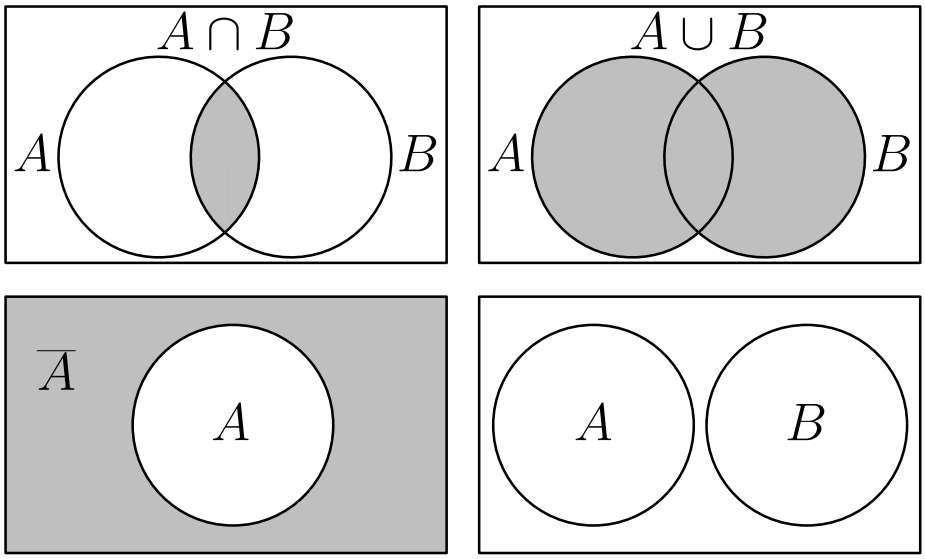


Figure 7: Venn diagrams demonstrating intersection (upper left), union (upper right), complement (lower left) and disjoint sets (lower right). The shaded area in the first three frames represents the sets of interest (i.e. intersection, etc.). See Problem 3 of § 2.4.

PE: Use Venn diagrams to demonstrate $A - B$ and $B - A$ for the following cases:
(**a**) $A \cap B \neq \varnothing$ with $A \not\subset B$ and $B \not\subset A$. (**b**) A is a proper subset of B. (**c**) $A = \overline{B}$.

4. Prove or disprove the following relations using Venn diagrams:
(**a**) $(A - B) \cup (B - A) = (A \cup B) - (A \cap B)$. (**b**) $A \cup (B - C) = (A - B) \cup (B - C)$.
Answer:
(**a**) In Figure 8 we constructed Venn diagrams (in stages) for the left hand side of this relation in the

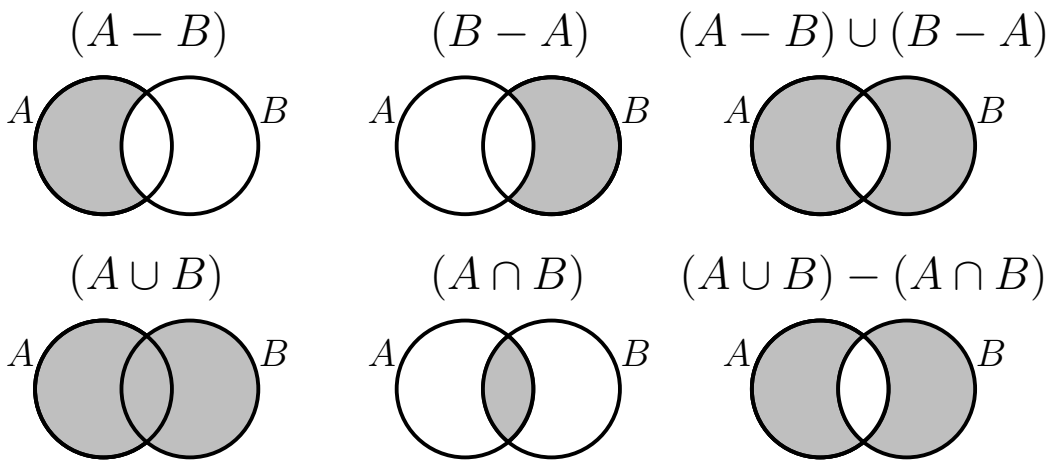


Figure 8: The Venn diagrams of part (a) of Problem 4 of § 2.4.

upper row and for the right hand side of this relation in the lower row. As we see, the Venn diagrams of the two sides are the same and hence this relation is correct.
(**b**) In Figure 9 we constructed Venn diagrams (in stages) for the left hand side of this relation in the upper row and for the right hand side of this relation in the lower row. As we see, the Venn diagrams of the two sides are not the same and hence this relation is incorrect (in general).

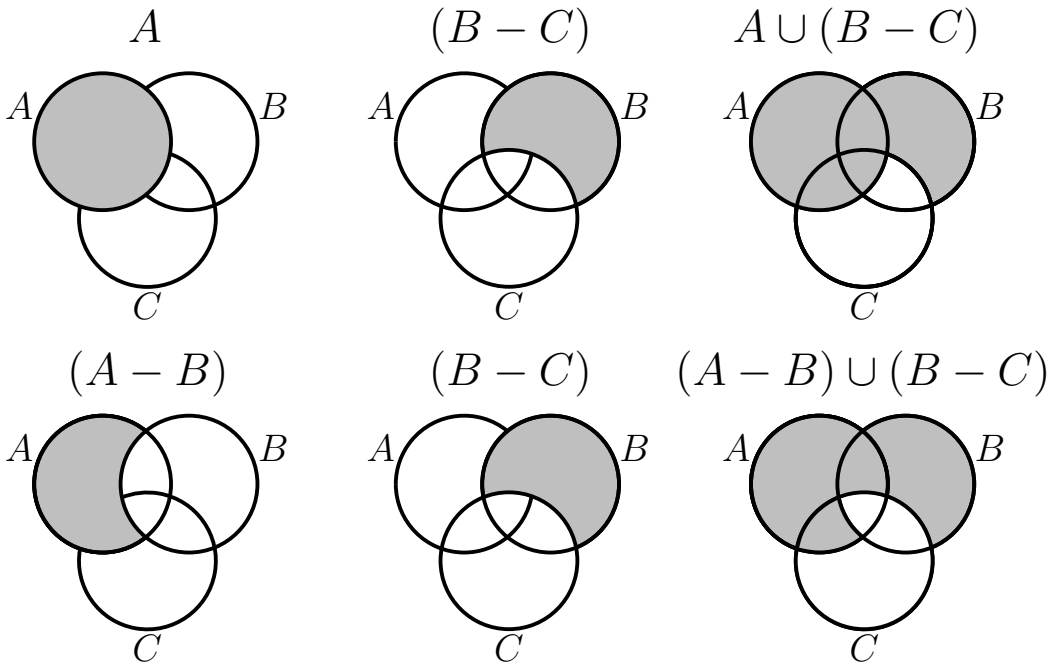


Figure 9: The Venn diagrams of part (b) of Problem 4 of § 2.4.

PE: Prove or disprove the following relations using Venn diagrams:
(**a**) $(B \cap C) - A = (B - A) \cap (C - A)$. (**b**) $\overline{(B - A)} \cup \overline{(A - B)} = U$.

5. A college has 3 main departments: mathematics, science and engineering. The mathematics department is divided to 2 branches: pure and applied. The science department is divided to 4 branches:

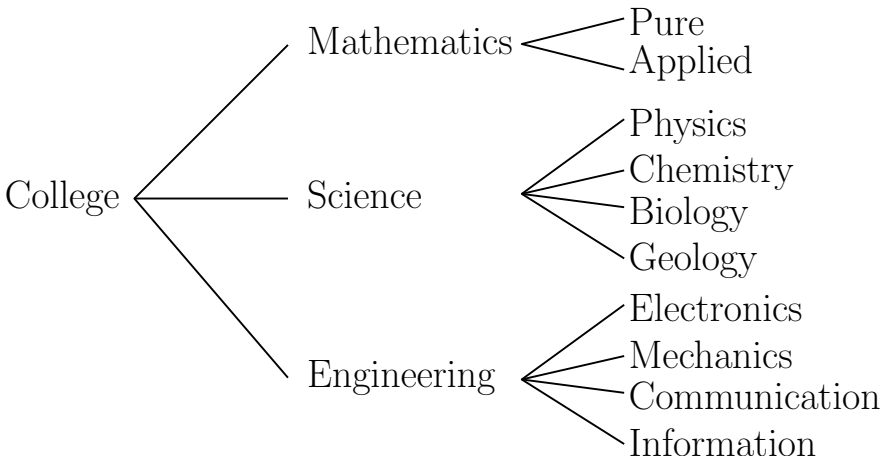


Figure 10: The tree diagram of Problem 5 of § 2.4.

physics, chemistry, biology and geology. The engineering department is divided to 4 branches: electronics, mechanics, communication and information. Make a tree diagram to demonstrate the structure of this college.

Answer: See Figure 10.

PE: Put the following in a simple tree diagram: fish, mammals, animals, birds, reptiles, cold-blooded, amphibians, warm-blooded.

6. Make a simple tree diagram to demonstrate the finity/infinity and countability of sets.
   Answer: See Figure 11.

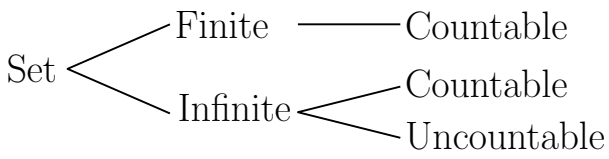


Figure 11: The tree diagram of Problem 6 of § 2.4.

PE: Put the following in a simple tree diagram: commutative, set operations, union, non-commutative, difference, intersection.

7. List all the 3-digit permutations of the set $\{1, 2, 3, 4\}$ through the construction of a tree diagram.
   Answer: See Figure 12. As we see, the enumeration and listing of these permutations through the tree diagram make the job easy and less prone to error because it is visual and systematic. Also see Problem 1.
   PE: List all the 4-letter permutations of the set $\{a, b, c, d\}$ through the construction of a tree diagram.
8. A typical Venn diagram is generally divided into 2^n distinct regions (with regard to the inclusion and exclusion of members) where $n = 1, 2, \ldots$ is the number of sets represented by the diagram. Explain this.
   Answer: If the diagram contains only one set (i.e. $n = 1$) then we have only two distinct regions: either inside or outside (the closed curve representing) the set and hence we have 2^1 distinct regions. If the diagram contains only two sets (i.e. $n = 2$) then we have four distinct regions: inside both sets, or outside both sets, or inside only the first set, or inside only the second set, and hence we have 2^2 distinct regions. By induction we can generalize this pattern and conclude that Venn diagram is generally divided into 2^n distinct regions.
   PE: Link this (i.e. having 2^n distinct regions in a typical Venn diagram) to the binomial theorem and

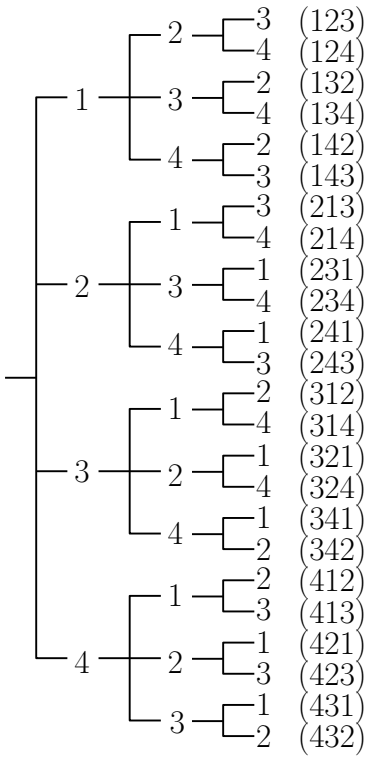


Figure 12: The tree diagram of Problem 7 of § 2.4.

to an identity investigated in Problem 25 of § 2.2.

9. Classify the 2^n regions found in Problem 8 and find the number of regions in each type.
   Answer: We can classify these regions into $n+1$ types:
   • Regions not under any set: we have $C_0^n = 1$ region of this type.
   • Regions (each of which is) under exactly one set: we have $C_1^n = n$ regions of this type.
   • Regions (each of which is) under exactly two sets: we have $C_2^n = \frac{n(n-1)}{2}$ regions of this type.
   ⋮
   • Regions (each of which is) under exactly $n-1$ sets: we have $C_{n-1}^n = n$ regions of this type.
   • Regions (each of which is) under all sets: we have $C_n^n = 1$ region of this type.
   As a check, we have (see part j of Problem 25 of § 2.2):

$$C_0^n + C_1^n + C_2^n + \cdots + C_{n-1}^n + C_n^n = \sum_{k=0}^{n} C_k^n = 2^n$$

   Note: as we will see (refer to § 3.2) events in the sample space of probability correspond to sets in the universal set. Accordingly, the above classification should apply to events in the sample space of any particular probability problem.
   PE: Based on the above findings, try to make a link between Venn diagrams, binomial theorem and binomial coefficients.

Chapter 3
Mathematics of Probability

In this chapter we present the basic mathematics of probability theory which we need and use in this book to solve probability problems.

3.1 Axioms of Probability Theory

The entire theory of probability can be constructed from a few axioms (or hypotheses). The number and content of these axioms may differ between authors. However, the following set of three axioms (with minor variations between authors) seems to be the most commonly used in recent times:[54]

1. Probabilities are real numbers between 0 and 1, i.e. $0 \leq P \leq 1$.
2. The probability of the sample space[55] is 1, i.e. $P(S) = 1$.
3. The probabilities of disjoint events add, i.e. $P(A \cup B \cup \ldots) = P(A) + P(B) + \cdots$.

In the following Problems we will briefly discuss the role of axioms in constructing a theory (in general) as well as some general aspects about the characteristics of axioms. Moreover, we will give a few examples for the role and use of the axioms of probability in the construction of probability theory in the next section (i.e. § 3.2).

Problems

1. Outline some of the roles that axioms can play in the construction and representation of a given theory.
   Answer: For example:
   • The axioms provide the starting points or propositions to derive the other propositions and premises of the theory.
   • The axioms provide the rational justification for the structure of the theory (in general) and its individual parts and components (in particular).
   • The axioms can play a role in organizing, classifying and structuring the given facts (such as observations) which make the body and content of the theory.
   • The axioms can reflect the spirit and basic essence of the theory and hence they can provide a simple and brief outline of the theory to be used for making general judgments and concise assessments.
   PE: Give more potential roles that axioms can play in the construction and representation of a given theory.
2. Is there any Problem in having more than one set of axioms for a given theory?
   Answer: Not at all. The purpose of any set of axioms is to construct the given theory and put it on a rational footing, and hence any set of axioms that can achieve this should be fine. In general, there is no unique way for constructing a theory and hence there is no unique set of axioms for constructing a specific theory. A given theory is like a house or a building which can be constructed in many different ways using different designs or/and construction materials, and all these ways should fulfill the objective of the required construction as long as they achieve and realize the construction.
   PE: What can you conclude from the possibility of having more than one set of axioms for a given theory?
3. Give some factors that favor/disfavor some sets of axioms against other sets of axioms for constructing a given theory.
   Answer: In our view, the main factors that should be considered in constructing or choosing a set of axioms to build a given theory are:

[54] We note that P symbolizes the probability (of a random occurrence or event) which is a single-valued real function of random variable (or variables).

[55] As we will see in § 3.2, the sample space is the set of all possible outcomes of a given trial.

• **Clarity**: it should be obvious that clarity is an advantage since it reduces the required work and minimizes confusion and mistakes, and hence more clear sets of axioms should be favored against less clear sets.
• **Simplicity**: so the set of axioms that leads to the construction and derivation of the theory more easily and simply is favored in comparison to another set of axioms which is less simple and hence it causes more complications and difficulties.
• **Intuitivity**: more intuitive sets of axioms should be favored against less intuitive sets. In fact, this factor can be included in the factors of clarity and simplicity.
• **Minimality**: this means being less in number and more concise. So, in principle a set of axioms that is less in number or/and content and hence it is more concise should be favored (from this perspective) against other sets. Again this is an advantage in general since it usually reduces the required work and enhances clarity and simplicity.
• **Richness**: some sets of axioms may lead to results and consequences that cannot be obtained from other rival sets (although both sets can construct the main body of the given theory in general). So, richness should be considered as an advantageous factor in constructing and choosing a set of axioms for the construction of a given theory.
PE: Give more potential factors that favor/disfavor some sets of axioms against rival sets of axioms for constructing a given theory.

4. Starting from different sets of axioms could lead to the same theory and could lead to different theories.[56] Comment on this.
Answer: Some kinds of differences between sets of axioms do not affect the essence (i.e. they are virtually superficial) and hence they lead to the same theory, while other kinds of differences between sets of axioms do affect the essence (i.e. they are essential) and hence they lead to different theories.
PE: Give some authentic examples (from mathematics or science) of different sets of axioms that lead to the same theory and other examples of different sets of axioms that lead to different theories.

3.2 The Basics of Probability Theory

In simple words, **probability** is a measure of the likelihood of something to occur (or not). It also reflects our partial ignorance of the surrounding conditions and circumstances which leads to our inability to predict events and occurrences definitely and deterministically (see Problem 1). Quantitatively, the probability is normally expressed as a real number between 0 and 1 (inclusive) where 0 represents certainty of non-occurrence and 1 represents certainty of occurrence. An **experiment** (or **trial**) in the context of probability theory is an action that produces one of a number of possible outcomes (see Problem 2).[57]

The **sample space** is the set of all possible (individual) outcomes in a given trial, while an **event** is a subset of the sample space. The individual outcomes in the sample space are usually described as **points** of the sample space (or **sample points**). The elements (or points) of the sample space must be exhaustive (which we indicated by "all") and mutually exclusive (which we indicated by "individual"). The individual points in the sample space may be equally likely (i.e. have the same probability) and may be not (where in the former/latter case the sample space is described as **uniform**/**non-uniform**). It is worth noting that the sample space can be **discrete** (made of countable points) and can be **continuous** (whose sample points constitute a continuum), and the discrete can be **finite** (made of finite number of sample points) and can be **infinite** (made of infinite number of sample points). We also note that the sample points may be described as **elementary events** or **simple events** because they cannot be split to simpler events.

An event **occurs** if the outcome of the trial belongs to that event (i.e. the outcome is a member of the event). The probability of occurrence of a given event A in a given trial with a sample space of a *finite*

[56] For example, different sets of axioms (in the context of probability) lead to the same probability theory (i.e. in essence), while different sets of axioms (in the context of 2D geometry) lead to different geometries (e.g. Euclidean and non-Euclidean geometries).
[57] We note that "experiment" and "trial" may be used differently, e.g. experiment is a series of trials.

number of *equally likely* sample points is given by:

$$P(A) = \frac{N_A}{N_S} \qquad (N_A \geq 0,\ N_S > 0,\ N_A \leq N_S) \tag{38}$$

where $P(A)$ is the probability of A, N_A is the number of elements of A and N_S is the number of elements of the sample space.[58] If the points of the sample space are not equally likely then the probability of an event A is the sum of the probabilities of all the sample points that belong to A.[59] So, in this case we need to define and assign beforehand (non-equal) probabilities to the individual points of the sample space (e.g. by experiment or guess or hypothesis) under the condition that their sum is 1 (as required by axiom 2 of probability theory; see § 3.1).

An event is described as **impossible** (or **empty**) if it contains no element of the sample space and described as **certain** (or **entire**) if it contains all the elements of the sample space. The probability of the impossible event is 0 (i.e. it cannot occur) and the probability of the certain event is 1 (i.e. it must occur) while the probability of other types of event (which we may call **probable event**, i.e. neither impossible nor certain) is between 0 and 1 (i.e. it can occur but not necessarily). Accordingly, we can write:

$$0 \leq p \leq 1 \qquad \text{or} \qquad 0 \leq P \leq 1 \tag{39}$$

where p and P stand for the probability of an event (say A). As indicated already, the sum of probabilities of all the sample points is equal to 1.

Two (or more) events that make the entire sample space and they cannot occur simultaneously are called **complementary events**.[60] The probability of occurrence of complementary events (i.e. one of them unspecified) is 1 since they exhaust the sample space (noting that the probability of the sample space is 1 as indicated already). Accordingly, we have:

$$P(A) + P(\overline{A}) = 1 \tag{40}$$

where A and $\overline{A}$ are complementary events.

The probability of occurrence of "A AND B" is symbolized by $P(A \cap B)$ and the probability of occurrence of "A OR B" is symbolized by $P(A \cup B)$. We note that:

$$P(A \cap B) = P(B \cap A) \qquad \text{and} \qquad P(A \cup B) = P(B \cup A) \tag{41}$$

This is because of the commutativity property of intersection and union of sets, i.e. $A \cap B = B \cap A$ and $A \cup B = B \cup A$ (see Eqs. 14 and 15).

Two events are **mutually exclusive** (or **disjoint** or **incompatible**) if they cannot occur simultaneously (i.e. if one occurs the other cannot occur at the same time). Accordingly, if A and B are mutually exclusive events then:

$$P(A \cap B) = 0 \tag{42}$$

According to the **addition law of probability**, if A and B are two events then the probability of occurrence of "A OR B" is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{43}$$

The subtraction of $P(A \cap B)$ in this equation should be obvious because it is counted twice in $P(A)+P(B)$, i.e. once in $P(A)$ and once in $P(B)$ (also see part e of Problem 14).

According to the **addition law for mutually exclusive events**, if A and B are mutually exclusive events then the probability of occurrence of "A OR B" is:

$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \varnothing) \tag{44}$$

[58] In fact, this primarily represents what we may call the repetition or counting method for defining and quantifying probabilities. We should also note that being elements of the sample space also implies being exhaustive and mutually exclusive (as indicated already). Also see § 1.2 and § 1.3 as well as Problem 2.

[59] In fact, this definition is general and hence it applies even to the case of equally likely sample points.

[60] In fact, this is based on the definition of complement of event A (symbolized as $\overline{A}$) which is the set of outcomes (in the sample space) that do not belong to A (see § 2.1).

This equation can be obtained from Eq. 43 noting that for mutually exclusive events $P(A \cap B) = 0$ (see Eq. 42).[61]

The **conditional probability** $P(B\,|A)$ of event B given that event A occurred is:

$$P(B\,|A) = \frac{P(A \cap B)}{P(A)} \qquad \Big[P(A) \neq 0\Big] \tag{45}$$

From this equation we get the following relation (which is the **multiplication law for associated events**):

$$P(A \cap B) = P(A)\,P(B\,|A) \tag{46}$$

Now, if we exchange the symbols of A and B in the last equation then we get $P(B \cap A) = P(B)\,P(A\,|B)$. If we also note (see Eq. 41) that $P(A \cap B) = P(B \cap A)$ then we get the following relation:

$$P(A \cap B) = P(A)\,P(B\,|A) = P(B)\,P(A\,|B) \tag{47}$$

As we will see (refer to § 6.1), this is the origin of the Bayes theorem. It is important to note that unlike $P(A \cup B)$ and $P(B \cap A)$ where the order of events does not matter (see Eq. 41), the order of events in the equation of conditional probability is important and hence in general we have:

$$P(A\,|B) \neq P(B\,|A) \tag{48}$$

Yes, they are equal *iff* $P(A) = P(B)$ as can be seen from Eq. 47.

Two (or more) events are **independent** if the occurrence (or not) of any one of them has no effect on the probability of occurrence (or not) of the other(s). In technical terms, event B is independent of event A *iff*

$$P(B\,|A) = P(B) \tag{49}$$

Now, if we combine this equation with Eq. 46 we get the following relation for the probability of independent events:

$$P(A \cap B) = P(A)\,P(B) \tag{50}$$

In fact, the last relation is the necessary and sufficient condition for the independence of two (non-empty) events, i.e. A and B are independent *iff* $P(A \cap B) = P(A)\,P(B)$ (see part d of Problem 14). As we see, the independence of two events means there is no involvement of conditional probability between them.[62]

Problems

1. Comment on the above statement: "It also reflects our partial ignorance of the surrounding conditions and circumstances which leads to our inability to predict events and occurrences definitely and deterministically".

   Answer: We note the following:

   • This statement is true classically but not necessarily in general, e.g. the probability at the quantum level may be intrinsic to the phenomena and not because of our partial ignorance of the surrounding conditions and circumstances.

   • Probability "reflects our partial ignorance" when attributed to specific events and occurrences in the presence of an observer, but in general (considering repetitive trials in similar circumstances with no consideration of observer) it reflects in a sense indeterminism in reality either because the reality itself is not determined or because the reality is not completely defined and identified (i.e. it is surrounded by an element of ambiguity or generality) to be completely determined. The readers are also advised to read § 1.3 carefully to avoid potential misunderstanding and confusion.[63]

   PE: Compare (and hence justify the seeming difference) between the above statement and what we presented in § 1.3 about the distinction between subjective and objective probabilities.

[61] In fact, if we consider the axioms of the probability theory (which we investigated in § 3.1) then this is an axiom rather than being derived or obtained from Eq. 43. Actually, this is axiom 3 which is used in the derivation of Eq. 43 (see part e of Problem 14).

[62] It is important to distinguish between "independent" and "disjoint" events (noting that "disjoint" could wrongly suggest being independent). Also see Problem 29.

[63] Some of these issues are investigated in our book "The Epistemology of Quantum Physics".

2. Give more details about the outcomes considered in the trials of probability theory and their nature and conditions.
   Answer: In general, we assume the following conditions about the outcomes in the trials of probability theory:
   • The outcomes should be well defined and reproducible, i.e. the experiment will generate the same outcomes if repeated.
   • Each one of the outcomes should have a definite and fixed (initial) probability and hence if the experiment is repeated then the outcomes will be produced with the same presumed probabilities (or frequencies).
   • The outcomes should be produced individually in each single trial (as indicated in the text by "produces one").
   Note: the purpose of some of these conditions is to include (non-deterministic) *random* occurrences and phenomena and exclude (non-deterministic) *haphazard* occurrences and phenomena (assuming the existence of such occurrences and phenomena) which do not follow any deterministic statistical pattern or regularity (unlike random ones which follow such pattern and regularity). We should also note that "experiment" could also include "observation".
   PE: Investigate thoroughly the difference between *random* events and *haphazard* events, and hence explain why *haphazard* events (unlike *random* events) are not subject to the probability theory.
3. Is the sample space of a given trial unique?
   Answer: In general, the sample space of a given trial is not unique since it depends on the involved considerations and categorizations and the intended observations and objectives in the given trial (and this appears more vividly and naturally in the cases where the trial has more than 2 possible outcomes). For example, the sample space of throwing a die once could be getting one of the numbers $\{1, 2, 3, 4, 5, 6\}$ or {getting a number < 3, getting a number ≥ 3} among many other possibilities (restricted by the condition that the sum of the probabilities of the points of the sample space must always be 1 as indicated earlier; see axiom 2 in § 3.1). However, a single possibility (or more) for the identification of the sample space is usually necessitated or favored by the nature of the problem and its requirements and objectives. For instance, if the objective of the experiment in the above example is to obtain the numbers on the faces of the die then the sample space is $\{1, 2, 3, 4, 5, 6\}$ (see Problem 9), while if the objective of the experiment in the above example is to obtain a winning number in a game of luck (where the winning number is set to be < 3) then the sample space is {getting a number < 3, getting a number ≥ 3}. Accordingly, in the Problems and examples of this book we generally assume and adopt a single sample space determined uniquely by the given conditions and requirements and the intended objectives of these Problems and examples.
   PE: Is it correct to say: given all the surrounding conditions and the intended objectives the sample space of a given trial is unique? If this is true, what about the labeling of the sample space and its points?
4. Discuss briefly how initial probabilities are attached to the discrete and continuous sample spaces.
   Answer: In the case of discrete sample space the initial probabilities are attached to the points of the sample space directly subject to the condition that their sum is unity.[64] In the case of continuous sample space the initial probabilities are attached to infinitesimal intervals of the sample space in the form of a distribution function subject to the condition that the integral of the distribution function over the entire sample space is unity.
   Note: for simplicity, the definitions and formulations that we gave in the text of this section are mainly phrased and presented in the terms and forms of discrete sample space. However, they are more general and should be understood so.
   PE: Can the sequence $1/2^n$ $(n = 1, 2, \ldots)$ represents the initial probabilities of an infinite discrete sample space? If so, can you give an example of a sample space whose initial probabilities (i.e. the probabilities of its points) are given by this sequence?
5. Referring to § 1.2 and Problem 4 of the present section, identify the relation between initial and

[64] It should be noted that this applies to infinite sample space as to finite sample space noting that the series of probabilities in the former case must be convergent (as demanded by being normalized to unity which we indicated already).

ultimate probabilities as well as their relation to the sample space.
Answer: In a discrete sample space, each sample point is given a specific initial probability (where the sum of all these initial probabilities is unity). The ultimate probability of any event (representing a number of sample points) is the sum of the initial probabilities of its sample points.
In a continuous sample space, each sample (infinitesimal) interval is given a specific initial probability (where the integral of all these initial probabilities is unity). The ultimate probability of any event (representing an interval) is the integral of the initial probabilities over that interval.[65]
Note: as indicated earlier, the initial probabilities are not required to be equal. We also note that initial probabilities are *probabilities* and hence they satisfy the condition $0 \leq p \leq 1$. However, if we assume that a sample point (or interval) can have a probability of 0 then it should be redundant in the calculations of probabilities and hence it can be eliminated from the sample space.
PE: Give a few examples of sample spaces with their initial probabilities (attached to their sample points) and ultimate probabilities (assigned to possible events). Consider both discrete and continuous cases.

6. What is the sample space for tossing a coin (considering the order in the case of multiple tossing):
(**a**) One time. (**b**) Two times. (**c**) Three times.
Answer: We symbolize sample space, head and tail with S, H and T respectively.
(**a**) $S = \{H, T\}$.
(**b**) $S = \{HH, HT, TH, TT\}$.
(**c**) $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.
PE: What is the sample space for tossing a coin four times?
7. Give an example of a (discrete) sample space with an infinite number of sample points.
Answer: Consider the experiment of throwing a die and counting the number of throws until "1" is obtained. The sample space of this experiment is: $S = \{1, 2, 3, \ldots\}$ which has an infinite number of sample points. This is because the required number of throws to get "1" can be any positive integer (i.e. we may get "1" in the first throw or in the second throw or ... etc.).
PE: Give another example of a (discrete) sample space with an infinite number of sample points.
8. Give examples of discrete and continuous sample space.
Answer: The sample spaces of Problem 6 are discrete (finite). The sample space of Problem 7 is discrete (infinite). The sample space representing the lifetime of an excited state of an atom (or the lifetime of a radioactive nucleus) is continuous.
PE: Give more examples of discrete and continuous sample space (both finite and infinite).
9. What is the sample space S for the numbers obtained when throwing a die (considering the order in the case of multiple throws):
(**a**) One time. (**b**) Two times. (**c**) Three times.
Answer: A die has 6 faces which we symbolize with $1, 2, 3, 4, 5, 6$.
(**a**) $S = \{1, 2, 3, 4, 5, 6\}$.
(**b**) If "14" for instance means getting "1" in the first throw and getting "4" in the second throw then:

$$\begin{aligned} S \;=\; & \{11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, 31, 32, 33, 34, 35, 36, \\ & 41, 42, 43, 44, 45, 46, 51, 52, 53, 54, 55, 56, 61, 62, 63, 64, 65, 66\} \end{aligned}$$

We may also express this more compactly as: $S = \Big\{ab \text{ where } a, b \in \{1, 2, 3, 4, 5, 6\}\Big\}$.
(**c**) $S = \Big\{abc \text{ where } a, b, c \in \{1, 2, 3, 4, 5, 6\}\Big\}$.
PE: Describe how you get the sample space for throwing a die 3 times from the sample space for throwing a die 2 times (assuming listing the individual sample points, rather than characterizing them in a compact form as we did in part c of this Problem).
10. Give examples of experiments whose sample spaces are given by:
(**a**) $S = \{5, 10, 15, \ldots\}$. (**b**) $S = \{\alpha : \ \pi \leq \alpha \leq 4\pi\}$. (**c**) $S = \{\text{red, green, blue}\}$.

[65] For simplicity and clarity we use "the integral of the initial probabilities over that interval" which is loose.

Answer:
(**a**) Throwing 2 dice simultaneously until a sum of 12 appears in attempt 5 or 10 or 15 (or a multiple of 5).
(**b**) Plotting a circle of radius r $(1 \leq r \leq 2)$ randomly and calculating its area α.
(**c**) Drawing a ball from a bag containing 5 red, 9 green and 3 blue balls.
PE: Repeat the Problem for the following sample spaces:
(**a**) $S = \{2, 3, 5, 7, 11, \ldots\}$. (**b**) $S = \{t : \ 0 \leq t < \infty\}$. (**c**) $S = \{\text{dog, cat, rabbit}\}$.

11. List a number of classifications for the sample space.
Answer: For example:
• Sample spaces may be classified as uniform and non-uniform (as explained already).
• Sample spaces may be classified as discrete and continuous (as explained already).
• We may also classify sample spaces as simple (corresponding to simple experiments) and composite (corresponding to composite experiments). For example, the sample space associated with an experiment in which a coin is tossed once is simple because the experiment is simple, while the sample space associated with an experiment in which a coin is tossed twice (or an experiment in which a die and a coin are tossed) is composite because the experiment is composite since it is a combination of a coin-tossing experiment with another coin-tossing experiment (or it is a combination of a die-tossing experiment with a coin-tossing experiment).
Note: composite sample space (of a given composite experiment) is usually obtained by the Cartesian product (see Problem 22 of § 2.1) of the sample spaces of the simple experiments that form the given composite experiment.[66]
PE: Discuss the sample spaces of Problem 9 in the light of the classifications given in the present Problem.

12. Identify all the events associated with tossing a coin 1 time.
Answer: The sample space in this case is $S = \{H, T\}$. So, the events are all the subsets of S, that is: $\{\}$, $\{H\}$, $\{T\}$, $\{H, T\}$.
PE: Identify 8 events associated with tossing a coin 2 times.

13. Justify the equation of conditional probability (i.e. Eq. 45) by explaining its rationale.
Answer: It is obvious that in conditional probability the sample space is reduced by the condition, i.e. the occurrence of A. In more explicit terms, by the occurrence of A we have a new sample space which is A. This should explain why we take the intersection of B with A and divide its probability by $P(A)$, since the conditional probability of B is the probability of the part of B that belongs to A divided by the probability of A which is the new sample space. So, what we actually do by dividing the probability of the part of B that belongs to A by $P(A)$ is to normalize this probability (i.e. the probability of the part of B that belongs to A).
PE: The multiplication law for associated events (i.e. Eq. 46) is a transformed form of the rule of conditional probability (i.e. Eq. 45). Explain the rationale of the multiplication law (in a similar or different way to the explanation of the conditional probability which is given in the answer of this Problem). Also try to find a similarity between the multiplication law and another law or principle which we met in counting (see § 2.2).

14. Show the following:
(**a**) $P(\overline{A}) = 1 - P(A)$.
(**b**) $P(A\,|B) = P(B\,|A)$ *iff* $P(A) = P(B)$.
(**c**) If B is independent of A then A is independent of B.
(**d**) A and B are independent *iff* $P(A \cap B) = P(A)\,P(B)$.
(**e**) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, i.e. the addition law of probability (see Eq. 43).
(**f**) $P(A \cap B) = P(A)\,P(B\,|A)$, i.e. the multiplication law of probability (see Eq. 46).
(**g**) If A_i $(i = 1, 2, \ldots)$ are disjoint and exhaustive events then $\sum_i P(A_i) = 1$.
Answer:
(**a**) As the occurrence of an outcome (in the sample space) is certain, we should have $P(A) + P(\overline{A}) = 1$

[66] In fact, there are many details about this issue (none of which is within our scope).

(noting that complementary events are mutually exclusive; see § 3.1 and Eq. 44), and this leads to $P(\overline{A}) = 1 - P(A)$. In fact, $P(A) + P(\overline{A}) = 1$ is a direct result of axiom 2 and axiom 3 (see § 3.1) noting that complementary events are mutually exclusive (and hence they follow axiom 3) and exhaustive (and hence they follow axiom 2).
(**b**) This can be seen from Eq. 47 $\big[$i.e. $P(A)\,P(B\,|A) = P(B)\,P(A\,|B)\big]$ because if $P(A\,|B) = P(B\,|A)$ then they can be canceled from both sides (assuming neither is zero) and we get $P(A) = P(B)$. Similarly, if $P(A) = P(B)$ then they can be canceled from both sides (assuming neither is zero) and we get $P(A\,|B) = P(B\,|A)$.
(**c**) If B is independent of A then we have:

$$\begin{aligned} P(B\,|A) &= P(B) && \text{(Eq. 49)} \\ P(A)\,P(B\,|A) &= P(A)\,P(B) && \\ P(B)\,P(A\,|B) &= P(A)\,P(B) && \text{(Eq. 47)} \\ P(A\,|B) &= P(A) && \big[P(B) \neq 0\big] \end{aligned}$$

i.e. A is independent of B.
(**d**) If A and B are independent then:

$$\begin{aligned} P(A)\,P(B) &= P(A)\,P(B\,|A) && (B \text{ is independent of } A, \text{ Eq. 49}) \\ &= P(A \cap B) && \text{(Eq. 47)} \end{aligned}$$

On the other hand, if $P(A \cap B) = P(A)\,P(B)$ then:

$$\begin{aligned} P(A \cap B) &= P(A)\,P(B) && \text{(given)} \\ P(A\,|B)\,P(B) &= P(A)\,P(B) && \text{(Eq. 47)} \\ P(A\,|B) &= P(A) && \big[P(B) \neq 0\big] \end{aligned}$$

i.e. A is independent of B (see Eq. 49). This similarly applies to B (or we use the result of part c), and hence we conclude that if $P(A \cap B) = P(A)\,P(B)$ then A and B are independent.
(**e**) We have $A = (A \cap B) \cup (A - B)$ and $B = (A \cap B) \cup (B - A)$ (see part e of Problem 20 of § 2.1). Now, if we note that $(A \cap B) \cap (A - B) = \varnothing$ and $(A \cap B) \cap (B - A) = \varnothing$ (see part d of Problem 20 of § 2.1), then by axiom 3 of probability (see § 3.1) we have:

$$\begin{aligned} P(A) &= P\big[(A \cap B) \cup (A - B)\big] = P(A \cap B) + P(A - B) \\ \text{and} \quad P(B) &= P\big[(A \cap B) \cup (B - A)\big] = P(A \cap B) + P(B - A) \end{aligned}$$

On adding these equations side by side we get

$$\begin{aligned} P(A) + P(B) &= P(A \cap B) + P(A - B) + P(A \cap B) + P(B - A) \\ P(A) + P(B) - P(A \cap B) &= P(A - B) + P(A \cap B) + P(B - A) \\ P(A) + P(B) - P(A \cap B) &= P\Big[(A - B) \cup (A \cap B) \cup (B - A)\Big] \\ P(A) + P(B) - P(A \cap B) &= P(A \cup B) \end{aligned}$$

where in line 3 we used axiom 3 of probability (noting that $A - B$, $A \cap B$ and $B - A$ are disjoint; see footnote [24] on page 22), and in line 4 we used the result of part (f) of Problem 20 of § 2.1.
(**f**) In Problem 13 we rationalized the equation of conditional probability (i.e. Eq. 45), and Eq. 46 is no more than a transformed form of Eq. 45. This should be enough as a proof (or justification) for Eq. 46.
(**g**) We have:

$$\sum_i P(A_i) = P\left(\cup_i A_i\right) = P(S) = 1$$

where equality 1 is because A_i are disjoint (see axiom 3 in § 3.1), equality 2 is because A_i are exhaustive, and equality 3 is because of axiom 2.

PE: Show the following by using the axioms of probability theory (see § 3.1) when convenient:
(**a**) If $A \subset B$ then $P(A) \leq P(B)$ (where A and B are events).
(**b**) $P(\varnothing) = 0$ (where $\varnothing$ represents the empty event).

15. If a single card is drawn randomly from a standard deck of cards, what is the probability of being a diamond?
Answer: In a standard deck of cards we have 52 cards all of which are equally likely to be drawn (as indicated by "randomly"). Moreover, the possible outcomes are exhaustive and disjoint. Hence, we can use Eq. 38 (noting that there are 13 diamonds in the deck), that is:

$$P(\text{diamond}) = \frac{\text{number of diamonds}}{\text{number of cards}} = \frac{13}{52} = \frac{1}{4}$$

PE: What is the probability of the drawn card to be a non-heart?

16. A 3-digit number (i.e. a number between 100 and 999) is chosen randomly. What is the probability that the sum of its digits is 3?
Answer: We have 900 3-digit numbers. Out of these 900 numbers, we have 6 numbers that satisfy this condition, i.e. 111, 102, 120, 201, 210, and 300. Hence, from Eq. 38 we get:

$$P(\text{sum of digits} = 3) = \frac{6}{900} = \frac{1}{150}$$

PE: Repeat the Problem for a sum of 4 (instead of 3).

17. A die is thrown randomly in a large number (say 6000) of identical trials.[67] The results were as follows:
face 1 is obtained in 995 trials, face 2 is obtained in 1004 trials, face 3 is obtained in 673 trials, face 4 is obtained in 994 trials, face 5 is obtained in 997 trials, and face 6 is obtained in 1337 trials. Estimate the probabilities of obtaining $1, 2, 3, 4, 5, 6$ for this die. Is this die fair?
Answer: If P_1 is the probability of obtaining face 1 (and the rest are similarly defined) then we have:

$$P_1 = \frac{995}{6000} \simeq \frac{1}{6} \qquad P_2 = \frac{1004}{6000} \simeq \frac{1}{6} \qquad P_3 = \frac{673}{6000} \simeq \frac{1}{9}$$

$$P_4 = \frac{994}{6000} \simeq \frac{1}{6} \qquad P_5 = \frac{997}{6000} \simeq \frac{1}{6} \qquad P_6 = \frac{1337}{6000} \simeq \frac{2}{9}$$

It is obvious that this die is not fair because (some of) the probabilities of the different faces are not equal. Also see § 1.2.
PE: How do you deal with such situations (i.e. estimating the probabilities statistically as done above) if the variable is continuous (such as distance or time) rather than discrete (as in the above example of throwing a die)? Would you suggest, for instance, dividing the continuous variable to small intervals and binning the outcomes (according to their values) in these intervals?

18. Give a simple example of impossible event, certain event and probable event.
Answer: If nine cards are numbered from 1 to 9 and one of these cards is drawn at random then getting a number larger than 9 is an impossible event, and getting a number between 1 and 9 (inclusive) is a certain event, while getting an odd number is a probable event.
PE: Do the following:
(**a**) Repeat the Problem for throwing a die 1 time.
(**b**) Give another example for impossible, certain and probable events from the nine cards example given in the answer.

[67] "Identical trials" should imply that the physical conditions of the die did not change during (and possibly by) these repetitive trials. It should also imply that the ambient conditions (including the individual, or machine, who throws the die and his physical and mental conditions) remained the same during these trials.

19. Give some simple examples of complementary events.
Answer: Examples of complementary events are:
• "Getting odd number" and "getting even number" in a die-throwing experiment.
• "Getting head" and "getting tail" in a coin-tossing experiment.
• "Passing the test" and "failing the test" (assuming they are the only possible outcomes of an exam).
PE: Give some examples of complementary events in science (e.g. physics) related to probability.
20. Give simple examples of mutually exclusive and non-"mutually exclusive" events.
Answer: In a die-throwing experiment, "getting odd number" and "getting even number" are mutually exclusive, while "getting odd number" and "getting number > 3" are not mutually exclusive.
PE: Give two more examples of mutually exclusive and non-"mutually exclusive" events from a die-throwing experiment.
21. Make a clear distinction between complementary events and mutually exclusive events.
Answer: Complementary events are necessarily mutually exclusive, but mutually exclusive events are not necessarily complementary. This is because complementary events are those mutually exclusive events that make the entire sample space. For example, "getting 1" and "getting 3" in a die-throwing experiment are mutually exclusive (because they cannot occur simultaneously) but they are not complementary (because they do not represent the entire sample space). Yes, "getting 1" and "not getting 1" are complementary (as well as mutually exclusive). Also, "getting 1", "getting 2", ... , "getting 6" are complementary (as well as mutually exclusive).
PE: Classify the following events as complementary or/and mutually exclusive or not (considering a trial in which we draw a ball from an urn containing 10 red, 6 black and 7 green balls):
(**a**) "Getting green ball" and "not getting black ball".
(**b**) "Getting black ball" and "getting red or green ball".
(**c**) "Getting red ball" and "getting black ball".
22. The medical staff of a small clinic consists of two male doctors and one female doctor as well as four female nurses and two male nurses. If a member of medical staff is picked randomly, symbolize the following probabilities: being a male doctor, being a female nurse, being a male or nurse, and being a doctor or female.
Answer: If D, N, M, F stand for doctor, nurse, male, female then these probabilities should be symbolized as follows:

$$P(M \cap D) \qquad P(F \cap N) \qquad P(M \cup N) \qquad P(D \cup F)$$

PE: Repeat the Problem for the following probabilities: being neither male nor nurse, being a female but not doctor, being a male nurse, and being a doctor or nurse. Try to correlate these probabilities (or some of them) to the above probabilities, i.e. those given in the Problem.
23. Calculate the probabilities of Problem 22.
Answer: Due to the simplicity of this Problem we solve it by simple count[68] instead of using formulae (noting that the formulae will be used for verifying these probabilities in Problem 25).
• We have two male doctors (out of nine staff) and hence $P(M \cap D) = 2/9$.
• We have four female nurses (out of nine staff) and hence $P(F \cap N) = 4/9$.
• We have eight members who are male or/and nurse and hence $P(M \cup N) = 8/9$. We may also say: we have only one female doctor (i.e. neither male nor nurse) and hence $P(M \cup N) = (9-1)/9 = 8/9$.
• We have seven members who are doctor or/and female and hence $P(D \cup F) = 7/9$. We may also say: we have only two male nurses (i.e. neither female nor doctor) and hence $P(D \cup F) = (9-2)/9 = 7/9$.
PE: Calculate the probabilities of the PE of Problem 22.
24. Which of the events of the 4 pairs of events in Problem 22 are independent and which are not?
Answer: The probabilities of D, N, M, F are:

$$P(D) = \frac{3}{9} = \frac{1}{3} \qquad P(N) = \frac{6}{9} = \frac{2}{3} \qquad P(M) = \frac{4}{9} \qquad P(F) = \frac{5}{9}$$

[68] We note that **simple count** is a basic and intuitive method for calculating probabilities and is based on the basic definition of probability (see Eq. 38). However, it is limited in applicability to very simple problems and limited in validity to the cases of equal initial probabilities.

To test if the events in the pair M, D (for instance) are independent or not we use the condition of Eq. 50, i.e. they are independent if $P(M \cap D) = P(M)\,P(D)$ and they are not independent if $P(M \cap D) \neq P(M)\,P(D)$ [see Eq. 50 and part (d) of Problem 14]. This test similarly applies to the other pairs. Accordingly:[69]

$$\begin{aligned} P(M)\,P(D) &= \frac{4}{9} \times \frac{1}{3} = \frac{4}{27} \neq \frac{2}{9} = P(M \cap D) \\ P(F)\,P(N) &= \frac{5}{9} \times \frac{2}{3} = \frac{10}{27} \neq \frac{4}{9} = P(F \cap N) \\ P(M)\,P(N) &= \frac{4}{9} \times \frac{2}{3} = \frac{8}{27} \neq \frac{2}{9} = P(M \cap N) \\ P(D)\,P(F) &= \frac{1}{3} \times \frac{5}{9} = \frac{5}{27} \neq \frac{1}{9} = P(D \cap F) \end{aligned}$$

Therefore, we conclude that the events in none of these pairs are independent.

PE: Repeat the Problem for the 4 pairs of events in the PE of Problem 22.

25. Use the multiplication and addition laws of probability to verify the probabilities $P(M \cap D)$, $P(F \cap N)$, $P(M \cup N)$ and $P(D \cup F)$ which we obtained in Problem 23 by simple count.

Answer: For $P(M \cap D)$ and $P(F \cap N)$ we use the multiplication law (i.e. Eq. 46), that is:

$$\begin{aligned} P(M \cap D) &= P(M)\,P(D\,|M) = \frac{4}{9} \times \frac{2}{4} = \frac{2}{9} \\ P(F \cap N) &= P(F)\,P(N\,|F) = \frac{5}{9} \times \frac{4}{5} = \frac{4}{9} \end{aligned}$$

For $P(M \cup N)$ and $P(D \cup F)$ we use the addition law (i.e. Eq. 43), that is:

$$\begin{aligned} P(M \cup N) &= P(M) + P(N) - P(M \cap N) = \frac{4}{9} + \frac{2}{3} - \frac{2}{9} = \frac{8}{9} \\ P(D \cup F) &= P(D) + P(F) - P(D \cap F) = \frac{1}{3} + \frac{5}{9} - \frac{1}{9} = \frac{7}{9} \end{aligned}$$

These results are identical to the results of Problem 23.

PE: Repeat the Problem for the 4 pairs of events in the PE of Problem 23.

26. Give some examples of disjoint events related to the trial of Problem 22.

Answer: It is obvious that the events M, F are disjoint because no member of staff can be male and female. Also, the events D, N are disjoint because no member of staff can be doctor and nurse. All other pairs are not disjoint.[70]

PE: Referring to Problem 22, give examples of events which are:

(**a**) Non-disjoint. (**b**) Non-complementary. (**c**) Disjoint but non-complementary.

27. Referring to Problem 22, if the member of staff that we picked randomly was a male, find the probability of (**a**) being a nurse and (**b**) being a doctor.

Answer:

(**a**) We use the rule of conditional probability (i.e. Eq. 45), that is (see Problem 24 for the employed probabilities):

$$P(N\,|M) = \frac{P(N \cap M)}{P(M)} = \frac{2/9}{4/9} = \frac{1}{2}$$

This is reasonable because we have four males in the staff two of them are nurse and hence if the picked person is male then the probability of being a nurse should be $2/4 = 1/2$.

[69] We note that $P(M \cap D) = 2/9$ and $P(F \cap N) = 4/9$ (see Problem 23). Also, by using the simple count method (which we used in Problem 23) we have $P(M \cap N) = 2/9$ and $P(D \cap F) = 1/9$.

[70] In fact, we have six pairs which we may label compactly as: DN, DM, DF, NM, NF, MF. As we see, only the pair DN and the pair MF are disjoint. However, all this is about pairs; otherwise we have other disjoint events such as being male doctor and being female doctor (or female nurse).

(**b**) We can use the rule of conditional probability (as we did in part a) but if we note that N and D are "complementary" (i.e. in the new sample space which is the sample space made of males) then we can immediately conclude that $P(D\,|M) = 1 - P(N\,|M) = 1/2$. This should also be reasonable because we have four males two of them are doctor and hence if the picked person is male then the probability of being a doctor should be $2/4 = 1/2$.

PE: Referring to Problem 22, if the chosen member of staff was a nurse, find the probability of being a female.

28. Referring to Problem 22, explain the meaning of: $P\Big[(N \cap \overline{M}) \cup (M \cap \overline{N})\Big]$.

    Answer: It means the probability of picking a female nurse or a male doctor.

    PE: Referring to Problem 22, what is the meaning of:

    (**a**) $P\Big[(\overline{D} \cap F) \cup (\overline{F} \cap D)\Big]$. (**b**) $P\Big[(N \cap F) \cup (\overline{F} \cap \overline{N})\Big]$. (**c**) $P\Big[(F \cap D) \cup (M \cap N)\Big]$.

29. Determine the requirement that a pair of events (say A and B) should satisfy to be independent AND disjoint.

    Answer: If A and B are independent then $P(A \cap B) = P(A)\,P(B)$, and if A and B are disjoint then $P(A \cap B) = 0$. On combining these conditions we obtain $P(A)\,P(B) = 0$, i.e. A and B are independent AND disjoint if $P(A) = 0$ or $P(B) = 0$ (i.e. one of them is impossible). In fact, this should indicate that disjoint events are highly dependent (since the occurrence of one means the negation of the other). For instance, if A and B are two disjoint events where $P(A) = 1/2$ and $P(B) = 1/3$ then $P(A\,|B) = 0 \neq P(A) = 1/2$ because $P(A \cap B) = 0$. Similarly, $P(B\,|A) = 0 \neq P(B) = 1/3$ because $P(A \cap B) = 0$. As we see, $P(A\,|B) \neq P(A)$ and $P(B\,|A) \neq P(B)$ which means that these events are dependent. So in brief, (non-trivial) disjoint events must be dependent.

    Note: the dependence of disjoint events can be shown in other ways. For example, if we use the method of Problem 24 for testing the dependence of events then we can write (assuming neither A or B is impossible):

    $$P(A \cap B) = 0 \neq P(A)\,P(B)$$

    i.e. if A and B are disjoint $\big[$as required by $P(A \cap B) = 0\big]$ then they cannot be independent $\big[$because $P(A)\,P(B) \neq 0\big]$ noting that if A and B are independent then we must have $P(A \cap B) = P(A)\,P(B)$.

    PE: Let A and B be two dependent events. List all the possibilities for the relationship between them.

30. Two pregnant women gave birth to two babies (i.e. each woman gave birth to a single child). What is the probability that at least one of the babies is a boy (assuming that the probabilities of having boy and having girl are equal and supposing that "boy" and "girl" are the only possibilities)?

    Answer: We have 4 equal possibilities: BB, BG, GB, and GG (where B stands for boy and G for girl and noting that the order in BG and GB does matter since the position of B and G corresponds to a specific woman). As we see, in 3 of these 4 possibilities we have at least one B and hence the probability that at least one of the babies is a boy is 3/4. We can also argue that "at least one of the babies is a boy" and "both babies are girl" are complementary and since the probability of "both babies are girl" is obviously 1/4 then the probability of "at least one of the babies is a boy" should be $1 - (1/4) = 3/4$ (see Eq. 40 and part a of Problem 14).

    PE: If one woman in this Problem gave birth to a twin (i.e. two babies) while the other gave birth to a single baby, what is the probability that all the babies are girls (assuming equal probabilities and dual possibilities for the gender of babies)?

31. The target in a firing test is a square board of area 0.25m^2 inside which a circle of radius 0.1m is inscribed. To pass the test, the bullet should hit the interior of the circle. Assuming that hitting the square is certain (or the square is hit already) and the probability of hitting any part of it is the same, what is the probability of passing the test?

    Answer: The probability of passing the test should equal the ratio of the area of the circle to the area of the square. The area of the circle is $\pi \times 0.1^2 = 0.01\pi\text{m}^2$. Hence, the probability of passing the test is $0.01\pi/0.25 = \pi/25 \simeq 0.126$.

    PE: Repeat the Problem but replace the square by a circle of radius 0.25m.

32. An integer number between 5 and 50 (inclusive) is chosen randomly. What is the probability of "being prime AND multiple of 3"?

Answer: The event of "being prime AND multiple of 3" is impossible and hence the probability of this is 0.

PE: Repeat the Problem for the probability of "being prime OR multiple of 3".

33. An integer number between 1 and 50 (inclusive) is chosen randomly. What is the probability of "being divisible by 2 OR divisible by 5"?

Answer: Let d_2 and d_5 mean divisible by 2 and divisible by 5 respectively. We have 25 numbers divisible by 2 (i.e. the even numbers $2, 4, \ldots, 50$) and hence $P(d_2) = 25/50 = 1/2$. We have 10 numbers divisible by 5 (i.e. the multiples of five $5, 10, \ldots, 50$) and hence $P(d_5) = 10/50 = 1/5$. We also have 5 numbers divisible by 2 AND divisible by 5 (i.e. the multiples of ten $10, 20, \ldots, 50$) and hence $P(d_2 \cap d_5) = 5/50 = 1/10$. On using the addition law of probability (see Eq. 43) we get:

$$P(d_2 \cup d_5) = P(d_2) + P(d_5) - P(d_2 \cap d_5) = \frac{1}{2} + \frac{1}{5} - \frac{1}{10} = \frac{3}{5}$$

PE: Repeat the Problem for the probability of being:

(**a**) Prime. (**b**) Perfect square. (**c**) Prime OR perfect square.

(**d**) Prime AND perfect square. (**e**) Perfect square AND even. (**f**) Perfect square AND odd.

(**g**) Prime AND odd. (**h**) Prime OR odd. (**i**) Prime AND even.

34. A die and a coin are thrown simultaneously. What is the probability of getting T (on the coin) and 2 (on the die)?

Answer: It is obvious that the outcomes of the die and the coin are independent. Hence, we can use the multiplication law for independent events [noting that $P(T) = 1/2$ and $P(2) = 1/6$ assuming they are fair], that is:

$$P(T \cap 2) = P(T)\, P(2) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}$$

We may also use simple count, i.e. we have 12 equally likely outcomes which are:

$$H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6$$

only one of which (i.e. $T2$) meets the requirement and hence $P(T \cap 2) = 1/12$.

PE: Repeat the Problem for the probability of getting:

(**a**) H (on the coin) and even number (on the die).

(**b**) "H AND odd number" OR "T AND number less than 4".

35. An electronic device requires a transistor whose chance of being defective is 3% and a resistor whose chance of being defective is 1%. The device works only if both these components are perfect. Assuming that all the other components in the device are perfect and the defects of the transistor and resistor are independent, what is the probability of the device being defective.

Answer: If T and R stand for perfect transistor and perfect resistor respectively, then $P(T) = 0.97$ and $P(R) = 0.99$, and hence by the multiplication law for independent events (see Eq. 50) the probability of the device being perfect is:

$$P(T \cap R) = P(T)\, P(R) = 0.97 \times 0.99 = 0.9603$$

Accordingly, the probability of the device being defective is:

$$P(\overline{T \cap R}) = 1 - P(T \cap R) = 1 - 0.9603 = 0.0397$$

where the bar means "defective" (and noting that "perfect" and "defective" are complementary; see Eq. 40 and part a of Problem 14).

We may also use the addition law of probability (see Eq. 43 as well as Eq. 21), that is:

$$P(\overline{T \cap R}) = P(\overline{T} \cup \overline{R}) = P(\overline{T}) + P(\overline{R}) - P(\overline{T} \cap \overline{R}) = 0.03 + 0.01 - (0.03 \times 0.01) = 0.0397$$

PE: Repeat the Problem assuming this time that $P(\overline{T}) = 0.02$ and $P(\overline{R}) = 0.04$.

36. Referring to Problem 27 of § 2.2, what is the probability of any specific distribution (where "specific distribution" means for instance balls 2, 5, 6 are in the 3-ball bag, balls 1, 4, 8, 11, 12 are in the 5-ball bag and the other balls are in the 8-ball bag)?
Answer: According to the result of Problem 27 of § 2.2 we have 720720 possibilities and hence the probability of any specific distribution (i.e. the probability of any one of these possibilities) is 1/720720 (noting that all these possibilities are presumably equally likely).
PE: Repeat the Problem for the PE of Problem 27 of § 2.2.
37. Referring to Problem 32 of § 2.2:
(**a**) What is the probability that the last 2 drawers remain vacant?
(**b**) What is the probability that at least one of the last 2 drawers is filled?
Answer: We assume that all ways of storing are equally likely.
(**a**) The event "the last 2 drawers remain vacant" is equivalent to the event "the shirts are stored in the first 5 drawers". Now, we have 5! possibilities for filling the first 5 drawers (i.e. the first drawer can store any one of the 5 shirts, the second drawer can store any one of the remaining 4 shirts, and so on), which means that we have 5! possibilities for the last 2 drawers being vacant. Also, from Problem 32 of § 2.2 we know that the total number of possibilities for storing these shirts is 2520. Hence, the probability that the last 2 drawers remain vacant is $5!/2520 \simeq 0.0476$.
We may also argue that "the last 2 drawers remain vacant" is just one possibility of the C_5^7 (equally likely) possibilities of selecting 5 drawers (out of 7) for storage and hence the probability that the last 2 drawers remain vacant is $1/C_5^7 = 1/21 \simeq 0.0476$.
(**b**) The event "the last 2 drawers remain vacant" and the event "at least one of the last 2 drawers is filled" are complementary. Hence, the probability that at least one of the last 2 drawers is filled is $1 - (5!/2520) \simeq 0.9524$.
PE: Referring to the PE of Problem 32 of § 2.2, what is the probability that the second, fourth and ninth garages remain vacant?
38. Referring to Problem 37 of § 2.2, a byte is selected randomly. What is the probability that the sum of its bits is 3?
Answer: A byte contains 8 bits each of which can be 0 or 1 and hence by the fundamental principle of counting there are $2^8 = 256$ different bytes. Also, from the method of Problem 37 of § 2.2 we know that we have $C_3^8 = 56$ different bytes that contain exactly three 1's (and hence the sum of their bits is 3). Therefore, the required probability is $56/256 \simeq 0.2188$.
PE: Repeat the Problem to find the probability that the sum is greater than 5.
39. Show that if we distribute n balls in n bags randomly (where there is no limit on the capacity or storage of bags), then for relatively large n the probability of each bag containing 1 ball is approximately $e^{-n}\sqrt{2\pi n}$.
Answer: Referring to part (a) of Problem 34 of § 2.2, balls are like distinguishable particles and bags are like states (with no limit on their occupancy) and hence we can use the Maxwell-Boltzmann statistics. Accordingly, we have n^n possibilities for the distribution of n balls in n bags. Moreover, we have $n!$ possibilities for distributing n balls in n bags such that each bag contains 1 ball (since the 1^{st} bag can contain any one of the n balls, the 2^{nd} bag can contain any one of the remaining $n-1$ balls and so on, and hence we have $n!$ possibilities). Therefore, the probability of each bag containing 1 ball is:

$$\frac{n!}{n^n} \simeq e^{-n}\sqrt{2\pi n} \qquad \text{(large } n\text{)}$$

where we used in making this approximation the Stirling formula for approximating factorials for large n (see Eq. 37). In fact, the last equation gives a good approximation even for small n. To demonstrate the accuracy of this approximation we present in Table 1 a sample of values that n can take and compare the exact value of probability (i.e. $n!/n^n$) with its approximate value (i.e. $e^{-n}\sqrt{2\pi n}$).
PE: Build a spreadsheet in which you calculate $n!/n^n$ and $e^{-n}\sqrt{2\pi n}$ for a range of values of n (e.g. $n = 2$ to $n = 100$) and compare the two by plotting them (or their ratio or relative difference) on a chart.

Table 1: The table of Problem 39 of § 3.2.

	n				
	2	**5**	**10**	**15**	**20**
$\boldsymbol{n!/n^n}$	0.5	0.03840	3.629×10^{-4}	2.986×10^{-6}	2.320×10^{-8}
$\boldsymbol{e^{-n}\sqrt{2\pi n}}$	0.47975	0.03777	3.599×10^{-4}	2.970×10^{-6}	2.311×10^{-8}

3.3 Extensions and Generalizations

Most of the relationships and laws of probability involving two events (which we investigated and stated earlier in § 3.2) can be extended and generalized to more than two events. In this section we present a number of these extensions and generalizations.

The **addition law of probability for three events** A, B, C is given by:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) \tag{51}$$

By induction, this formula can be extended to n events, that is:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum_{i \neq j} P(A_i \cap A_j) + \\ &\qquad \sum_{i \neq j \neq k} P(A_i \cap A_j \cap A_k) - \cdots (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n) \end{aligned} \tag{52}$$

where $\cup_{i=1}^{n} A_i$ means the union of all the A_i's (i.e. $\cup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \ldots \cup A_n$), and all indices run over n.[71] If A_i are **pairwise exclusive events**, the last equation (i.e. Eq. 52) becomes:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) \qquad (A_j \cap A_k = \varnothing,\ j \in n,\ k \in n,\ j \neq k) \tag{53}$$

The **multiplication law of probability for three events** A, B, C is given by:

$$P(A \cap B \cap C) = P(A)\ P(B\,|A)\ P\Big[C\,|(A \cap B)\Big] \tag{54}$$

This equation can be easily extended to more than three events (by noting the pattern of Eq. 54), that is:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)\ P(A_2\,|A_1)\ P\Big[A_3\,|(A_1 \cap A_2)\Big] \cdots P\Big[A_n\,|(A_1 \cap A_2 \cap \cdots \cap A_{n-1})\Big] \tag{55}$$

The **multiplication law of probability for three mutually independent events** A, B, C is given by:

$$P(A \cap B \cap C) = P(A)\,P(B)\,P(C) \qquad (A, B, C \text{ are mutually independent}) \tag{56}$$

This formula can be extended to n mutually independent events, that is:

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i) \qquad (A_i \text{ are mutually independent}) \tag{57}$$

where $\cap_{i=1}^{n} A_i$ means the intersection of all the A_i's (i.e. $\cap_{i=1}^{n} A_i = A_1 \cap A_2 \cap \ldots \cap A_n$).[72]

[71] We note that $(-1)^{n-1}$ in the last term of Eq. 52 is specific to the last term. The sign of each sum of terms is $(-1)^{m-1}$ where m is the number of events considered in the probabilities of that sum, i.e. 1 event for $\sum_{i=1}^{n} P(A_i)$, 2 events for $\sum_{i \neq j} P(A_i \cap A_j)$, and so on.

[72] We note that Eqs. 56 and 57 should be seen as part of the definition of mutual independence (see Problems 1 and 2).

If event A is a union of n pairwise exclusive events and B is another event (in the same sample space) then from Eq. 53 we have:

$$P(A \cap B) = \sum_{i=1}^{n} P(A_i \cap B) \tag{58}$$

Moreover, if A represents the entire sample space then from Eq. 58 we get:

$$P(B) = P(A \cap B) = \sum_{i=1}^{n} P(A_i \cap B) \tag{59}$$

where we used Eq. 7 in the first equality (noting that A is the universal set since it represents the entire sample space).

Problems

1. Distinguish between pairwise independence and mutual independence of a number of events (i.e. ≥ 3 events).
   Answer: We describe a collection of events as **pairwise independent** if any two events in the collection are independent, and we describe the collection as **mutually (or jointly) independent** if any number of events in the collection are independent. For example, if A, B, C are three events then A, B, C are pairwise independent if:

   $$P(A \cap B) = P(A)\,P(B) \qquad \& \qquad P(A \cap C) = P(A)\,P(C) \qquad \& \qquad P(B \cap C) = P(B)\,P(C)$$

   However, A, B, C are mutually independent if in addition to these three conditions we also have:

   $$P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$$

   Accordingly, mutual independence is stronger than pairwise independence, i.e. if a number of events are mutually independent then they are pairwise independent but if they are pairwise independent then they are not necessarily mutually independent.
   Note: pairwise independence of random events A_i ($i = 1, 2, \ldots, n$ where n can be infinite) is expressed mathematically by the condition:

   $$P(A_i \cap A_j) = P(A_i)\,P(A_j) \qquad\qquad (i, j = 1, 2, \ldots, n \text{ and } i \neq j)$$

   while mutual independence of random events A_i is expressed mathematically by the combination of conditions:[73]

   $$\begin{aligned} P(A_i \cap A_j) &= P(A_i)\,P(A_j) && (i, j = 1, 2, \ldots, n \text{ and } i \neq j) \\ P(A_i \cap A_j \cap A_k) &= P(A_i)\,P(A_j)\,P(A_k) && (i, j, k = 1, 2, \ldots, n \text{ and } i \neq j \neq k) \\ &\vdots \\ P\left(\bigcap_{i=1}^{n} A_i\right) &= \prod_{i=1}^{n} P(A_i) \end{aligned}$$

   PE: A, B, C, D are four events. Write the conditions for their (**a**) pairwise independence and (**b**) mutual independence.
2. Give examples for the following:
   (**a**) Events which are not pairwise independent (and hence they are not mutually independent).
   (**b**) Events which are pairwise independent but not mutually independent.
   (**c**) Events which are mutually independent (and hence they are pairwise independent).

[73] To put it in plain words, this combination of conditions means: a number of events are mutually independent *iff* every combination of these events (involving any number of these events) is independent.

Answer: Let A, B, C be three events.
(**a**) If we draw a card (randomly) from a pack of 6 cards (numbered $1, 2, 3, 4, 5, 6$) where:

$$A = \{1, 2, 3\} \qquad B = \{3, 4, 5\} \qquad C = \{1, 5, 6\}$$

then we have:

$$\begin{aligned} P(A)\,P(B) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq \frac{1}{6} = P(A \cap B) \\ P(A)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq \frac{1}{6} = P(A \cap C) \\ P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \neq \frac{1}{6} = P(B \cap C) \\ P(A)\,P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \neq 0 = P(A \cap B \cap C) \end{aligned}$$

i.e. A, B, C are neither pairwise independent nor mutually independent.
(**b**) If we draw a card (randomly) from a pack of 12 cards (numbered $1, 2, \ldots, 12$) where:

$$A = \{1, 2, 3, 4, 5, 6\} \qquad B = \{4, 5, 6, 7, 8, 9\} \qquad C = \{1, 2, 3, 7, 8, 9\}$$

then we have:

$$\begin{aligned} P(A)\,P(B) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(A \cap B) \\ P(A)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(A \cap C) \\ P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(B \cap C) \\ P(A)\,P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \neq 0 = P(A \cap B \cap C) \end{aligned}$$

i.e. A, B, C are pairwise independent but not mutually independent.
(**c**) If we draw a card (randomly) from a pack of 16 cards (numbered $1, 2, \ldots, 16$) where:

$$A = \{1, 2, 3, 4, 5, 6, 7, 8\} \qquad B = \{5, 6, 7, 8, 9, 10, 11, 12\} \qquad C = \{3, 4, 7, 8, 9, 10, 13, 14\}$$

then we have:

$$\begin{aligned} P(A)\,P(B) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(A \cap B) \\ P(A)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(A \cap C) \\ P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} = P(B \cap C) \\ P(A)\,P(B)\,P(C) &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} = P(A \cap B \cap C) \end{aligned}$$

i.e. A, B, C are pairwise independent and mutually independent.
Note: as explained above, to have mutual independence we need to have pairwise independence (as well as higher-multiplicity independence). In other words, higher-multiplicity independence does not imply or guarantee pairwise independence and hence we need to verify pairwise independence independently of verifying higher-multiplicity independence. For example, if A, B, C are three events in a trial of drawing a card (randomly) from a pack of 16 cards (numbered $1, 2, \ldots, 16$) where:

$$A = \{1, 2, 3, 4, 5, 6, 7, 8\} \qquad B = \{7, 8, 9, 10, 11, 12, 13, 14\} \qquad C = \{6, 7, 8, 9, 10, 14, 15, 16\}$$

then we have:

$$\begin{aligned} P(A)\,P(B) &= \frac{1}{2}\times\frac{1}{2}=\frac{1}{4}\neq\frac{1}{8}=P(A\cap B)\\ P(A)\,P(C) &= \frac{1}{2}\times\frac{1}{2}=\frac{1}{4}\neq\frac{3}{16}=P(A\cap C)\\ P(B)\,P(C) &= \frac{1}{2}\times\frac{1}{2}=\frac{1}{4}\neq\frac{5}{16}=P(B\cap C)\\ P(A)\,P(B)\,P(C) &= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=\frac{1}{8}=P(A\cap B\cap C) \end{aligned}$$

As we see, although the condition $P(A\cap B\cap C)=P(A)\,P(B)\,P(C)$ is satisfied, the events A, B, C are not mutually independent because they are not pairwise independent. So in brief, the condition $P(A\cap B\cap C)=P(A)\,P(B)\,P(C)$ does not guarantee (or imply) pairwise independence and hence it is not sufficient (although it is necessary) for mutual independence. This similarly applies to more than three events where higher-multiplicity independence does not guarantee pairwise or lower-multiplicity independence.

This example should also show that the condition of mutual independence in Eqs. 56 and 57 is stronger than we need for the validity of these equations; in other words it is sufficient but not necessary. We also note that this example, in addition to the example of part (b) of this Problem, should show that pairwise independence (in itself) is neither sufficient nor necessary for the validity of Eqs. 56 and 57 (even though mutual independence, which implies pairwise independence, was imposed in these equations to ensure their validity).

PE: Repeat this Problem by giving other examples for the three parts as well as an example similar to the example given in the note.

3. Verify Eq. 54 for the events of:

(**a**) Part (a) of Problem 2.

(**b**) Part (b) of Problem 2.

(**c**) Part (c) of Problem 2.

(**d**) The note of Problem 2.

Answer:

(**a**)

$$\begin{aligned} P(A\cap B\cap C)=0 &= \frac{1}{2}\times\frac{1}{3}\times 0=P(A)\;P(B\,|A)\;P\Big[C\,|(A\cap B)\Big]\\ =0 &= \frac{1}{2}\times\frac{1}{3}\times 0=P(B)\;P(C\,|B)\;P\Big[A\,|(B\cap C)\Big]\\ =0 &= \frac{1}{2}\times\frac{1}{3}\times 0=P(C)\;P(A\,|C)\;P\Big[B\,|(C\cap A)\Big] \end{aligned}$$

(**b**)

$$\begin{aligned} P(A\cap B\cap C)=0 &= \frac{1}{2}\times\frac{1}{2}\times 0=P(A)\;P(B\,|A)\;P\Big[C\,|(A\cap B)\Big]\\ =0 &= \frac{1}{2}\times\frac{1}{2}\times 0=P(B)\;P(C\,|B)\;P\Big[A\,|(B\cap C)\Big]\\ =0 &= \frac{1}{2}\times\frac{1}{2}\times 0=P(C)\;P(A\,|C)\;P\Big[B\,|(C\cap A)\Big] \end{aligned}$$

(**c**)

$$\begin{aligned} P(A\cap B\cap C)=\frac{1}{8} &= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=P(A)\;P(B\,|A)\;P\Big[C\,|(A\cap B)\Big]\\ =\frac{1}{8} &= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=P(B)\;P(C\,|B)\;P\Big[A\,|(B\cap C)\Big]\\ =\frac{1}{8} &= \frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}=P(C)\;P(A\,|C)\;P\Big[B\,|(C\cap A)\Big] \end{aligned}$$

(**d**)

$$\begin{aligned} P(A\cap B\cap C)=\frac{1}{8} &= \frac{1}{2}\times\frac{1}{4}\times 1= P(A)\ P(B\,|A)\ P\Big[C\,|(A\cap B)\Big] \\ =\frac{1}{8} &= \frac{1}{2}\times\frac{5}{8}\times\frac{2}{5}= P(B)\ P(C\,|B)\ P\Big[A\,|(B\cap C)\Big] \\ =\frac{1}{8} &= \frac{1}{2}\times\frac{3}{8}\times\frac{2}{3}= P(C)\ P(A\,|C)\ P\Big[B\,|(C\cap A)\Big] \end{aligned}$$

PE: Repeat this Problem for the examples of the PE of Problem 2.

4. What you notice from the results of Problems 2 and 3?
   Answer: We notice that Eq. 54 is always true but Eq. 56 is not always true because it is conditioned by mutual independence (noting that "always" here is within the investigated cases although we know from other evidence that it is true in general).
   PE: Do you recommend using Eq. 54 in all cases (i.e. of three events and its alike for more than three events; see Eq. 55) especially when there is some confusion about the nature of the events with regard to their independence?
5. Distinguish between mutual exclusivity and pairwise exclusivity.
   Answer: Let define mutual exclusivity on the style of mutual independence, that is: a collection of events (i.e. ≥ 3 events) are mutually exclusive if any number of events in the collection are exclusive (i.e. their intersection is empty and hence the probability of the intersection is zero). Accordingly, if a number of events are mutually exclusive then they should be pairwise exclusive. We also note that if a number of events are pairwise exclusive then they should be mutually exclusive.
   PE: Why if a number of events are pairwise exclusive then they should be mutually exclusive?
6. Referring to Problem 5, can we say: mutual exclusivity and pairwise exclusivity are equivalent?
   Answer: Yes, we can say this but at the practical level (i.e. if a collection of events are pairwise/mutually exclusive then they should be mutually/pairwise exclusive). However, at the conceptual level they are not equivalent because mutual exclusivity is a stronger (or richer) condition than pairwise exclusivity since it embeds multiple-events exclusivity (in addition to pairwise exclusivity). To clarify the idea further, let define *inclusivity* as the opposite of exclusivity. Accordingly, both:

$$A=\{1,2,3\} \qquad B=\{3,4,5\} \qquad C=\{1,5,6\}$$

   and

$$D=\{1,2,3,7\} \qquad E=\{3,4,5,7\} \qquad F=\{1,5,6,7\}$$

   are pairwise *inclusive*, but only D, E, F are mutually *inclusive* because $D\cap E\cap F=\{7\}\neq\varnothing$ while $A\cap B\cap C=\varnothing$ (which is an extra condition imposed by mutual exclusivity but not by pairwise exclusivity).
   PE: Compare pairwise/mutual exclusivity with pairwise/mutual independence and try to justify any difference.
7. Prove the addition law of probability for three events A, B, C (i.e. Eq. 51).
   Answer:

$$\begin{aligned} P(A\cup B\cup C) &= P\Big[A\cup(B\cup C)\Big] \\ &= P(A)+P(B\cup C)-P\Big[A\cap(B\cup C)\Big] \\ &= P(A)+P(B)+P(C)-P(B\cap C)-P\Big[A\cap(B\cup C)\Big] \\ &= P(A)+P(B)+P(C)-P(B\cap C)-P\Big[(A\cap B)\cup(A\cap C)\Big] \\ &= P(A)+P(B)+P(C)-P(B\cap C)-P(A\cap B)-P(A\cap C) \\ &\qquad +P\Big[(A\cap B)\cap(A\cap C)\Big] \\ &= P(A)+P(B)+P(C)-P(B\cap C)-P(A\cap B)-P(A\cap C)+P(A\cap B\cap A\cap C) \end{aligned}$$

$$= \quad P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

where in equality 1 we use Eq. 18, in equality 2 we apply Eq. 43 on $P\Big[A \cup (B \cup C)\Big]$,[74] in equality 3 we apply Eq. 43 on $P(B \cup C)$, in equality 4 we apply Eq. 19, in equality 5 we apply Eq. 43 on $P\Big[(A \cap B) \cup (A \cap C)\Big]$, in equality 6 we apply Eq. 17, and in equality 7 we apply Eq. 5 (in association with Eq. 14).

PE: Derive the addition law of probability for four events A, B, C, D.

8. Prove the multiplication law of probability for three events A, B, C (i.e. Eq. 54).
Answer:

$$\begin{aligned} P(A \cap B \cap C) &= P\Big[(A \cap B) \cap C\Big] && \text{(Eq. 17)} \\ &= P(A \cap B)\ P\Big[C\,|(A \cap B)\Big] && \text{(Eq. 47)} \\ &= P(A)\ P(B\,|A)\ P\Big[C\,|(A \cap B)\Big] && \text{(Eq. 47)} \end{aligned}$$

PE: Can we write $P\Big[(B \cap C)\,|A\Big] = P(B\,|A)\ P\Big[C\,|(A \cap B)\Big]$?

9. Derive the multiplication law of probability for four events A, B, C, D.
Answer:

$$\begin{aligned} P(A \cap B \cap C \cap D) &= P\Big[(A \cap B \cap C) \cap D\Big] && \text{(Eq. 17)} \\ &= P(A \cap B \cap C)\ P\Big[D\,|(A \cap B \cap C)\Big] && \text{(Eq. 47)} \\ &= P(A)\ P(B\,|A)\ P\Big[C\,|(A \cap B)\Big]\ P\Big[D\,|(A \cap B \cap C)\Big] && \text{(Eq. 54)} \end{aligned}$$

PE: Derive the multiplication law of probability for five events A, B, C, D, E.

10. In a class made of 50 students, 20% have a height of less than 160cm, 64% a height between 160cm-180cm, and 16% have a height greater than 180cm. If 5 students are selected at random from this class, what is the probability of all 5 being less than 160cm tall?
Answer: We have C_5^{10} combinations for selecting 5 students out of the 10 students whose height is less than 160cm, and we have C_5^{50} combinations for selecting 5 students out of the 50 students. Hence:

$$P(\text{height of all } 5 < 160\text{cm}) = \frac{C_5^{10}}{C_5^{50}} = \frac{252}{2118760} \simeq 0.0001189$$

We may also argue (differently) that if we select the 5 students successively then the probabilities of the $1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}$ student to have a height less than 160cm are $(10/50), (9/49), (8/48), (7/47), (6/46)$ and hence by the multiplication law of probability (for dependent events)[75] we have:

$$P(\text{height of all } 5 < 160\text{cm}) = \frac{10}{50} \times \frac{9}{49} \times \frac{8}{48} \times \frac{7}{47} \times \frac{6}{46} \simeq 0.0001189$$

PE: Find the probability of:
(**a**) Exactly 1 student of the selected 5 being less than 160cm tall, and exactly 1 student of the selected 5 being greater than 180cm tall.
(**b**) Exactly 3 students of the selected 5 being greater than 180cm tall.

11. Let U be the set of lower case English alphabet, $A = \{a, b, d, k, r, t, z\}$, $B = \{a, c, d, m, r, t, w, x\}$, and $C = \{a, m, n, q, r, s, t\}$. If a lower case English letter is selected randomly, verify Eq. 51 for $P(A \cup B \cup C)$ in this case.
Answer: The union $A \cup B \cup C = \{a, b, c, d, k, m, n, q, r, s, t, w, x, z\}$ contains 14 elements. Moreover,

[74] We note that Eq. 43 was derived in part (e) of Problem 14 of § 3.2.
[75] In fact, we are using Eq. 55 which is an extended form of Eq. 54.

we have 26 lower case English letters. Hence, by simple count we have $P(A \cup B \cup C) = 14/26 = 7/13$. Regarding Eq. 51 we have [following a simple count approach as we did already with $P(A \cup B \cup C)$]:

$P(A) = 7/26$ $P(B) = 8/26$ $P(C) = 7/26$ $P(A \cap B) = 4/26$

$P(A \cap C) = 3/26$ $P(B \cap C) = 4/26$ $P(A \cap B \cap C) = 3/26$

Hence, from Eq. 51 we get:

$$P(A \cup B \cup C) = \frac{7}{26} + \frac{8}{26} + \frac{7}{26} - \frac{4}{26} - \frac{3}{26} - \frac{4}{26} + \frac{3}{26} = \frac{14}{26} = \frac{7}{13}$$

which is identical to what we got earlier by simple count. So, Eq. 51 for $P(A \cup B \cup C)$ is verified in this case.

PE: Do the following:

(**a**) Create and solve a Problem similar to this using other U (such as a set of numbers, or a set of plane geometric shapes, or a set of chemical elements in the periodic table).

(**b**) If $D = \{a, b, f, j, k, p, t, w, y, z\}$ (and A, B, C are as above), verify Eq. 52 for $P(A \cup B \cup C \cup D)$.

12. What is the probability that a group of n persons have different birthdays (i.e. no two of them have the same birthday)? Also find the minimum n for which this probability is less than 2/3.

Answer: We assume the following:

• The year has 365 days.

• The birthdays of the group of people in question are totally independent of each other (e.g. we do not have twins).

• Any person in the group has equal probability to be born in any day of the year and hence the probability of him being born in a specific day of the year is 1/365 (e.g. we do not have people selected on the basis of their birth date which favors or disfavors certain birthdays such as those in the summer or spring).

Now, if we select one person from the group randomly then there is no restriction on his birthday. If we repeat the selection then the second person has a 1/365 probability of being born in the day of the birthday of the first person and hence the probability that the second person has a birthday different from the birthday of the first person is $1 - (1/365)$. Similarly, the third person has a probability of $1 - (2/365)$ that he has a birthday different from the first two persons. So, in general the n^{th} person has a probability of $1 - \frac{n-1}{365}$ that he has a birthday different from the $n-1$ previously-selected persons. Accordingly, by the law of multiplication (for dependent events),[76] the probability P_{dbn} that a group of n persons have different birthdays is:

$$P_{dbn} = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{n-1}{365}\right) = \prod_{k=2}^{n}\left(1 - \frac{k-1}{365}\right) \tag{60}$$

We note that for $n > 365$ this probability becomes 0 according to this formula (as it should be).

Regarding the minimum n for which this probability is less than 2/3, it can be found simply by trial using for instance a spreadsheet. Following this method we found that the minimum n is 18.

Note: starting from Eq. 60, we can obtain another (and rather simpler) expression for P_{dbn}, that is:

$$\begin{aligned} P_{dbn} &= \prod_{k=2}^{n}\left(1 - \frac{k-1}{365}\right) = \prod_{k=1}^{n}\left(1 - \frac{k-1}{365}\right) = \prod_{k=1}^{n}\left(\frac{365-k+1}{365}\right) \\ &= \frac{1}{365^n}\prod_{k=1}^{n}(365-k+1) = \frac{1}{365^n}P_n^{365} = \frac{365!}{365^n\,(365-n)!} \end{aligned} \tag{61}$$

PE: Do the following:

• Create a spreadsheet to find the minimum n for which this probability is less than 1/2.

• Repeat this Problem for different *birthmonths* (instead of birthdays).

[76] As before, we are using Eq. 55 which is an extended form of Eq. 54.

- Find the probability that in a group of n persons at least two persons have the same birthday.
- Make an argument to derive Eq. 61 directly rather than obtaining it from Eq. 60.

13. A box contains 10 balls: 4 black and 6 white. If 3 balls are drawn randomly, what is the probability of being all black?
Answer: We have C_3^4 combinations for drawing 3 black balls (out of 4 black balls), and C_3^{10} combinations for drawing 3 balls (out of 10 balls). Hence:

$$P(\text{all 3 balls are black}) = \frac{C_3^4}{C_3^{10}} = \frac{4}{120} = \frac{1}{30}$$

Alternatively, if we select the 3 balls successively then the probabilities of the $1^{st}, 2^{nd}, 3^{rd}$ ball to be black are $(4/10), (3/9), (2/8)$ and hence by the multiplication law of probability (for dependent events using Eq. 54) we have:

$$P(\text{all 3 balls are black}) = \frac{4}{10} \times \frac{3}{9} \times \frac{2}{8} = \frac{1}{30}$$

PE: Repeat the Problem for drawing 4 balls which are:
(**a**) All white. (**b**) All black. (**c**) 2 white and 2 black. (**d**) 1 white and 3 black.

14. Let event A be a union of n pairwise exclusive events $(A_1, A_2, \ldots, A_n)$ and B and C are other events in the same sample space. Prove the following:

(**a**) $P(A\,|B) = \sum_{i=1}^{n} P(A_i\,|B)$

(**b**) $P(B) = \sum_{i=1}^{n} P(A_i)\,P(B\,|A_i)$ (A represents the entire sample space)

(**c**) $P(B\,|C) = \sum_{i=1}^{n} P(A_i\,|C)\,P\Big[B\,|(A_i \cap C)\Big]$ (A represents the entire sample space)

Answer:
(**a**)

$$P(A \cap B) = \sum_{i=1}^{n} P(A_i \cap B) \qquad \text{(Eq. 58)}$$

$$\frac{P(A \cap B)}{P(B)} = \sum_{i=1}^{n} \frac{P(A_i \cap B)}{P(B)} \qquad \Big[\text{dividing by } P(B) \neq 0\Big]$$

$$P(A\,|B) = \sum_{i=1}^{n} P(A_i\,|B) \qquad \text{(Eq. 45)}$$

We note that this equation is called the **addition law for conditional probabilities**.
(**b**)

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) \qquad \text{(Eq. 59)}$$

$$P(B) = \sum_{i=1}^{n} P(A_i)\,P(B\,|A_i) \qquad \text{(Eq. 47)} \qquad (62)$$

(**c**)

$$P(B \cap C) = \sum_{i=1}^{n} P\Big[A_i \cap (B \cap C)\Big] \qquad \text{(Eq. 59)}$$

$$P(B \cap C) = \sum_{i=1}^{n} P\Big[A_i \cap B \cap C\Big] \qquad \text{(Eq. 17)}$$

$$P(B \cap C) = \sum_{i=1}^{n} P(C)\, P(A_i\,|C)\, P\Big[B\,|(A_i \cap C)\Big] \qquad \text{(Eq. 54)}$$

$$\frac{P(B \cap C)}{P(C)} = \sum_{i=1}^{n} P(A_i\,|C)\, P\Big[B\,|(A_i \cap C)\Big] \qquad \Big[\text{dividing by } P(C) \neq 0\Big]$$

$$P(B\,|C) = \sum_{i=1}^{n} P(A_i\,|C)\, P\Big[B\,|(A_i \cap C)\Big] \qquad \text{(Eq. 45)}$$

PE: What is the significance of the results of this Problem? Give some real-life examples of these results.

3.4 Abstract Devices in Probability

There are several types of abstract devices and techniques which are commonly used in tackling and solving probability problems. These devices are used for various purposes like explanation, illustration, proving, testing and verification. These include **Venn diagrams** and **tree diagrams** which are widely used for explanation and illustration and can be crucial for setting, clarifying and formulating the problems.[77] **Tables** and **graphs** (e.g. **histograms**) may also be used for these purposes. **Computer simulation** can also be used for verifying the solutions obtained analytically or by other methods (see § 1.6).[78] In the following Problems we demonstrate the use of these devices in probability.

Problems

1. Draw a Venn diagram representing the case of Problem 22 of § 3.2.
   Answer: See Figure 13.
   PE: Referring to part (b) of Problem 9 of § 3.2, draw a Venn diagram for the sample space with A being the set of "sum of the dice is greater than 8" and B being the set of "at least one of the dice is 5".
2. Make a tree diagram for part (c) of Problem 6 of § 3.2.
   Answer: See Figure 14.
   PE: Make a tree diagram for the case of Problem 22 of § 3.2 (noting that you have more than one possibility).
3. Create a simple table that represents the case of Problem 22 of § 3.2 and can be used to answer the questions related to that Problem.
   Answer: See Table 2.

Table 2: The table of Problem 3 of § 3.4.

	Male	**Female**	**Sum**
Doctor	2	1	3
Nurse	2	4	6
Sum	4	5	9

PE: Make a simple table displaying the possibilities of the outcome of a trial in which a coin and a die are thrown simultaneously (noting that the sample space of throwing a coin is $\{H, T\}$ and the sample space of throwing a die is $\{1, 2, 3, 4, 5, 6\}$).

[77] We note that Venn diagrams and tree diagrams were introduced in § 2.4.

[78] We note that simulation is used in several parts of this book including the present section. We also note that simulation may be used for purposes other than verification (e.g. for solving problems initially).

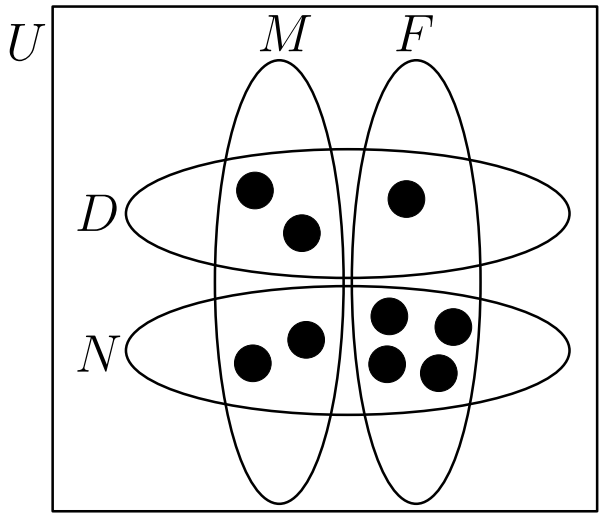


Figure 13: The Venn diagram of Problem 1 of § 3.4 (noting that D, N, M, F stand for doctor, nurse, male, female and each filled circle represents a member of medical staff).

Table 3: The table of Problem 4 of § 3.4.

11	12	13	14	15	16
21	22	23	24	25	(**26**)
31	32	33	34	(**35**)	36
41	42	43	(**44**)	45	46
51	52	(**53**)	54	55	56
61	(**62**)	63	64	65	66

4. Make and use a table to find the probability of getting a sum of 8 in a trial of rolling two (fair and independent) dice.
   Answer: Referring to Table 3 (noting that "25" for instance means getting 2 on the first die and 5 on the second die), we note that out of 36 entries in the table we have 5 entries whose sum is 8 (i.e. the bold and parenthesized ones). Now, if we note that the dice are fair and independent (and hence all the possibilities represented by these entries are equally likely), then we can conclude (using Eq. 38) that the probability of getting a sum of 8 in this trial is 5/36.
   PE: Use Table 3 to find the probability of:
   (**a**) Getting a double, e.g. 22.
   (**b**) Getting odd number on both dice.
   (**c**) Getting a sum of less than 3 or greater than 8.
   (**d**) Getting a prime sum.
   (**e**) Getting at least one prime number on the dice.
   (**f**) Getting a sum which is even.
   (**g**) Getting an odd number on one die and an even number on the other.
5. Use Table 3 to create a new table showing a sample space whose points represent the possible sums of the numbers on the two dice with their probabilities.
   Answer: The sums (which range between 2 and 12) and their probabilities can be obtained from the diagonals (from upper right to lower left) of Table 3. For instance, the sum of 4 is represented by the diagonal made of the elements 13, 22, and 31 and hence this sum has a probability of 3 (possibilities)

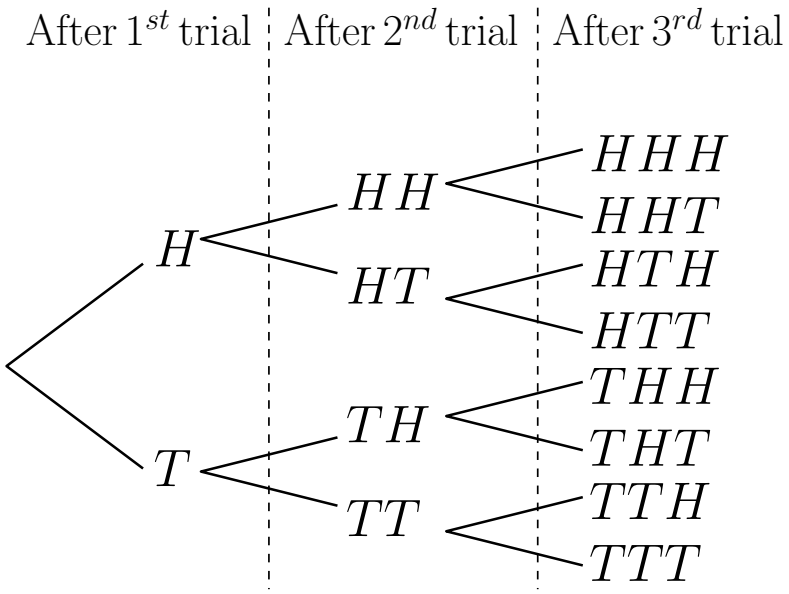


Figure 14: The tree diagram of Problem 2 of § 3.4.

Table 4: The table of Problem 5 of § 3.4.

Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

out of 36 (possibilities) and thus its probability is 3/36. The results are given in Table 4.
Note: in Problem 4 we have an example of uniform sample space and in the present Problem we have an example of non-uniform sample space (see § 3.2).
PE: Use Table 3 to create a new table (similar to Table 4) showing a sample space whose points represent the following:
(**a**) Possible sums s: $s < 4$, $4 \leq s \leq 6$, $7 \leq s \leq 10$, and $s > 10$.[79]
(**b**) Possible products α: $\alpha < 6$, $6 \leq \alpha \leq 13$, $14 \leq \alpha \leq 24$, and $\alpha > 24$.

6. Repeat Problem 5 but this time create a graph (histogram).
Answer: See Figure 15.
PE: Create graphs (histograms) for parts (a) and (b) of the PE of Problem 5.
7. Find the probability of getting a sum of 7 or an odd number in a trial of rolling two (fair and independent) dice (where "odd number" means a 2-digit number made of combining the numbers of the two dice, e.g. if the first die is 1 and the second is 3 then we have 13).
Answer: This Problem can be easily solved by a glimpse to Table 3 where we see 18 odd numbers (i.e. $11, 13, \ldots, 65$) and 3 even numbers whose digits add up to 7 (i.e. 16, 34 and 52) and hence we have 21 possibilities (out of 36 possibilities) that meet the requirements.[80] Therefore, the probability of getting a sum of 7 or an odd number is $21/36 = 7/12$.
As a check, let use the addition law of probability (where S_7 stands for sum of 7 and O stands for

[79] Investigate the possibility of using Table 4.
[80] We note that "odd numbers" and "even numbers whose digits add up to 7" are disjoint. In fact, we are effectively using Eq. 44.

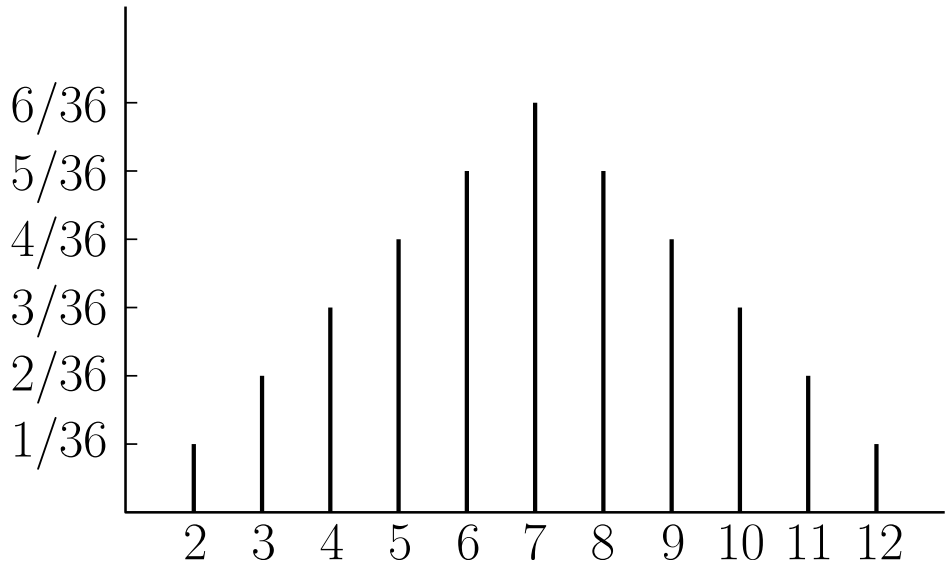


Figure 15: The histogram of Problem 6 of § 3.4. The horizontal axis represents the sums while the vertical axis represents their probabilities.

odd), that is (see Eq. 43):

$$P(S_7 \cup O) = P(S_7) + P(O) - P(S_7 \cap O) = \frac{6}{36} + \frac{18}{36} - \frac{3}{36} = \frac{21}{36} = \frac{7}{12}$$

PE: Repeat the Problem for the following:

(**a**) Getting a sum of 10 or a perfect square. (**b**) Getting a product of 12 or a prime number.

8. Mention some advantages for the use of tables and graphs in presenting data and information in probability theory.
   Answer: Tables and graphs present data and information in a visual and compact form and hence they provide effective and efficient methods for understanding and communicating the problems. For example, if we want to find the probability of having a sum between 4 and 6 in the trial of rolling two (fair and independent) dice then by a glimpse to Table 4 or Figure 15 we can easily find this probability by summing the corresponding probabilities, i.e. $(3/36) + (4/36) + (5/36) = 12/36 = 1/3$.
   PE: Mention some disadvantages for using tables and graphs. Also, mention some advantages and disadvantages for using Venn and tree diagrams and computer simulation.
9. In a national park there are three types of big cats whose numbers and maturity state are given in Table 5. If one of these big cats is killed randomly (e.g. by poachers), use the table to find the probability of being:
   (**a**) Juvenile leopard. (**b**) Mature lion. (**c**) Mature. (**d**) Juvenile cheetah OR mature leopard.
   Answer: The total number of cats is $156 + 68 + 94 + 179 + 59 + 101 = 657$. We label lion, leopard, cheetah, juvenile, mature with L, D, C, J, M.
   (**a**) $P(J \cap D) = 68/657 \simeq 0.1035$.
   (**b**) $P(M \cap L) = 179/657 \simeq 0.2725$.
   (**c**) $P(M) = (179 + 59 + 101)/657 \simeq 0.5160$.
   (**d**) $P\Big[(J \cap C) \cup (M \cap D)\Big] = (94 + 59)/657 \simeq 0.2329$ (noting that $J \cap C$ and $M \cap D$ are disjoint).
   PE: Find the probability of being:
   (**a**) Not cheetah. (**b**) Lion OR leopard. (**c**) Leopard OR juvenile lion.
   (**d**) Non-cheetah juvenile. (**e**) Mature OR leopard. (**f**) Lion OR juvenile.

Table 5: The table of Problem 9 of § 3.4.

	Lion (L)	**Leopard** (D)	**Cheetah** (C)
Juvenile (J)	156	68	94
Mature (M)	179	59	101

10. Test the result of Problem 13 of § 3.3 by simulation.
    Answer: See Box4B6W.cpp code.
    PE: Modify the Box4B6W.cpp code to simulate the results of the PE of Problem 13 of § 3.3.

Chapter 4
Probability Functions

In this chapter we investigate probability functions of random variables. However, before we go through this investigation we need to introduce some definitions and preliminaries.

Random variable is a variable that can take various values each of which is associated with a given probability. For example, the integers $1, 2, 3, 4, 5, 6$ obtained in a die-throwing trial is a random variable where each of its values is associated with a probability of $1/6$ (assuming the die is fair). Similarly, the weight of newborn babies is a random variable where each of its values is associated with a certain probability (depending on many factors like ethnicity, social and economic status, etc.). Random variable is of two main types: discrete and continuous. **Discrete random variable** is characterized by taking a countable number of values (e.g. the integers $1, 2, 3, 4, 5, 6$ obtained in a die-throwing trial), while **continuous random variable** is characterized by taking a continuous range of values (e.g. the height of students in a physics class).[81] It is worth noting that continuous random variables may be treated as discrete (and vices versa) for various purposes and objectives (e.g. for convenience or for comparison), and this can take several shapes and forms. Instances of such treatment can be found in this book (as well as in the literature of probability in general).

Probability distribution function (or probability distribution for brevity) is a mathematical relation that gives probability as a function of a random variable. It is obvious that probability distributions are real-valued functions (noting that probability is real). As we have discrete and continuous random variables, we have **probability mass function** which is a probability distribution of a discrete random variable and **probability density function** which is a probability distribution of a continuous random variable. Probability distributions of discrete random variables are usually presented graphically by a histogram or a bar graph (see for instance Figure 15), while probability distributions of continuous random variables are usually presented graphically by an ordinary graph (i.e. continuous curve).[82]

Probability distribution can be a function of a **single random variable** and can be a function of **multiple random variables**. However, in this chapter (and in this book in general) we generally deal with probability distributions of a single (real) random variable (noting that the generalization to the probability distributions of multiple random variables is generally straightforward). We should also note that any probability distribution must be normalized to unity since it represents probability.[83]

Cumulative distribution function is a real-valued function of a real-valued random variable that gives the probability of the random variable taking a value less than or equal to a given value. As cumulative distribution functions represent cumulative probability they are monotonically increasing functions (noting that probability is non-negative). As we have discrete and continuous random variables, we have cumulative distributions of discrete random variables and cumulative distributions of continuous random variables.

A **Bernoulli trial** (which may also be called **binomial trial**) is a random experiment that has only two possible outcomes (which may be labeled as success and failure). **Bernoulli trials** (i.e. plural which may also be called **Bernoulli process**) is a number of *mutually-independent* Bernoulli trials in which the probability of the outcomes is *the same* in all trials.

Problems

[81] Although "height" in this example (as well as many other types of continuous variable) is primarily and intrinsically continuous, in practical reality it is discrete due to the limits on accuracy and units of measurement. We also note that random variable may also be mixed, i.e. it is partially discrete and partially continuous (noting that we will not study this type of random variable in this book).

[82] As indicated in the previous footnote, random variables can be mixed and so their distributions.

[83] It should be obvious that normalization here means scaling the sum of all probabilities in the distribution to 1.

1. Propose some possible random variables for the following processes:
   (**a**) Throwing 2 dice once. (**b**) Throwing a coin repeatedly.
   (**c**) Observation of an animal species. (**d**) Observation of a type of stars.
   Answer: Examples of possible random variables are:
   (**a**) The sum (or product) of the two numbers obtained. The difference between the first number and the second number. The quotient of the first number to the second number. The absolute value of the difference between the two numbers.
   (**b**) The number of throws until an H is obtained. The number of throws until an HTH sequence is obtained.
   (**c**) Weight. Height. Age. Sex.[84]
   (**d**) Brightness. Mass. Size. Surface temperature.
   PE: Repeat the Problem for the following processes:
   (**a**) Throwing a die repeatedly. (**b**) Throwing 5 coins together one time.
   (**c**) Observation of late-night travelers. (**d**) Observation of a type of chemical reaction.

2. Discuss briefly the issue of transformation of random variables.
   Answer:[85] This is a big and complicated issue and hence it is out of the scope of this book. However, in the following points we make a few remarks about this issue which we will need in the future (as well as being useful and potentially necessary knowledge in general):
   • A random variable can be transformed deterministically or stochastically to another variable which is generally a random variable too. However, our interest (as well as the interest of other authors in general) is about deterministic transformations using analytical functions such as linear functions. For example, if x is a random variable with a probability distribution function $f(x)$ then the random variable y obtained from x by the linear transformation $y(x) = ax + b$ could be of interest to us and we may need to know its probability distribution function $g(y)$ and how it is related to $f(x)$.
   • If x is a random variable and y is a random variable obtained from x by a given (deterministic) transformation then it is reasonable (with certain conditions) to assign the probabilities of the values of x to the corresponding values of y. So loosely speaking, if x_1 is a given element of x with a probability $P(x_1)$ and y_1 is the image of x_1 (as obtained by the given transformation) then we can write generically $P(y_1) = P(x_1)$. Such an assignment of the probabilities of y (i.e. from the probabilities of x) should (under certain conditions) preserve the characteristics and requirements of probability (such as normalization).
   • Let x be a continuous random variable with a probability distribution (or density) function $f(x)$, and let y be a random variable obtained by a differentiable and strictly-increasing transformation function T and hence we can write $y = y(x)$, i.e. y is a function of x. Noting that T is one-to-one it should have an inverse and hence we can also write $x = x(y)$, i.e. x is also a function of y. Now, if $g(y)$ is the (presumed) probability distribution (or density) function of y then based on the previous point we can sensibly write:

$$\begin{aligned} g(y)\,dy &= f(x)\,dx \\ \text{and hence} \qquad g(y) &= f\big[x(y)\big]\,\frac{dx}{dy} \end{aligned} \tag{63}$$

   where we note that $x(y)$ is differentiable and it is explicitly expressed in terms of y (and hence the right hand side of Eq. 63 is purely in terms of y). This means that we can (under these conditions) obtain $g(y)$ from $f(x)$ with the help of the conditions imposed on x and y and the relation between them.
   As an example, let T be the (strictly-increasing) linear transformation $y = ax + b$ where $a > 0$ and x

[84] Random variables which are not numeric (like sex) should be expressed numerically (e.g. 1 for male and 2 for female).
[85] The readers should be aware that for the sake of clarity we follow in this answer a simplified approach and hence it is not sufficiently rigorous.

(which is a variable over a real interval satisfying the normalization condition) has the density function $f(x)$. So, from Eq. 63 we can write [noting that $x = (y-b)/a$ and $dx/dy = 1/a$]:

$$g(y) = f\big[x(y)\big]\,\frac{dx}{dy} = f\left(\frac{y-b}{a}\right) \times \frac{1}{a} = \frac{1}{a}\,f\left(\frac{y-b}{a}\right)$$

As a second example, let T be the (strictly-increasing) quadratic transformation $y = ax^2$ where $a > 0$, $0 < x < (3/a)^{1/3}$ and x has the density function $f(x)$. So, from Eq. 63 we can write [noting that $x = \sqrt{y/a}$ and $dx/dy = 1/\sqrt{4ay}$]:

$$g(y) = f\big[x(y)\big]\,\frac{dx}{dy} = f\left(\sqrt{\frac{y}{a}}\right) \times \frac{1}{\sqrt{4ay}} = \frac{1}{\sqrt{4ay}}\,f\left(\sqrt{\frac{y}{a}}\right)$$

PE: Give two more examples (like the two given examples) about the transformation of random variables.

3. Provide more explanation about the term "cumulative distribution function" as used in the literature (or as it can be used in general regardless of the literature).
   Answer: The primary meaning of the term "cumulative distribution function" is as defined above (i.e. a real-valued function of a real-valued random variable that gives the probability of the random variable taking a value less than or equal to a given value). However, this term may be used rather differently to label similar types of "cumulative distribution functions". For instance, the term may be used to label a "function" of a random variable that gives the probability of the random variable taking a value between two given values, e.g. $P(1 \leq k \leq 7)$.[86] This type of cumulative distribution function may be distinguished from the "ordinary" cumulative distribution function by the label "*inner* cumulative distribution function". In fact, we can identify several types of "cumulative distribution function" which include for instance: $P(x \leq a)$, $P(x > a)$ and $P(a \leq x < b)$ where x is a random variable and a and b are given numbers. These (and other types) have a common feature of being "cumulative" since they represent the added probability of more than one value of the random variable (where this added probability is obtained by summation or integration) and hence they are "cumulative".
   PE: List all the possible types of "cumulative distribution function" in the broad sense of this term.

4.1 Probability Mass Functions

As indicated earlier, probability mass function is defined for discrete variables. If x is a random variable that can assume discrete distinct values x_i $(i = 1, 2, \ldots)$ with probabilities $P(x_i) = p_i$ then we have the following properties for the mass function $P(x)$:[87]

$$P(x_i) \geq 0 \tag{64}$$

$$P(x_j \cap x_k) = 0 \qquad (j \neq k) \tag{65}$$

$$P(x_j \cup x_k) = p_j + p_k \qquad (j \neq k) \tag{66}$$

$$P(\cup_i\, x_i) = \sum_i P(x_i) = \sum_i p_i = 1 \tag{67}$$

In the subsections of this section we investigate some of the well known and commonly used probability mass functions.

Problems

1. Justify the properties of the mass function (as expressed by Eqs. 64-67).
   Answer:

[86] We note that if the two given values are specific then this is not really a "function", so this label is used to represent this type in general considering the two limits (or values) as variables.

[87] Probably it is more appropriate to describe Eqs. 64 and 67 as conditions (although they are properties as well) and describe Eqs. 65 and 66 as properties.

• Eq. 64 is justified by the axioms of probability (see axiom 1 of § 3.1) noting that $P(x_i)$ is a probability.
• Eq. 65 is justified by the fact that x_j and x_k are mutually exclusive events and hence they are subject to Eq. 42.
• Eq. 66 is justified by the addition law of mutually exclusive events (i.e. Eq. 44 or rather axiom 3 of § 3.1).[88]
• Eq. 67 is justified by the axioms of probability (see axiom 2 of § 3.1).
Note: from Eqs. 64 and 67 we can conclude another property for the mass function, that is $P(x_i) \leq 1$ although this should be obvious from the fact that $P(x_i)$ is a probability. In fact, this property (like the property of Eq. 64) can be obtained directly from the axioms of probability (see axiom 1 of § 3.1).
PE:
(**a**) Can we say that the above properties of mass function are essentially an expression or demonstration of the fact that mass functions satisfy the conditions set by the axioms of probability?
(**b**) Give some real examples (e.g. from daily life or from science) of probability mass function, and show that these examples satisfy the properties of probability mass function (i.e. Eqs. 64-67).

2. Let have a discrete random variable which has k distinct and independent outcomes x_i each with a probability p_i ($i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$), and let have n independent trials involving this random variable. What is the probability $P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k)$ that in n_1 of these n trials we get the x_1 outcome, in n_2 of these n trials we get the x_2 outcome, ..., and in n_k of these n trials we get the x_k outcome (noting that $\sum_{i=1}^{k} n_i = n$)?
Answer: Referring to Problem 28 of § 2.2, we have $C^n_{n_1,n_2,\ldots,n_k}$ ways for getting "n_1 times of x_1, n_2 times of x_2, ..., and n_k times of x_k". Also, by the multiplication rule for independent events, each one of these $C^n_{n_1,n_2,\ldots,n_k}$ ways has a probability of $p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_k^{n_k}$. So, by the addition law for mutually exclusive events we have:

$$P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k) = C^n_{n_1,n_2,\ldots,n_k} \times p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_k^{n_k} = \frac{n!}{n_1!\, n_2! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i} \quad (68)$$

Note: it is worth noting that (where the sum is taken for all $n_1 + n_2 + \cdots + n_k = n$):

$$\begin{aligned}
\sum P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k) &= \sum C^n_{n_1,n_2,\ldots,n_k} \times p_1^{n_1} \times p_2^{n_2} \times \cdots \times p_k^{n_k} && \text{(Eq. 68)} \\
&= (p_1 + p_2 + \cdots + p_k)^n && \text{(Eq. 34)} \\
&= 1^n && \left(\sum_i p_i = 1\right) \\
&= 1
\end{aligned}$$

So, it meets the normalization condition.
PE: What is the relation between Eq. 33 and Eq. 34? Try to correspond the symbols in Eq. 33 to the symbols in Eq. 34.

3. Give some examples of (discrete) functions that can be probability mass functions (i.e. they meet the conditions of Eqs. 64-67).
Answer: The following functions can be probability mass functions because they meet the conditions of Eqs. 64-67:
(**a**) $P(k) = \frac{6-|k-7|}{36}$ $(k = 2, 3, \ldots, 12)$, $P(k) = 0$ (otherwise).
(**b**) $P(k) = \frac{1}{22}$ $(k = -1.5, -1, -0.5, \ldots, 9)$, $P(k) = 0$ (otherwise).
(**c**) $P(k) = \frac{40k}{283(k-1)}$ $(k = 2, 3, 5, 6, 9)$, $P(k) = 0$ (otherwise).
(**d**) $P(k) = \frac{4}{5^k}$ $(k = 1, 2, \ldots, \infty)$, $P(k) = 0$ (otherwise).
PE: Give more examples like the given ones.

[88] We note that Eq. 66 extends to more than two values of the random variable, e.g. $x_j.x_k, x_l$.

4. Which of the following can be a probability mass function (noting that we use P tentatively at this stage):
   (**a**) $P(k) = \frac{1}{k}$ $(k = 10, 11, \ldots, 30)$, $P(k) = 0$ (otherwise).
   (**b**) $P(k) = \frac{k}{40}$ $(k = -5, -4, \ldots, 10)$, $P(k) = 0$ (otherwise).
   (**c**) $P(k) = \frac{e^{-1}}{k!}$ $(k = 0, 1, 2, \ldots, \infty)$, $P(k) = 0$ (otherwise).
   (**d**) $P(k) = \frac{k}{210}$ $(k = 1, 2, \ldots, 20)$, $P(k) = 0$ (otherwise).
   Answer:
   (**a**) It cannot, because it does not satisfy the condition of Eq. 67.
   (**b**) It cannot, because it does not satisfy the condition of Eq. 64.
   (**c**) It can, because it satisfies the conditions of Eqs. 64-67.
   (**d**) It can, because it satisfies the conditions of Eqs. 64-67.
   PE: Verify the answer of this Problem by verifying the conditions of Eqs. 64-67 in each case.
5. Find the constant c in the following probability mass functions:
   (**a**) $P(k) = c\,2^k$ $(k = 1, 2, 3, 4, 5)$. (**b**) $P(k) = c\,k^{-2}$ $(k = 1, 2, 3, 4, 5, 6, 7)$.
   (**c**) $P(k) = c\,2^{-k}$ $(k = 1, 2, \ldots, \infty)$. (**d**) $P(k) = c\,k^{-2}$ $(k = 1, 2, \ldots, \infty)$.
   Answer: Any probability mass function must satisfy Eq. 67 and this can be used (i.e. by summing the function over its entire range and equating the result to 1) to infer the value of c in each case, that is:
   (**a**) $\sum_{k=1}^{5} c\,2^k = 1 \quad \rightarrow \quad 62c = 1 \quad \rightarrow \quad c = \frac{1}{62}$.
   (**b**) $\sum_{k=1}^{7} c\,k^{-2} = 1 \quad \rightarrow \quad \frac{266681}{176400}c = 1 \quad \rightarrow \quad c = \frac{176400}{266681}$.
   (**c**) $\sum_{k=1}^{\infty} c\,2^{-k} = 1 \quad \rightarrow \quad 1 \times c = 1 \quad \rightarrow \quad c = 1$.
   (**d**) $\sum_{k=1}^{\infty} c\,k^{-2} = 1 \quad \rightarrow \quad \frac{\pi^2}{6}c = 1 \quad \rightarrow \quad c = \frac{6}{\pi^2}$.
   PE: Find the constant c in the following probability mass functions:
   (**a**) $P(k) = c\,3^k$ $(k = 1, 2, 3, 4, 5, 6)$. (**b**) $P(k) = c\,k^{-3}$ $(k = 1, 2, 3, 4, 5)$.
   (**c**) $P(k) = c\,3^{-k}$ $(k = 1, 2, \ldots, \infty)$. (**d**) $P(k) = c\,k^{-3}$ $(k = 1, 2, \ldots, \infty)$.

4.1.1 Uniform Distribution

The discrete uniform distribution (which is the simplest mass function) is given by:

$$P(x = x_k) = \begin{cases} 1/n & (k = 1, 2, \ldots, n) \\ 0 & (\text{otherwise}) \end{cases} \tag{69}$$

where x is a discrete random variable that takes n different values (i.e. $x_1, x_2, \ldots, x_n$). For example, selecting a ball randomly from a bag containing exactly n identical balls is subject to this distribution because the probability of selecting any specific ball is $1/n$. For the uniform distribution $\big[$of the form $P(k) = 1/n$ where $k = 1, 2, \ldots, n\big]$ the mean μ and the variance V are given by (see Problem 9 of § 5.1 and Problem 6 of § 5.2):

$$\mu = \frac{1+n}{2} \qquad\qquad V = \frac{n^2 - 1}{12} \tag{70}$$

Problems

1. Give some examples of processes and experiments whose outcome is subject to the discrete uniform distribution.
   Answer:

- Tossing a fair coin (to get H, T).
- Rolling a fair die (to get $1, 2, 3, 4, 5, 6$).
- Drawing a card from a deck of cards.
- Giving birth (to boy or girl).
- Selecting a season (i.e. spring, summer, autumn, winter) at random in a year.
- Selecting an integer at random (to get an even or odd number).

PE: Give more examples of processes and experiments whose outcome is subject to the discrete uniform distribution.

2. Show that the discrete uniform distribution is normalized.
Answer: From Eq. 69 we have:

$$\sum_k P(x_k) = \sum_{k=1}^{n} \frac{1}{n} = \frac{1}{n} \sum_{k=1}^{n} 1 = \frac{1}{n} \times n = 1$$

PE: Explain and justify all the details of this derivation.

4.1.2 Binomial Distribution

The binomial distribution is probably the most important discrete probability distribution. The mass function of this distribution (which is used to model the probability distribution in a series of Bernoulli trials) is given by:

$$P(x = k) = P(n, p, k) = C_k^n p^k (1 - p)^{n-k} \qquad (k = 0, 1, \ldots, n) \tag{71}$$

where C_k^n is the binomial coefficient, p is the probability of occurrence, n (which is a non-negative integer) is the number of trials and k is the number of occurrences ($k = 0, 1, \ldots, n$). For example, getting k heads in a trial of throwing a coin n times (where the probability of getting head in each throw is p) is subject to this kind of distribution. For the binomial distribution the mean μ and the variance V are given by (see Problem 9 of § 5.1 and Problem 6 of § 5.2):

$$\mu = np \qquad V = np(1 - p) \tag{72}$$

It is important to note the following about the binomial distribution:
• For the binomial distribution to apply, p must be the same in all trials and the outcomes of the trials must be independent of each other (i.e. the trials are a Bernoulli process which we defined in the preamble of this chapter and indicated at the beginning of this subsection).
• The "binomial" designation is because only two possible outcomes are considered in this type of distribution, i.e. occurrence (or "success"; see next point) with probability p and non-occurrence (or "failure") with probability $(1 - p)$. The "binomial" designation is also because the probabilities are given by the terms of the binomial theorem (see Problem 20 of § 2.2).
• It is common in the literature of probability to label one of the outcomes in the binomial distribution (which is the occurrence of the event of primary interest in that distribution) as "success" and to label the other outcome as "failure". However, the reader should note that "success" (like "failure") is a label rather than a real success (and hence it could be a disaster like the occurrence of death or explosion) and that is why we prefer to use "occurrence" and "non-occurrence" instead of "success" and "failure" (although "success" and "failure" are widely used in the literature).
• The Bernoulli distribution is a special case of the binomial distribution corresponding to $n = 1$. However, the reader should be aware that "Bernoulli distribution" may be used to label the binomial distribution in general. In fact, the terminology about this issue (as almost about any other issue in science and mathematics) is not universal and hence attention is required when reading the literature.

Problems

1. Describe the binomial distribution and outline its main features.
   Answer: The binomial distribution is a probabilistic model for processes made of a number of identical and independent trials where each trial has only two possible outcomes (which are usually labeled as success and failure) and the probability of these outcomes is the same in all trials. The random variable in the binomial distribution represents the number of occurrences (or successes) k in a given number of trials n (where k can take any one of the integer values $k = 0, 1, \ldots, n$).
   PE: Suggest some reasons for why the binomial distribution is one of the most important discrete probability distribution functions (and possibly the most important of all).
2. Give some examples of the binomial distribution.
   Answer:
   • Getting k evens (and $9 - k$ odds) in a trial of throwing a die 9 times (where $k = 0, 1, \ldots, 9$).
   • Getting k defective items in n items ($0 \leq k \leq n$) manufactured in a production line whose probability of defect is p.
   PE: Give more examples for the binomial distribution.
3. Derive (or rather justify) Eq. 71 using the assumptions (outlined earlier) which the binomial distribution is based on.
   Answer: Because the trials are independent, the probability of k successes (each with probability p) and $n - k$ failures (each with probability $1 - p$) in a given order is $p^k(1-p)^{n-k}$ (see Eq. 57). Now, the number of different orders in which this can occur is given by $C^n_{k,n-k} = C^n_k$ (which is the number of permutations with repetitions where k objects are identical and $n-k$ objects are identical; see Problem 30 of § 2.2). So, according to the addition law of disjoint events (see Eq. 53) the total probability of k successes (each with probability p) and $n - k$ failures (each with probability $1 - p$) is $C^n_k p^k (1-p)^{n-k}$ which is what is given by Eq. 71.
   PE: Justify the use of the addition law of disjoint events (as given by Eq. 53) in the above argument.
4. Calculate the following probabilities of the binomial distribution given in the form $P(n, k, p)$:
   (**a**) $P(12, 6, 0.12)$. (**b**) $P(47, 33, 0.65)$. (**c**) $P(276, 201, 0.91)$.
   (**d**) $P(1672, 1569, 0.26)$. (**e**) $P(3927, 56, 0.99)$. (**f**) $P(5920, 4935, 0.43)$.
   Answer: From Eq. 71 we have:

   (**a**) $P(12, 6, 0.12) = C^{12}_6 \times 0.12^6 \times (1 - 0.12)^{12-6} \simeq 0.00128131373155$
   (**b**) $P(47, 33, 0.65) = C^{47}_{33} \times 0.65^{33} \times (1 - 0.65)^{47-33} \simeq 0.0947691317661$
   (**c**) $P(276, 201, 0.91) = C^{276}_{201} \times 0.91^{201} \times (1 - 0.91)^{276-201} \simeq 1.5350300806 \times 10^{-18}$
   (**d**) $P(1672, 1569, 0.26) = C^{1672}_{1569} \times 0.26^{1569}(1 - 0.26)^{1672-1569} \simeq 1.69204310134 \times 10^{-765}$
   (**e**) $P(3927, 56, 0.99) = C^{3923}_{56} \times 0.99^{56} \times (1 - 0.99)^{3927-56} \simeq 9.99983922704 \times 10^{-7617}$
   (**f**) $P(5920, 4935, 0.43) = C^{5920}_{4935} \times 0.43^{4935} \times (1 - 0.43)^{5920-4935} \simeq 1.2159880926 \times 10^{-894}$

   PE: Repeat the Problem for the following:
   (**a**) $P(59, 11, 0.36)$. (**b**) $P(183, 77, 0.05)$. (**c**) $P(592, 18, 0.18)$.
   (**d**) $P(927, 725, 0.67)$. (**e**) $P(3380, 2451, 0.22)$. (**f**) $P(6964, 2369, 0.85)$.
5. Show that the binomial distribution is normalized.
   Answer: From Eq. 71 we have:

$$\sum_k P(k) = \sum_{k=0}^{n} C^n_k p^k (1-p)^{n-k} = \Big[(1-p) + p\Big]^n = \Big[1\Big]^n = 1$$

   where we used the binomial theorem (see Eq. 33) in the second step.
   PE: Explain and justify all the details of this derivation.
6. Show that the probability of having at least one success in a binomial distribution is $1 - (1 - p)^n$.
   Answer: "Having at least one success" is equivalent to $k > 0$. Hence:

$$P(k > 0) = P(k = 0, 1, \ldots, n) - P(k = 0) = 1 - P(k = 0) = 1 - C^n_0 p^0 (1-p)^{n-0} = 1 - (1-p)^n$$

PE: Do the following:
(**a**) Justify each step of the above derivation.
(**b**) Show that in a binomial distribution $P(n, p, k)$ the number of trials n required to have at least one success with a probability greater than or equal to P $(0 < P < 1)$ is:

$$n \geq \frac{\ln(1-P)}{\ln(1-p)}$$

7. Investigate some of the limiting cases of the binomial distribution which may look dubious or problematic.
Answer: We investigate some of these limiting cases in the following points:
• $n = 0$ with $p \neq 0$ and $p \neq 1$: in this case the above formula (i.e. Eq. 71) is valid because:

$$P(0) = C_0^0 p^0 (1-p)^{0-0} = 1 \times 1 \times 1 = 1$$

which is correct because the probability of 0 occurrence in 0 number of trials is certainty.
• $n = 1$ with $p \neq 0$ and $p \neq 1$: in this case the above formula (i.e. Eq. 71) is valid because:

$$\begin{aligned} P(0) &= C_0^1 p^0 (1-p)^{1-0} = 1 \times 1 \times (1-p)^{1-0} = 1 \times 1 \times (1-p) = 1-p \\ P(1) &= C_1^1 p^1 (1-p)^{1-1} = 1 \times p \times (1-p)^0 = 1 \times p \times 1 = p \end{aligned}$$

which is correct because the first is the probability of non-occurrence and the second is the probability of occurrence.
• $p = 0$: in this case the above formula (i.e. Eq. 71) becomes problematic for $k = 0$ because:

$$P(0) = C_0^n p^0 (1-p)^{n-0} = 1 \times 0^0 \times 1^n \qquad (n = 0, 1, \ldots)$$

So, to salvage this formula in this case we need to adopt a convention that $0^0 = 1$. However, the formula should be OK for $k \neq 0$ including $k = n$ since in all these cases we have:

$$P(k) = C_k^n p^k (1-p)^{n-k} = C_k^n \times 0^k \times 1^{n-k} = 0 \qquad (k = 1, 2, \ldots, n)$$

which is correct because if $p = 0$ then the occurrence k times is impossible (noting that $k > 0$ which means that it does occur sometimes in contradiction with $p = 0$).
• $p = 1$: in this case the above formula (i.e. Eq. 71) becomes problematic for $k = n$ because:

$$P(n) = C_n^n p^n (1-p)^{n-n} = 1 \times 1^n \times 0^0 \qquad (n = 0, 1, \ldots)$$

So, to salvage this formula in this case we again need to adopt a convention that $0^0 = 1$. However, the formula should be OK for $k \neq n$ including $k = 0$ since in all these cases we have:

$$P(k) = C_k^n p^k (1-p)^{n-k} = C_k^n \times 1^k \times 0^{n-k} = 0 \qquad (k = 0, 1, 2, \ldots, n-1)$$

which is correct because if $p = 1$ then the occurrence k times is impossible (noting that $k < n$ which means that it does not occur sometimes in contradiction with $p = 1$).
PE: Investigate and analyze other potential limiting cases.
8. Investigate the variation of the shape of (the curve[89] representing) the binomial distribution with the variation of p for a given n (say $n = 100$) by plotting the binomial distribution with various values of p (say $p = 0.1, 0.3, 0.5, 0.7, 0.9$) on the same graph.
Answer: See Figure 16. As we see, as p increases the peak of the curve shifts to the right (since $\mu = np$ increases with increasing p). Regarding the height of the peak we note that it decreases with increasing p until we reach $p = 0.5$ and the curve steadily flattens with this increase of p (since the area under the curve should remain constant due to normalization), but this trend is reversed after $p = 0.5$.

[89] The use of "curve" (as well as other terms and expressions which are more appropriate for continuous distributions) in the case of binomial and other discrete distributions is for the sake of simplicity and clarity.

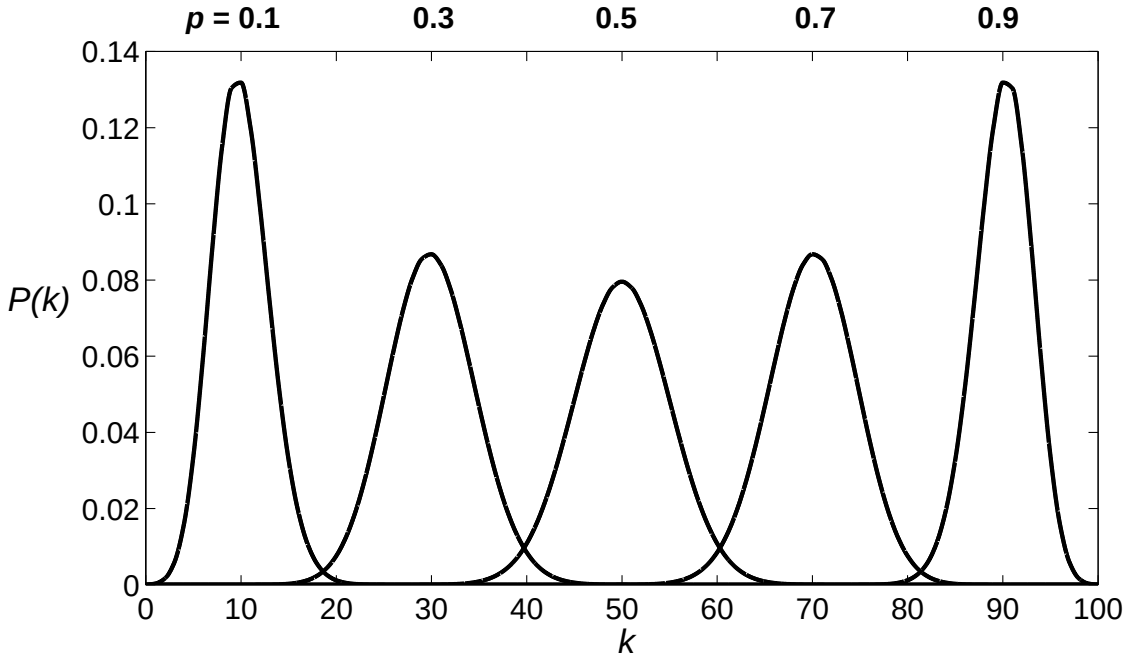


Figure 16: The plot of Problem 8 of § 4.1.2 (noting that $n = 100$ for all these curves). For clarity, we use solid curves instead of discrete points. Each number on the top (i.e. $p = 0.1, 0.3, 0.5, 0.7, 0.9$) belongs to the profile beneath it.

The reason of this behavior is that the variance of the binomial distribution (see Eq. 72 as well as § 5.2) is $V = np(1-p)$ which varies in this way, i.e. it increases up to $p = 0.5$ and then decreases after $p = 0.5$.[90]

PE: Repeat this Problem for $n = 200$ and $p = 0.05, 0.20, 0.35, 0.50, 0.65, 0.80, 0.95$.

9. Investigate the variation of the shape of (the curve representing) the binomial distribution with the variation of n for a given p (say $p = 0.5$) by plotting the binomial distribution with various values of n on the same graph.

   Answer: Noting that for the binomial distribution we have $\mu = np$ and $V = np(1-p)$, it should be obvious that increasing n results in shifting the peak of (the curve representing) the distribution to the right and flattening (the curve representing) the distribution. In Figure 17 we plotted the binomial distribution for a number of n's (i.e. $n = 20, 60, 100, 140, 180$). As we see, as n increases the peak of the curve shifts to the right and the curve becomes more flat by decreasing its height and broadening its width.

   PE: Repeat this Problem for other p's (e.g. $p = 0.3$).

10. Find the probability of:

    (**a**) Getting the number 6 three times (exactly) in a series of trials of throwing a fair die 7 times.

    (**b**) Getting 10 defective items (exactly) in 10^5 items manufactured in a production line whose probability of defect is 0.0003.

    Answer: These are obviously instances of the binomial distribution.

    (**a**) Using Eq. 71 with $n = 7$, $k = 3$ and $p = 1/6$, we have:

$$P(3) = C_3^7 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^4 \simeq 0.0781$$

[90] We note that at $p = 0.5$ we have $dV/dp = n(1-2p) = 0$ and $d^2V/dp^2 = -2n < 0$ which means that V has a peak at $p = 0.5$.

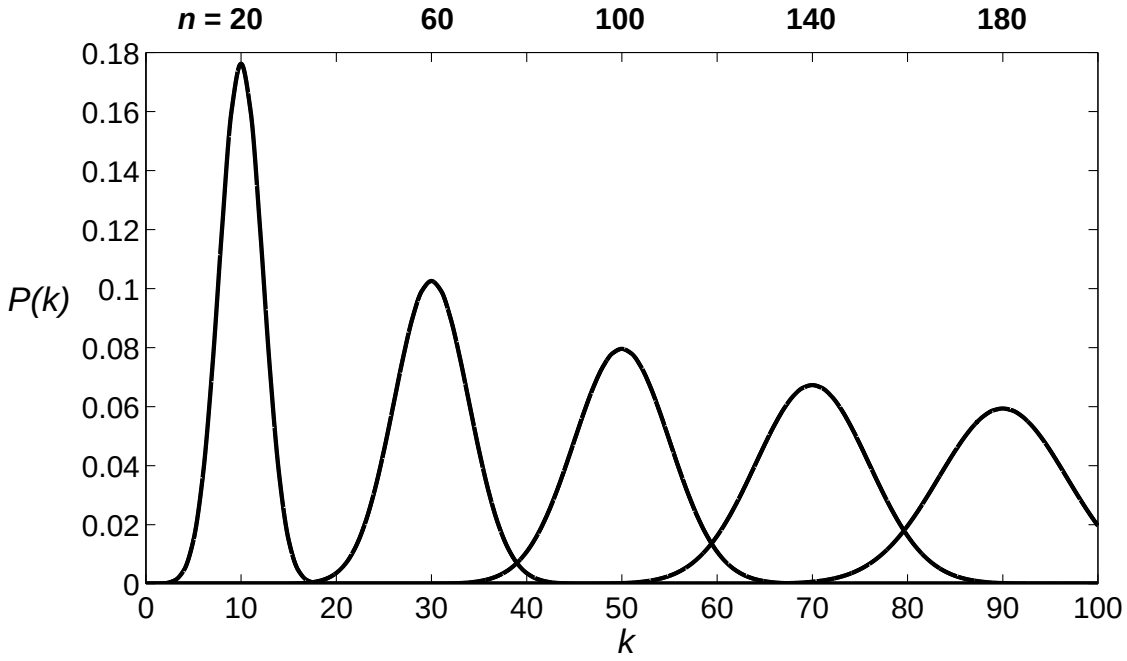


Figure 17: The plot of Problem 9 of § 4.1.2 (noting that $p = 0.5$ for all these curves). For clarity, we use solid curves instead of discrete points. Each number on the top (i.e. $n = 20, 60, 100, 140, 180$) belongs to the profile beneath it.

(**b**) Using Eq. 71 with $n = 10^5$, $k = 10$ and $p = 0.0003$, we have:

$$P(10) = C_{10}^{10^5} \ (0.0003)^{10} \ (0.9997)^{99990} \simeq 0.0000152$$

PE: Find the probability of:
(**a**) Getting a tail 2 times in a series of trials of throwing a fair coin 5 times.
(**b**) Getting zero defective items in 1000 items manufactured in a production line whose probability of defect is 0.007.

11. Make a 3D plot for the binomial distribution with $p = 0.5$ and $n = 5, 6, \ldots, 15$ (i.e. P as a function of n and k).
Answer: See Figure 18.
PE: Repeat the Problem with $p = 0.6$.
12. Write a simple program that calculates the probabilities of the binomial distribution.
Answer: See the BinomialProbability.cpp code which calculates the individual probabilities of the binomial distribution.
Note: to do extensive calculations of the binomial distribution (i.e. on a given range of k) conveniently, we wrote another code (see the BinomialDistribution.cpp code) in which we put the core of the BinomialProbability.cpp code into a k loop and output the results to a file. This code is especially useful for doing extensive calculations in extreme cases (such as those cases involving very large or/and very small numbers).
PE: Comment the BinomialProbability.cpp code explaining what each line is supposed to do.
13. Obtain a formula for the binomial probability $P(k+1)$ in terms of the binomial probability $P(k)$ and suggest useful applications and advantages of this formula.
Answer: From Eq. 71 we have (noting that $1 \leq k+1 \leq n$):

$$P(k+1) \quad = \quad C_{k+1}^{n} p^{k+1} (1-p)^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!} p^{k+1} (1-p)^{n-k-1}$$

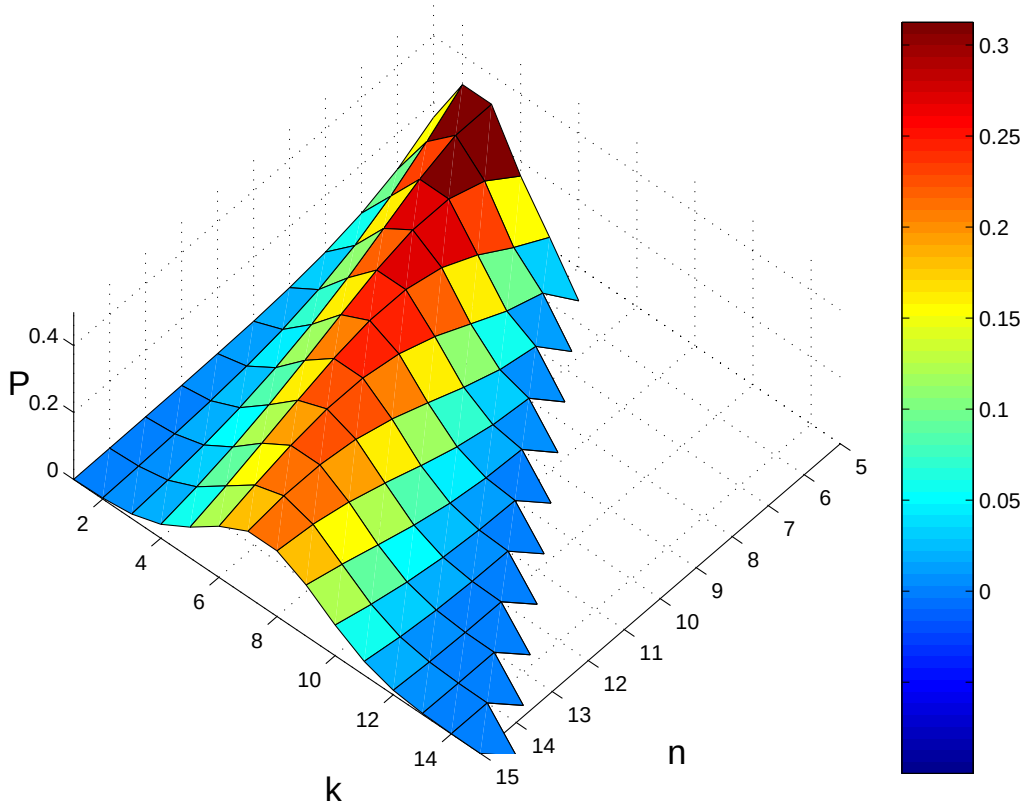


Figure 18: The plot of Problem 11 of § 4.1.2.

$$= \frac{p(n-k)}{(1-p)(k+1)}\left[\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}\right] = \frac{p(n-k)}{(1-p)(k+1)}P(k)$$

This recurrence formula is used to calculate successive binomial probabilities more easily and rapidly since this formula uses the previously calculated probability in the calculation of a given probability and hence it is more economic in terms of time and computing resources. Regarding the useful applications and advantages we can say (for example):

• The formula can alleviate some of the difficulties and hardships in the calculations of binomial probabilities involving large numbers by avoiding overflow in the calculations since these extreme calculations can be done in stages and hence the large numbers are balanced by scaling them down at each stage by multiplying them by tiny numbers which keeps the calculations manageable and under control (i.e. they do not exceed the limits and ceilings imposed by the calculation resources which lead to overflow).

• This formula can be especially important in the case of calculating cumulative binomial probabilities (see § 4.3) involving large numbers (e.g. $n = 10^6$) where the calculation of a long series of probabilities is required and hence it can reduce the time and computing resources needed to do these calculations.

• This formula can also be useful in analytical derivations and arguments where it could lead to simplifications or cancellations for instance.

PE: Do the following:

(**a**) Construct a spreadsheet (or write a computer code) in which you use this formula to do some relatively lengthy binomial distributions (e.g. with $n = 150$ and $p = 0.65$).

(**b**) Suggest other useful applications and advantages of this formula.

4.1.3 Multinomial Distribution

When we have a discrete random variable that has k distinct and independent outcomes x_i each with a probability p_i (where $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} p_i = 1$) and we made n trials involving this random variable, then the probability $P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k)$ that in n_i of these n trials we get the x_i outcome (where $\sum_{i=1}^{k} n_i = n$) is modeled by the multinomial distribution. In fact, the multinomial distribution was introduced and investigated briefly (without mentioning its name) in Problem 2 of § 4.1 (where it is used there as an introductory example or a case study for the probability mass functions). So, from the result of Problem 2 of § 4.1 the probability mass function of this distribution is given by:

$$P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k) = C^n_{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} = \frac{n!}{n_1!\, n_2! \ldots n_k!} \prod_{i=1}^{k} p_i^{n_i} = n! \prod_{i=1}^{k} \frac{p_i^{n_i}}{n_i!} \quad (73)$$

where $n = n_1 + n_2 + \cdots + n_k$ and $p_1 + p_2 + \cdots + p_k = 1$. For the multinomial distribution the mean μ and the variance V are given by (see Problem 9 of § 5.1 and Problem 6 of § 5.2):

$$\mu(x_i) = np_i \qquad V(x_i) = np_i(1 - p_i) \quad (74)$$

where $i = 1, 2, \ldots, k$.

It is important to note the following about the multinomial distribution:

• The "multinomial" label comes from the fact that the distribution is represented mathematically by the expansion of the multinomial theorem (see Problem 28 of § 2.2 and Eq. 34 in particular).

• The binomial distribution (see § 4.1.2) is a special case of the multinomial distribution[91] corresponding to $k = 2$ (and hence $n_1 + n_2 = n$ and $p_1 + p_2 = 1$). Now, if we note that the Bernoulli distribution is a special case of the binomial distribution (see the notes of § 4.1.2), then the Bernoulli distribution can also be seen as a special case of the multinomial distribution.

• Referring to Eq. 74, we have k means and k variances corresponding to the k possible outcomes, i.e. each outcome has a mean and a variance.

• The multinomial distribution can be considered and treated as a multivariate distribution (see § 4.4). However, in this book we avoid this approach and hence we mostly treat (and label) the k distinct and independent possibilities (i.e. x_i) as *outcomes*[92] of a discrete random variable x which has k distinct and independent outcomes x_i rather than being different random *variables* (although they can be similarly treated as different random variables). Our motivation for avoiding the multivariate approach is to avoid going through some details and complexities of the subject of multivariate distributions which is out of scope (although we will introduce this subject briefly in § 4.4).

Problems

1. Give some examples of probabilities that are subject to the multinomial distribution.
   Answer: For instance:
   • The probability of getting "1" once, "2" twice, and "3" five times in a trial of throwing 8 dice.
   • In a presidential election we have 5 candidates: the 1^{st} candidate got 41% of the vote, the 2^{nd} 27%, the 3^{rd} 15%, the 4^{th} 11%, and the 5^{th} 6%. If 10 voters are selected randomly, the probability that exactly 2 voters of these 10 voters have voted for each candidate is subject to the multinomial distribution.[93]
   PE: Give more examples of probabilities that follow the model of multinomial distribution.
2. In a game of gambling, the player throws 3 fair dice simultaneously and he wins if the sum is not less than 15. Find the probability of winning.

[91] Or alternatively, the multinomial distribution is a generalization of the binomial distribution where each trial has more than two possible outcomes (instead of the two possible outcomes assumed in the binomial distribution) with corresponding probabilities whose sum is 1.

[92] "Outcomes" here should not be understood as single (or individual) outcomes but rather as types (or categories) of outcomes.

[93] In fact, we should assume that the number of voters is very large (which is very realistic in a presidential election).

Answer: To win, he should get (3 sixes) or (2 sixes and 1 five) or (2 sixes and 1 four) or (2 sixes and 1 three) or (1 six and 2 fives) or (1 six and 1 five and 1 four) or (3 fives). From Eq. 73 we get:

$$\begin{aligned}
P(\text{3 sixes}) &= C^3_{0,0,0,0,0,3}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^3 = \left(\frac{1}{6}\right)^3 \\
P(\text{2 sixes and 1 five}) &= C^3_{0,0,0,0,1,2}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^2 = 3\left(\frac{1}{6}\right)^3 \\
P(\text{2 sixes and 1 four}) &= C^3_{0,0,0,1,0,2}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2 = 3\left(\frac{1}{6}\right)^3 \\
P(\text{2 sixes and 1 three}) &= C^3_{0,0,1,0,0,2}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2 = 3\left(\frac{1}{6}\right)^3 \\
P(\text{1 six and 2 fives}) &= C^3_{0,0,0,0,2,1}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^1 = 3\left(\frac{1}{6}\right)^3 \\
P(\text{1 six and 1 five and 1 four}) &= C^3_{0,0,0,1,1,1}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^1\left(\frac{1}{6}\right)^1 = 6\left(\frac{1}{6}\right)^3 \\
P(\text{3 fives}) &= C^3_{0,0,0,0,3,0}\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^0\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^0 = \left(\frac{1}{6}\right)^3
\end{aligned}$$

Now, these events are pairwise exclusive and hence from Eq. 53 we get:

$$\text{Probability of win } = 2\left(\frac{1}{6}\right)^3 + 4\times 3\left(\frac{1}{6}\right)^3 + 6\left(\frac{1}{6}\right)^3 = 20\left(\frac{1}{6}\right)^3 \simeq 0.09259$$

PE: Do you notice anything odd about the relation between n and k in the multinomial distribution formula (i.e. Eq. 73) which we used in the solution of this Problem?

3. Verify that the probability in the example of Problem 2 is normalized (i.e. the sum of the probabilities of all the possible outcomes is unity).
   Answer: When we throw 3 fair dice simultaneously we have three main possibilities:
   • All dice show different faces. The number of cases for this possibility is $P^6_3 = 120$ (because we are choosing 3 different faces out of 6). Now, the probability of each case is $(1/6)^3$ and the cases are pairwise exclusive, and hence the total probability of this possibility is $120 \times (1/6)^3$.
   • Only two dice show different faces. The number of cases for this possibility is $P^6_2 \times C^3_2 = 30 \times 3 = 90$ (because we are choosing 2 types of different faces out of 6 which is $P^6_2 = 30$; moreover one type is repetitive and hence we have $C^3_2 = 3$ options for choosing the 2 repetitive faces out of the 3 faces).[94] Now, the probability of each case is $(1/6)^3$ and the cases are pairwise exclusive, and hence the total probability of this possibility is $90 \times (1/6)^3$.
   • All dice show identical faces. The number of cases for this possibility is obviously 6 (because the faces could be 1 or 2 or 3 or 4 or 5 or 6). Now, the probability of each case is $(1/6)^3$ and the cases are pairwise exclusive, and hence the total probability of this possibility is $6 \times (1/6)^3$.
   On summing the probabilities of these three exhaustive and disjoint possibilities we get:

$$120 \times \left(\frac{1}{6}\right)^3 + 90 \times \left(\frac{1}{6}\right)^3 + 6 \times \left(\frac{1}{6}\right)^3 = \frac{216}{216} = 1$$

   PE: Justify the normalization shown in this Problem using a multinomial approach.

4. Calculate the following probabilities of the multinomial distribution given in the form:

$$P^{n_1,n_2,\ldots,n_k}_{p_1,p_2,\ldots,p_k} = C^n_{n_1,n_2,\ldots,n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

[94] Alternatively, we have 3 possibilities for the face that is different from the other two faces.

where $n = n_1 + n_2 + \ldots + n_k$ and $p_1 + p_2 + \ldots + p_k = 1$:

(**a**) $P^{6,9,15}_{0.34,0.39,0.27}$. (**b**) $P^{12,0,21,2,15}_{0.11,0.19,0.35,0.25,0.10}$. (**c**) $P^{8,2,65,1,33,99}_{0.02,0.26,0.12,0.36,0.08,0.16}$.

(**d**) $P^{3,19,17,5,6,108}_{0.2,0.1,0.3,0.07,0.1,0.23}$. (**e**) $P^{14,53,19,51,66,79}_{0.11,0.2,0.18,0.06,0.43,0.02}$. (**f**) $P^{9,83,78,197,150}_{0.13,0.09,0.69,0.06,0.03}$.

Answer: From Eq. 73 we have:

$$\begin{aligned}
\text{(a)}\quad P^{6,9,15}_{0.34,0.39,0.27} &= C^{30}_{6,9,15}\ 0.34^6\ 0.39^9\ 0.27^{15} \\
&\simeq 0.000739565021175 \\
\text{(b)}\quad P^{12,0,21,2,15}_{0.11,0.19,0.35,0.25,0.10} &= C^{50}_{12,0,21,2,15}\ 0.11^{12}\ 0.19^0\ 0.35^{21}\ 0.25^2\ 0.10^{15} \\
&\simeq 2.48247886983 \times 10^{-14} \\
\text{(c)}\quad P^{8,2,65,1,33,99}_{0.02,0.26,0.12,0.36,0.08,0.16} &= C^{208}_{8,2,65,1,33,99}\ 0.02^8\ 0.26^2\ 0.12^{65}\ 0.36^1\ 0.08^{33}\ 0.16^{99} \\
&\simeq 3.99728329044 \times 10^{-86} \\
\text{(d)}\quad P^{3,19,17,5,6,108}_{0.2,0.1,0.3,0.07,0.1,0.23} &= C^{158}_{3,19,17,5,6,108}\ 0.2^3\ 0.1^{19}\ 0.3^{17}\ 0.07^5\ 0.1^6\ 0.23^{108} \\
&\simeq 1.26258419896 \times 10^{-42} \\
\text{(e)}\quad P^{14,53,19,51,66,79}_{0.11,0.2,0.18,0.06,0.43,0.02} &= C^{282}_{14,53,19,51,66,79}\ 0.11^{14}\ 0.2^{53}\ 0.18^{19}\ 0.06^{51}\ 0.43^{66}\ 0.02^{79} \\
&\simeq 1.77512819128 \times 10^{-89} \\
\text{(f)}\quad P^{9,83,78,197,150}_{0.13,0.09,0.69,0.06,0.03} &= C^{517}_{9,83,78,197,150}\ 0.13^9\ 0.09^{83}\ 0.69^{78}\ 0.06^{197}\ 0.03^{150} \\
&\simeq 4.51844246930 \times 10^{-273}
\end{aligned}$$

PE: Repeat the Problem for the following:

(**a**) $P^{3,7,9,1,6}_{0.07,0.24,0.49,0.15,0.05}$. (**b**) $P^{150,0,23,33,12,8}_{0.23,0.15,0.16,0.14,0.13,0.19}$. (**c**) $P^{71,22,13,36,81,3}_{0.33,0.08,0.05,0.26,0.15,0.13}$.

5. Show that the multinomial distribution is normalized.

Answer: From Eq. 73 we have:

$$\begin{aligned}
\sum_{\forall\ n_1+n_2+\cdots+n_k=n} P(n_1, n_2, \ldots, n_k; p_1, p_2, \ldots, p_k) &= \sum_{\forall\ n_1+n_2+\cdots+n_k=n} C^n_{n_1,n_2,\ldots,n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k} \\
&= (p_1 + p_2 + \cdots + p_k)^n = (1)^n = 1
\end{aligned}$$

where we used the multinomial theorem (see Eq. 34) in the second step and used the fact that $\sum_{i=1}^{k} p_i = 1$ in the third step.

PE: Explain and justify (verbally) all the details of this derivation.

6. Find the probability for:

(**a**) The first example of Problem 1. (**b**) The second example of Problem 1.

Answer: We use in this answer the form $P^{n_1,n_2,\ldots,n_k}_{p_1,p_2,\ldots,p_k}$.

(**a**) Using Eq. 73 with $k = 6$ (corresponding to the 6 possible outcomes of each die), $n = 8$, $n_1 = 1$, $n_2 = 2$, $n_3 = 5$, $n_4 = n_5 = n_6 = 0$, $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$, we have:

$$P^{1,2,5,0,0,0}_{1/6,1/6,1/6,1/6,1/6,1/6} = C^8_{1,2,5,0,0,0} \times \left(\frac{1}{6}\right)^1 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{1}{6}\right)^5 \times \left(\frac{1}{6}\right)^0 \times \left(\frac{1}{6}\right)^0 \times \left(\frac{1}{6}\right)^0 \simeq 0.0001$$

(**b**) Using Eq. 73 with $k = 5$ (corresponding to the 5 candidates), $n = 10$, $n_1 = n_2 = n_3 = n_4 = n_5 = 2$, $p_1 = 0.41$, $p_2 = 0.27$, $p_3 = 0.15$, $p_4 = 0.11$, $p_5 = 0.06$, we have:

$$P^{2,2,2,2,2}_{0.41,0.27,0.15,0.11,0.06} = C^{10}_{2,2,2,2,2} \times 0.41^2 \times 0.27^2 \times 0.15^2 \times 0.11^2 \times 0.06^2 \simeq 0.00136$$

PE: Repeat:

(**a**) Part (a) of this Problem for $P^{1,1,0,3,2,1}_{1/6,1/6,1/6,1/6,1/6,1/6}$.

(**b**) Part (b) of this Problem for $P^{1,2,4,0,3}_{0.41,0.27,0.15,0.11,0.06}$.

7. Write a simple program that calculates the probabilities of the multinomial distribution.
   Answer: See the MultinomialProbability.cpp code which calculates the individual probabilities of the multinomial distribution.
   PE: Describe the method which the MultinomialProbability.cpp code uses to calculate the probabilities of the multinomial distribution.

4.1.4 Poisson Distribution

The Poisson distribution is given by:

$$P(x = k) = P(\lambda, k) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (k = 0, 1, 2, \ldots) \tag{75}$$

where $\lambda > 0$ is the Poisson parameter and k is the number of occurrences. The Poisson distribution is commonly used to model the probability of an event occurring a given number of times k within a given time period (or/and a given spatial region). For example, having k decays in a minute by a given sample of radioactive material is subject to this kind of distribution. For the Poisson distribution the mean μ and the variance V are equal to λ, that is (see Problem 9 of § 5.1 and Problem 6 of § 5.2):

$$\mu = V = \lambda \tag{76}$$

It is important to note the following points about the Poisson distribution:
• The Poisson distribution is based on the assumption that the probability of occurrence is constant throughout the process and the events are independent of each other. In fact, there are other assumptions related to the size of event rate and the total number of events.
• The Poisson distribution can be seen as a limiting case for the binomial distribution when the number of trials n becomes large ($n \to \infty$) and the probability of occurrence p becomes small ($p \to 0$) such that $\lambda = np$ stays finite and constant (see Problem 2). Therefore, the Poisson distribution can be used (and is used) as a substitute for the binomial distribution in this case (where the relative ease of calculating the Poisson compared to calculating the binomial makes the shift from the binomial to the Poisson advantageous).
• The specifications and conditions given in the previous point are rather generic and loose and they require more explanation and details (noting for instance that other factors like k and np affect the quality of the approximation of binomial by Poisson). So, a case-by-case assessment is required (or at least recommended) before using the Poisson distribution (primarily or as an approximation) to model a given probability problem. However, these details are not important to us and hence we ignore them (noting that these issues are not treated fairly and properly in the literature).
• Referring to the previous points, an excuse that is usually presented in the literature for shifting from the binomial to the Poisson (as an approximation to the binomial) is that calculating C_k^n in the binomial formula could become difficult when n is large. However, we note that the computers (associated with mathematical software packages and programming languages) these days usually avoid this difficulty. Nevertheless, the Poisson formula involves less (and usually smaller) factorials than the binomial formula and hence it offers more efficiency in computation and less difficulty in calculation and these factors are generally advantageous especially in the cases of extreme calculations which involve large numbers or/and many cases (e.g. calculating probabilities involving factorials of integers of order 10^4 for 10^6 times where computation time becomes significant even on modern computers).

Problems

1. Give some examples of probabilities modeled by the Poisson distribution.
   Answer:
   • Getting a given number of visitors to a clinic during a day.
   • Getting a given number of visitors to a website during a week.
   • Having a given number of infections by corona virus in a city during 2024.
   • Having a given number of decays in a given time period by a given sample of uranium.

• Observing a given number of spiral galaxies in a certain zone of the sky dome.

PE: Give more examples of the Poisson distribution.

2. What is the most distinctive feature that distinguishes the Poisson distribution from the binomial distribution (noting that both are discrete)?

   Answer: It is the fact that while the number of trials n in the binomial distribution is a given finite constant, the number of trials in the Poisson distribution is not identified or limited (and hence in the binomial distribution we have $k = 0, 1, \ldots, n$ while in the Poisson distribution we have $k = 0, 1, 2, \ldots$). In fact, this is related to what we stated in the second point in the preamble of this subsection.

   PE: According to the literature, the Poisson distribution represents probabilistic situations in which discrete events occur independently in a continuum at a rate of λ. Try to justify this.

3. Justify Eq. 75 (assuming that Poisson is a limit for the binomial when $n \to \infty$ and $p \to 0$ such that $\lambda = np$ stays finite and constant).

   Answer: Based on the given assumption, we start from Eq. 71, that is:

$$P(k) = C_k^n p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n) \tag{77}$$

   Now, if we assume that $k \ll n$ (noting that $n \to \infty$) then we have:

$$\frac{n!}{(n-k)!} = n(n-1)\cdots(n-k+1) \simeq n^k \tag{78}$$

   Moreover, if we note that $p \to 0$ then we have (noting that $p = \lambda/n$):

$$(1-p)^{n-k} = \frac{(1-p)^n}{(1-p)^k} \simeq \frac{\left(1-\frac{\lambda}{n}\right)^n}{1} \simeq \frac{e^{-\lambda}}{1} = e^{-\lambda} \tag{79}$$

   where we used the identity $e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ in the third step. On substituting from Eqs. 78 and 79 into Eq. 77 we get (noting that $p = \lambda/n$):

$$P(k) = \frac{1}{k!}\, n^k \, \left(\frac{\lambda}{n}\right)^k \, e^{-\lambda} = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (k = 0, 1, 2, \ldots)$$

   which is Eq. 75.

   PE: By inspecting and analyzing the above justification (and the stated assumptions in particular), try to set some broad rules (mainly of practical nature) about when the Poisson distribution can be used as a good approximation to the binomial distribution.

4. Calculate the following probabilities of the Poisson distribution given in the form $P(k, \lambda)$:

   (**a**) $P(4, 6.7)$. (**b**) $P(11, 392.7)$. (**c**) $P(481, 2.8)$.

   (**d**) $P(663, 7.9)$. (**e**) $P(7829, 6961.5)$. (**f**) $P(9947, 5915.3)$.

   Answer: Using Eq. 75 we get:

   (**a**) $P(4, 6.7) \simeq 1.033511 \times 10^{-1}$. (**b**) $P(11, 392.7) \simeq 2.432603 \times 10^{-150}$.

   (**c**) $P(481, 2.8) \simeq 8.139821 \times 10^{-870}$. (**d**) $P(663, 7.9) \simeq 1.444462 \times 10^{-993}$.

   (**e**) $P(7829, 6961.5) \simeq 1.254247 \times 10^{-25}$. (**f**) $P(9947, 5915.3) \simeq 2.273048 \times 10^{-497}$.

   PE: Repeat the Problem for the following:

   (**a**) $P(11, 9.1)$. (**b**) $P(37, 222.4)$. (**c**) $P(3611, 56.8)$.

   (**d**) $P(5390, 153.7)$. (**e**) $P(829, 8421.7)$. (**f**) $P(9381, 7429.4)$.

5. What is the effect of varying λ on the shape and position of (the curve representing) the Poisson distribution?

   Answer: Noting that for the Poisson distribution we have $\mu = V = \lambda$, increasing λ results in shifting the peak of (the curve representing) the distribution to the right and flattening (the curve

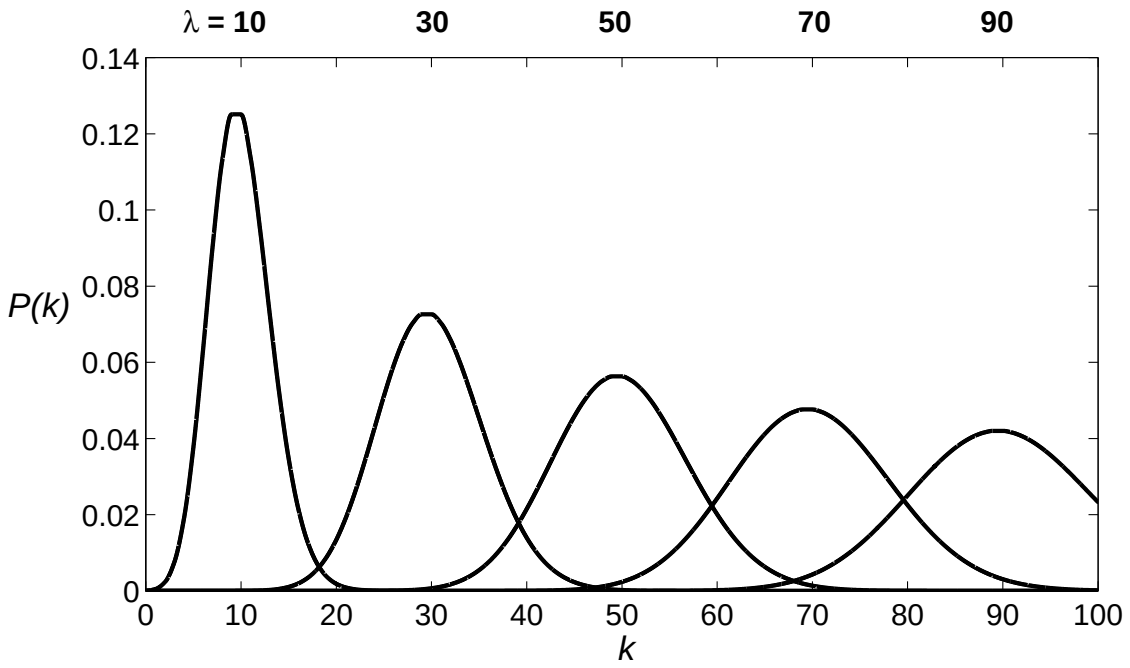


Figure 19: The plot of Problem 5 of § 4.1.4. For clarity, we use solid curves instead of discrete points. Each number on the top (i.e. $\lambda = 10, 30, 50, 70, 90$) belongs to the profile beneath it.

representing) the distribution. In Figure 19 we plotted the Poisson distribution for a number of λ's (i.e. $\lambda = 10, 30, 50, 70, 90$). As we see, as λ increases the peak of the curve shifts to the right (because $\mu = \lambda$) and the curve becomes more flat by decreasing its height and broadening its width (because $V = \lambda$).[95]

PE: Compare this behavior to the behavior of the corresponding binomial distribution (see Problems 8 and 9 of § 4.1.2) noting that for the binomial distribution $\mu = np$ and $V = np(1-p)$. Try to link the two behaviors to the relationship between the two distributions noting that for the Poisson distribution $\lambda = \mu = V$.

6. Show that the Poisson distribution is normalized.
   Answer: From Eq. 75 we have:

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

   where we used the exponential series in the third step.
   PE: Explain and justify (in words) all the details of this derivation.
7. A (large) specimen of a radioactive material (with very long half-life) is recorded to emit 15920 β-ray particles during one hour. Plot the probability of having k emissions per second for $k = 0, 1, \ldots, 15$ (assuming a Poisson distribution).
   Answer: We have $\lambda = \mu = 15920/3600 \simeq 4.42222$ emissions per second. Using Eq. 75, we calculated $P(k)$ for $k = 0, 1, \ldots, 15$ and plotted the results in Figure 20.
   PE: Repeat the Problem for 34197 emissions in 5 hours.
8. A population of 1592738 individuals were given a vaccine whose probability of causing blood clots is $p = 0.0000053$. Plot the probability of having k cases of blood clotting (for $k = 0, 1, \ldots, 20$) in this

[95] A feature of the Poisson distribution (which cannot be easily noticed in Figure 19) is that as λ increases the shape of the curve becomes more symmetric (about its value at the peak).

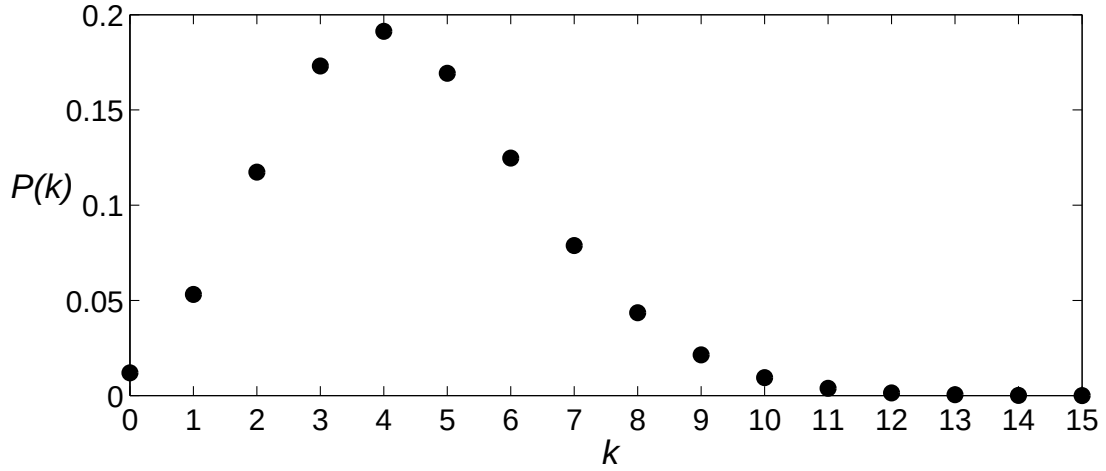


Figure 20: The plot of Problem 7 of § 4.1.4.

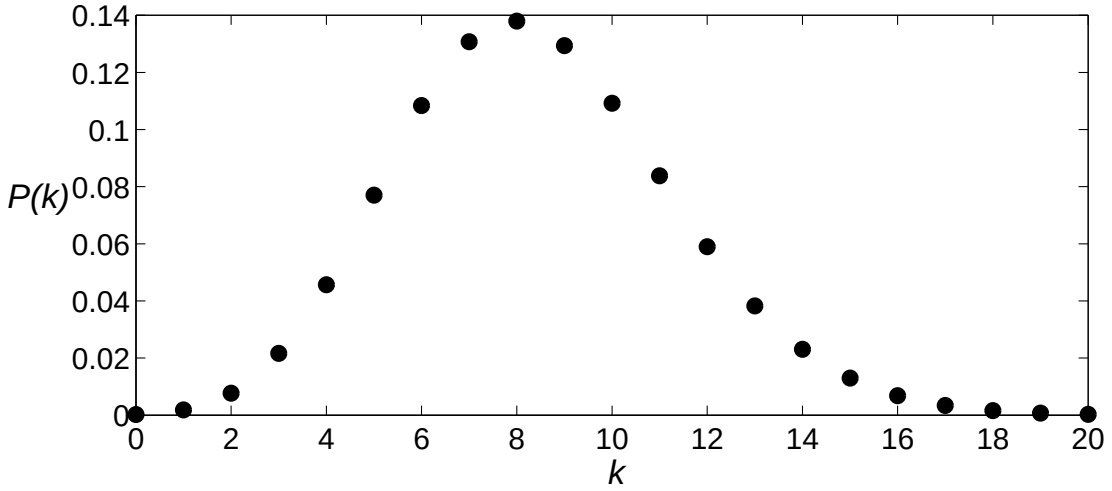


Figure 21: The plot of Problem 8 of § 4.1.4.

process of mass vaccination (assuming a Poisson distribution).
Answer: We have $\lambda = \mu = np = 1592738 \times 0.0000053 = 8.4415114$ cases of clotting.[96] Using Eq. 75, we calculated $P(k)$ for $k = 0, 1, \ldots, 20$ and plotted the results in Figure 21.
PE: Repeat the Problem for 2372042 individuals with probability of clotting $p = 0.0000037$.

9. Similar to what we did in Problem 13 of § 4.1.2, derive a recurrence formula for the Poisson distribution.
Answer: From Eq. 75 we have:

$$P(k+1) = \frac{\lambda^{k+1} e^{-\lambda}}{(k+1)!} = \frac{\lambda}{(k+1)} \left[\frac{\lambda^k e^{-\lambda}}{k!} \right] = \frac{\lambda}{(k+1)} P(k)$$

PE: Suggest useful applications and advantages of this recurrence formula.

10. Give an example showing that the Poisson distribution is generally more convenient in calculation than the corresponding binomial distribution and hence justify the use of Poisson as an approximation to the binomial (when the approximation of binomial by Poisson is valid).

[96] As we see, $\mu = np$ should remind us that the binomial is in the background.

Answer: For example, let have a binomial probability problem where we have to calculate the probability $P(n,p,k_1 \leq k \leq k_2) = P(200, 0.05, 6 \leq k \leq 8)$. Now, the binomial probability is:

$$\begin{aligned} P(200, 0.05, 6 \leq k \leq 8) &= C_6^{200}\, 0.05^6\, 0.95^{194} + C_7^{200}\, 0.05^7\, 0.95^{193} + C_8^{200}\, 0.05^8\, 0.95^{192} \\ &= \frac{200!}{6!\,194!}\, 0.05^6\, 0.95^{194} + \frac{200!}{7!\,193!}\, 0.05^7\, 0.95^{193} + \frac{200!}{8!\,192!}\, 0.05^8\, 0.95^{192} \\ &\simeq 0.0614 + 0.0896 + 0.1137 \simeq 0.2647 \end{aligned}$$

On the other hand, the corresponding Poisson probability $P(\lambda, k_1 \leq k \leq k_2)$ is (noting that $\lambda = np = 10$):

$$\begin{aligned} P(10, 6 \leq k \leq 8) &= e^{-10}\left(\frac{10^6}{6!} + \frac{10^7}{7!} + \frac{10^8}{8!}\right) \\ &\simeq 0.0631 + 0.0901 + 0.1126 \simeq 0.2657 \end{aligned}$$

As we see, the Poisson probability requires much less calculations than the corresponding binomial probability and the results are very similar with a minute relative percentage error of about 0.4% (noting that the conditions for the validity of this approximation are generally satisfied in this example).
PE: Repeat the Problem by giving another example.

11. Write a simple program that calculates the probabilities of the Poisson distribution.
Answer: See the PoissonProbability.cpp code which calculates the individual probabilities of the Poisson distribution.
Note: to do extensive calculations of Poisson distribution (i.e. on a given range of k) conveniently, we wrote another code (see the PoissonDistribution.cpp code) in which we put the core of the PoissonProbability.cpp code into a k loop and output the results to a file. This code is especially useful for doing extensive calculations in extreme cases (such as those cases involving very large or/and very small numbers).
PE: Comment the PoissonProbability.cpp code explaining what each line is supposed to do.

4.1.5 Geometric Distribution

The random variable in this probability distribution represents the number of (binomial or Bernoulli) trials required to obtain the first success. Accordingly, this distribution may be regarded as a special case of the binomial distribution (although this should be understood as a similarity rather than being so literally). It should be obvious that this distribution is given by:

$$P(x = k) = P(p, k) = (1-p)^{k-1} p \qquad (k = 1, 2, \ldots) \tag{80}$$

where k is the number of trials required to obtain the first success and p is the probability of success $(0 < p < 1)$.[97] For the geometric distribution the mean μ and the variance V are given by (see Problem 9 of § 5.1 and Problem 6 of § 5.2):

$$\mu = \frac{1}{p} \qquad V = \frac{1-p}{p^2} \tag{81}$$

It is important to note the following about the geometric distribution:
• The "geometric" label comes from the fact that the terms of this distribution represent a geometric sequence where each term is obtained from the previous one by multiplying it by a factor of $(1-p)$.
• There is another form of the geometric distribution which is used for representing the number of failures before the first success, and this form requires slight modifications to the above formulations and conditions. So, the readers should be aware of this to avoid confusion (see Problem 7).
• The geometric distribution is a special case of the negative binomial distribution (which will be investigated in § 4.1.6) corresponding to $r = 1$ (noting the difference in k between the two distributions).

[97] The latter condition may be stated as $0 < p \leq 1$ but we prefer to exclude the case of $p = 1$.

• The geometric distribution is commonly used to model the waiting time or the lifetime in probabilistic processes.[98]

Problems

1. Justify Eq. 80.
   Answer: The random variable in the geometric distribution represents the number k of binomial trials required to obtain the first success. This means that we should have $k-1$ failures (each with probability $1-p$) before we get the first success at the k^{th} trial (where the probability of success is p). Hence, by the multiplication law of independent events (see Eq. 57) the probability of obtaining the first success at the k^{th} trial should be $(1-p)^{k-1}p$, which is what Eq. 80 states.
   PE: What is the mass function of a probability distribution in which the random variable represents the number of binomial trials required to obtain the first failure?
2. Give some examples of processes that can be modeled by the geometric distribution.
   Answer:
   • Tossing a coin until a tail is obtained (where success is defined as "getting tail").
   • Throwing a die until 4 appears (where success is defined as "getting 4").
   PE: Give more examples of processes that can be modeled by the geometric distribution.
3. Show that the geometric distribution is normalized.
   Answer: From Eq. 80 we have:

$$\sum_k P(k) = \sum_{k=1}^{\infty}(1-p)^{k-1}p = p\sum_{k=1}^{\infty}(1-p)^{k-1} = p \times \frac{(1-p)^0}{1-(1-p)} = p \times \frac{1}{p} = 1$$

   where we used the well-known geometric series formula $\left[\text{i.e. } S = a_1/(1-r)\right]$ in the third equality.
   PE: Why the series $\sum_{k=1}^{\infty}(1-p)^{k-1}p$ should be convergent?
4. What is the effect of varying p on the shape of (the curve representing) the geometric distribution?
   Answer: We note that the first value of this distribution is p $\left[\text{i.e. } P(1) = p\right]$ and hence the curve representing this distribution starts high for high p (and low for low p). Now, since the area under the curve should be unity (due to normalization), we should expect the curve to drop faster for high p (and slower for low p).[99] These features are evident in Figure 22 where we plotted the geometric distribution for $p = 0.1, 0.3, 0.5, 0.7, 0.9$ between $k = 1$ and $k = 10$.
   PE: Plot similar curves for $p = 0.2, 0.4, 0.6, 0.8$.
5. In a game of chance, the player throws a die and a coin simultaneously until he gets a given face of the die (say 6) and a given face of the coin (say head H) *simultaneously*.[100] Find the probability distribution of this game and calculate the chance of winning (where win is determined by getting 6 and H before the fourth trial).
   Answer: This is an instance of the geometric distribution with:

$$p = P(6 \cap H) = P(6) \times P(H) = \frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$$

   where we use the multiplication law of independent events (see Eq. 50) since the events of getting 6 and getting H are independent. Accordingly, the probability distribution for this game is (see Eq. 80):

$$P(k) = (1-p)^{k-1}p = \left(1 - \frac{1}{12}\right)^{k-1} \times \frac{1}{12}$$

[98] This may be justified by the property of the geometric distribution (which is shared also by the exponential distribution; see § 4.2.3) that it is "memoryless".

[99] For clarity and simplicity we use a language more appropriate for continuous distributions (e.g. using "curve" and "area"). Alternatively, we may say (using a language more appropriate for discrete distributions): the successive values of this distribution (corresponding to successive k) are obtained by multiplication with $(1-p) < 1$ and hence larger p means smaller $(1-p)$ and hence faster drop while smaller p means larger $(1-p)$ and hence slower drop.

[100] Simultaneously means getting (6 AND H) at the same trial.

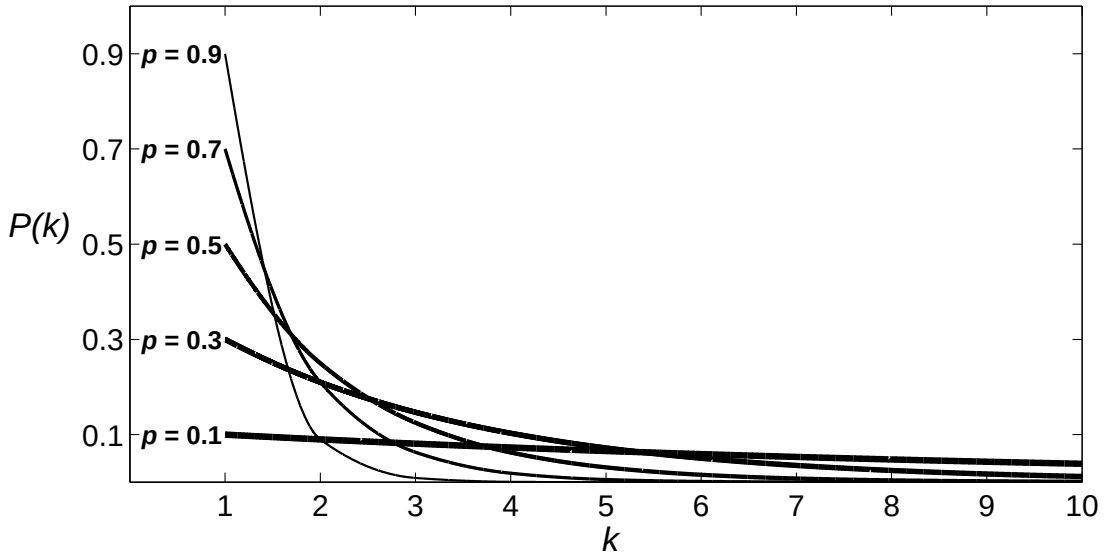


Figure 22: The plot of Problem 4 of § 4.1.5. For clarity, we use solid curves instead of discrete points.

The chance of winning is represented by the probability of getting 6 and H (*simultaneously*) in the 1^{st} or 2^{nd} or 3^{rd} trial, and hence it is obtained by summing these probabilities (of these disjoint events), that is:

$$P(\text{win}) = \sum_{k=1}^{3} \left(1 - \frac{1}{12}\right)^{k-1} \times \frac{1}{12} = \frac{1}{12} + \frac{11}{12^2} + \frac{11^2}{12^3} = \frac{397}{1728} \simeq 0.229745$$

PE: Repeat the Problem assuming now that win is determined by getting (*simultaneously*) an even number on the die and H on the coin before the third trial.

6. In a game of chance, the player throws a die and a coin simultaneously until he gets a given face of the die (say 6) and a given face of the coin (say head H) *simultaneously* or *separately*.[101] Find the probability distribution of this game and calculate the chance of winning (where win is determined by getting 6 and H before the fourth trial).

   Answer: The probability of getting 6 and H (*simultaneously* or *separately*) at the k^{th} trial can be obtained from the sum of the following probabilities (noting that the events of these probabilities are mutually exclusive):

   • $P(H_{\text{at } k} \cap 6_{\text{at or before } k})$ which is the probability of getting the first H at the k^{th} trial and getting 6 at or before the k^{th} trial.

   • $P(6_{\text{at } k} \cap H_{\text{before } k})$ which is the probability of getting the first 6 at the k^{th} trial and getting H before the k^{th} trial (noting that $k > 1$ in this case because there is no "before" in the first trial).

   Now, $P(H_{\text{at } k} \cap 6_{\text{at or before } k})$ is given by:

$$\begin{aligned} P(H_{\text{at } k} \cap 6_{\text{at or before } k}) &= P(H_{\text{at } k}) \times P(6_{\text{at or before } k}) && \text{(Eq. 50)} \\ &= \left[\left(1 - \frac{1}{2}\right)^{k-1} \frac{1}{2}\right] \times \left[\sum_{i=1}^{k} \left(1 - \frac{1}{6}\right)^{i-1} \frac{1}{6}\right] && \text{(Eq. 80)} \\ &= \left(\frac{1}{2}\right)^{k} \times \left[\sum_{i=1}^{k} \left(1 - \frac{1}{6}\right)^{i-1} \frac{1}{6}\right] \\ &= \left(\frac{1}{2}\right)^{k} \times \left[\frac{1}{6} \sum_{i=1}^{k} \left(\frac{5}{6}\right)^{i-1}\right] \end{aligned}$$

[101] Separately means getting the first 6 (or H) at the k^{th} trial and getting H (or 6) once or more before the k^{th} trial.

$$= \left(\frac{1}{2}\right)^k \times \left[\frac{1}{6} \times \frac{1-(5/6)^k}{1-(5/6)}\right] \qquad \text{(geometric series)}$$

$$= \left(\frac{1}{2}\right)^k \times \left[1-\left(\frac{5}{6}\right)^k\right] = \frac{6^k-5^k}{12^k} \qquad (k=1,2,\ldots)$$

Similarly, $P(6_{\text{at } k} \cap H_{\text{before } k})$ is given by (noting that $k>1$):

$$P(6_{\text{at } k} \cap H_{\text{before } k}) = P(6_{\text{at } k}) \times P(H_{\text{before } k}) \qquad \text{(Eq. 50)}$$

$$= \left[\left(1-\frac{1}{6}\right)^{k-1}\frac{1}{6}\right] \times \left[\sum_{i=1}^{k-1}\left(1-\frac{1}{2}\right)^{i-1}\frac{1}{2}\right] \qquad \text{(Eq. 80)}$$

$$= \left[\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}\right] \times \left[\sum_{i=1}^{k-1}\left(1-\frac{1}{2}\right)^{i-1}\frac{1}{2}\right]$$

$$= \left[\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}\right] \times \left[\sum_{i=1}^{k-1}\left(\frac{1}{2}\right)^{i}\right]$$

$$= \left[\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}\right] \times \left[\frac{1}{2} \times \frac{1-(1/2)^{k-1}}{1-(1/2)}\right] \qquad \text{(geometric series)}$$

$$= \left[\left(\frac{5}{6}\right)^{k-1}\frac{1}{6}\right] \times \left[1-\left(\frac{1}{2}\right)^{k-1}\right] = \frac{10^{k-1}-5^{k-1}}{6\times 12^{k-1}} \qquad (k=2,3,\ldots)$$

Accordingly, the probability of getting 6 and H (*simultaneously* or *separately*) at the k^{th} trial is [noting that $P(6_{\text{at } k} \cap H_{\text{before } k})$ as given by the last equation is equal to 0 for $k=1$]:

$$P(k) = P(H_{\text{at } k} \cap 6_{\text{at or before } k}) + P(6_{\text{at } k} \cap H_{\text{before } k}) = \frac{6^k-5^k}{12^k} + \frac{10^{k-1}-5^{k-1}}{6\times 12^{k-1}} \qquad (k=1,2,\ldots)$$

Regarding the chance of winning, it is represented by the probability of getting 6 and H (*simultaneously* or *separately*) in the 1^{st} or 2^{nd} or 3^{rd} trial, and hence it is obtained by summing these probabilities (of these disjoint events), that is:

$$P(\text{win}) = \sum_{k=1}^{3}\left(\frac{6^k-5^k}{12^k} + \frac{10^{k-1}-5^{k-1}}{6\times 12^{k-1}}\right)$$

$$= \frac{1}{12} + 0 + \frac{6^2-5^2}{12^2} + \frac{10-5}{6\times 12} + \frac{6^3-5^3}{12^3} + \frac{10^2-5^2}{6\times 12^2} = \frac{637}{1728} \simeq 0.368634$$

Note: this distribution is normalized, that is:

$$\sum_{k=1}^{\infty}\left(\frac{6^k-5^k}{12^k} + \frac{10^{k-1}-5^{k-1}}{6\times 12^{k-1}}\right) = \frac{2}{7} + \frac{1}{6}\frac{30}{7} = \frac{2}{7} + \frac{5}{7} = 1$$

PE: Repeat the Problem assuming now that win is determined by getting (*simultaneously* or *separately*) an odd number on the die and T on the coin before the fifth trial.

7. Outline the other form of the geometric distribution (which was indicated earlier in the notes).
Answer: We have (noting that in this form k represents the number of failures before the first success):

$$P(p,k) = (1-p)^k p \qquad \mu = \frac{1-p}{p} \qquad V = \frac{1-p}{p^2}$$

where $k = 0,1,2,\ldots$ and $0<p<1$.
PE: Justify the formulations and conditions given in this Problem for the other form of the geometric distribution (ignoring μ and V).

4.1.6 Other Discrete Distributions

There are many other discrete distributions some of which are outlined in the following points:
• **Negative binomial distribution**: this distribution represents the probability of the number of failures before the r^{th} success in a series of Bernoulli (or binomial) trials, e.g. the probability in a series of coin-tossing trials of having 5 H (where H represents failure) before obtaining 3 T (where T represents success). If r represents the number of successes (each with probability p) and k represents the number of failures (each with probability $1-p$) then the mass function of this distribution is given by:

$$P(k,r,p) = C_k^{k+r-1}\, p^r\, (1-p)^k \tag{82}$$

where $k = 0, 1, \ldots$ and $r = 1, 2, \ldots$ while $0 < p < 1$. The mean and variance of this distribution are given by:

$$\mu = \frac{r(1-p)}{p} \qquad\qquad V = \frac{r(1-p)}{p^2} \tag{83}$$

As indicated earlier, the geometric distribution (see § 4.1.5) is a special case of the negative binomial distribution corresponding to $r = 1$. We also note that the "negative binomial" label comes from the fact that the coefficient C_k^{r+k-1} in the equation of this distribution (see Eq. 82) is equal to the magnitude of C_k^{-r} (i.e. $C_k^{r+k-1} = \left|C_k^{-r}\right|$) which is the binomial coefficient in the binomial theorem expansion for negative integer powers (see Problem 46 of § 2.2).
• **Hypergeometric distribution**:[102] this distribution is similar to the binomial distribution but while the trials in the binomial are independent of each other and hence they have a fixed probability (i.e. they are Bernoulli trials), the trials in the hypergeometric are not independent (and hence they are not of Bernoulli type). In brief, the hypergeometric distribution represents the probability of the number of successes in a series of random trials in which objects are drawn *with no replacement* from a population (consisting of two types) and hence the probability of success (represented by one type) and failure (represented by the other type) in each draw depends on the outcomes of the previous trials. For example, if we draw consecutively with no replacement a number of balls from a box that contains given numbers of blue and red balls (where drawing blue is a success and drawing red is a failure) then it is obvious that the probability of drawing blue and red in each trial depends on the outcomes of the previous trials.[103] The mass function of this distribution is given by:

$$P(N,n,R,r) = \frac{C_r^R C_{n-r}^{N-R}}{C_n^N} \tag{84}$$

where N is the size of the population from which the choices are made, n is the number of trials made, R is the number of available successes within N, and r is the number of actual successes made in the trials.[104] The mean and variance of this distribution are given by:

$$\mu = \frac{nR}{N} \qquad\qquad V = n\left(\frac{R}{N}\right)\left(\frac{N-n}{N-1}\right)\left(1-\frac{R}{N}\right) \tag{85}$$

Problems

1. Justify Eq. 82.
   Answer: We have $k+r$ trials (since we have k failures and r successes) where the last trial is a definite

[102] The "hypergeometric" label comes from its moment generating function.
[103] For instance, if the box contains N balls (R blue and $N-R$ red) and we draw two balls then if the first draw is blue then the probability of drawing blue in the second draw is $(R-1)/(N-1)$, while if the first draw is red then the probability of drawing blue in the second draw is $R/(N-1)$. It is worth noting that being "consecutive" is for clarity rather than being a condition (i.e. this distribution applies even if the balls are drawn in one go noting that in this case the probability of the color of each ball in the selection depends on the color of the "previous" balls in the selection where "previous" here means "in consideration" rather than "in time").
[104] We note that all these numbers are finite. Moreover, $0 < R < N$ and $0 \leq r \leq R$ restricted by $r \leq n$. The readers are referred to Problem 6 of the present subsection and to Problem 2 of § 7.4 for examples of the hypergeometric distribution.

success. Hence, the number of permutations with repetitions (see Problem 30 of § 2.2) for the failures and successes (noting that the last trial is a definite success and hence we have k repetitive failures and $r-1$ repetitive successes that contribute to the permutations) is $C_{k,r-1}^{k+r-1} = C_k^{k+r-1}$. Moreover, the probability of each one of these permutations is $p^r(1-p)^k$ because we have r successes and k failures. Therefore, by the addition law of disjoint events (see Eq. 53) the total probability of r successes (each with probability p) and k failures (each with probability $1-p$) is $C_k^{k+r-1}\, p^r\, (1-p)^k$ which is what is given by Eq. 82.

PE: Justify the use of the addition law of disjoint events (as given by Eq. 53) in the above argument.

2. Show that the negative binomial distribution is normalized.

Answer: From Eq. 82 we have (where for brevity and clarity we use $q = 1-p$):

$$\begin{aligned}
\sum_{k=0}^{\infty} P(k,r,p) &= \sum_{k=0}^{\infty} C_k^{k+r-1}\, p^r\, q^k \\
&= p^r \sum_{k=0}^{\infty} C_k^{k+r-1}\, q^k && (r \text{ is fixed}) \\
&= p^r \sum_{k=0}^{\infty} (-1)^k C_k^{k+r-1}\, (-q)^k\ 1^{-r-k} \\
&= p^r\, (-q+1)^{-r} && (\text{Eq. } 35) \\
&= p^r\, \frac{1}{(1-q)^r} \\
&= p^r\, \frac{1}{p^r} && (1-q=p) \\
&= 1
\end{aligned}$$

PE: In the above derivation we treated r as fixed and k as variable. Justify this.

3. Justify Eq. 84.

Answer: Let justify this equation by using an example where we have a pack of N cards R of which are blue and the rest (i.e. $N-R$) are red. Now, if we draw n cards ($n \leq N$) then our chance of getting r blue cards ($r \leq R$) should be $C_r^R C_{n-r}^{N-R}/C_n^N$. This is because there are C_r^R ways for drawing r blues out of the available R blues, and there are C_{n-r}^{N-R} ways for drawing $(n-r)$ reds out of the available $(N-R)$ reds, and hence by the fundamental principle of counting (see § 2.2) there are $C_r^R C_{n-r}^{N-R}$ ways for drawing r blues and $(n-r)$ reds out of the available N cards. Moreover, we have a total of C_n^N ways for drawing n cards out of the available N cards. So, by the definition of probability (see § 3.2 and Eq. 38 in particular) the probability of getting n cards (r blue and $n-r$ red) out of the available N cards (R blue and $N-R$ red) should be $C_r^R C_{n-r}^{N-R}/C_n^N$ which is what is given by Eq. 84.

PE: Obtain an expression for $P(N,n,R,r)$ in terms of factorials.

4. Show that the hypergeometric distribution is normalized.

Answer: From Eq. 84 we have:

$$\sum_{r=0}^{n} P(N,n,R,r) = \sum_{r=0}^{n} \frac{C_r^R C_{n-r}^{N-R}}{C_n^N} = \frac{1}{C_n^N} \sum_{r=0}^{n} C_r^R C_{n-r}^{N-R} = \frac{1}{C_n^N}\, C_n^N = 1$$

where we used the Vandermonde identity (see part g of Problem 25 of § 2.2) in the third step.

PE: Investigate other methods for showing the normalization of the hypergeometric distribution.

5. Regarding the hypergeometric distribution, it is claimed that if n is negligible in comparison to N, R and $(N-R)$ then the hypergeometric distribution $P(N,n,R,r)$ can be replaced by the binomial distribution $P(n,r,p)$ where $p = R/N$. Justify this claim.

Answer: This claim sounds logical because the probability of success at a given trial is $(R-s)/(N-t)$ (where s and t represent respectively the number of successes and trials made before that trial) and

hence if n is negligible in comparison to R and N then s and t are also negligible in comparison to R and N and thus $(R-s)/(N-t)$ can be approximated by $p = R/N$ which is the (virtually) constant probability of success. Similarly, the probability of failure at a given trial is $(N-R-f)/(N-t)$ (where f and t represent respectively the number of failures and trials made before that trial) and hence if n is negligible in comparison to $(N-R)$ and N then f and t are also negligible in comparison to $(N-R)$ and N and thus $(N-R-f)/(N-t)$ can be approximated by $q = (N-R)/N$ which is the (virtually) constant probability of failure. Accordingly, the probability of both success and failure are (virtually) constant (where they add up to unity) and hence the trials are (virtually) independent of each other which means that they can be treated as Bernoulli trials (see the preamble of the present chapter) and thus their probability can be modeled by the binomial distribution as an approximation to the hypergeometric distribution.

Note: an obvious consequence of the above result is that when we sample from a very large population it does not make a difference (or rather considerable difference) whether we sample with or without replacement (as long as our sample is tiny in comparison to the population as given by the above conditions). This is because in both cases we can use the binomial distribution (i.e. as an exact model in the case of replacement and as an approximation to the hypergeometric in the case of no replacement). Also see Problem 4 of § 1.5.4. We also note that the binomial distribution is generally easier to evaluate than the hypergeometric distribution because the binomial has only one binomial coefficient while the hypergeometric has three.

PE: Create a table or a plot in which you compare (by an example) the binomial distribution to the hypergeometric distribution where the former can approximate the latter (i.e. by satisfying the above conditions).

6. The player in a game of chance draws successively and with no replacement 5 balls at random from a box containing 4 red balls and 6 blue balls. What is the probability of drawing $0, 1, 2, 3, 4$ red balls?[105]

   Answer: This is an instance of the hypergeometric distribution and hence we use Eq. 84 (noting that $N = 10$, $n = 5$, $R = 4$ and $r = 0, 1, 2, 3, 4$), that is:

$$\begin{aligned} P(10,5,4,0) &= C_0^4 C_5^6 / C_5^{10} = (1 \times 6)/252 \simeq 0.02381 \\ P(10,5,4,1) &= C_1^4 C_4^6 / C_5^{10} = (4 \times 15)/252 \simeq 0.23810 \\ P(10,5,4,2) &= C_2^4 C_3^6 / C_5^{10} = (6 \times 20)/252 \simeq 0.47619 \\ P(10,5,4,3) &= C_3^4 C_2^6 / C_5^{10} = (4 \times 15)/252 \simeq 0.23810 \\ P(10,5,4,4) &= C_4^4 C_1^6 / C_5^{10} = (1 \times 6)/252 \simeq 0.02381 \end{aligned}$$

   As we see, the sum of these probabilities is 1 as it should be.[106]

   PE: Repeat the Problem with $N = 15$, $n = 7$, $R = 3$ and $r = 0, 1, 2, 3$.

4.2 Probability Density Functions

As indicated earlier, probability density function is defined for continuous variables as a real function $f(x)$ that gives the probability value as a function of the (continuous) random variable x, i.e. $f(x_i)\,dx$ is the probability that x lies in the interval $x_i \leq x \leq x_i + dx$.[107] The two main properties of the probability density function (which are similar to corresponding properties of the probability mass function; see Eqs. 64 and 67) are:

$$f(x) \geq 0 \qquad (-\infty < x < +\infty) \tag{86}$$

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 \tag{87}$$

[105] We note that "successively and with no replacement" is for clarity rather than being a condition (i.e. it could be "at once" instead).

[106] We note that by convention $C_m^n = 0$ for $0 < n < m$ and hence $P(10,5,4,5) \equiv C_5^4 C_0^6 / C_5^{10} = 0$ since $C_5^4 = 0$.

[107] "Probability value" should be understood in this sense. More clarifications about this issue will follow.

The actual probabilities are obtained by taking the integral of the probability density function over a given range of the random variable (i.e. the range whose probability is required). For example, the probability of x being between -1 and $+1$ is given by:

$$P(-1 \leq x \leq +1) = \int_{-1}^{+1} f(x)\,dx \tag{88}$$

In the subsections of this section we investigate some of the well known and commonly used probability density functions. It is noteworthy that probability density functions have usually an advantage over the corresponding mass functions by being more manageable analytically and computationally since dealing with continuous variables (e.g. through differentiation and integration) is generally easier than dealing with discrete variables (e.g. through differencing and summation).

Problems

1. Justify the properties of the density function (as expressed by Eqs. 86 and 87).
   Answer:
   • Eq. 86 is justified by the axioms of probability (see axiom 1 of § 3.1) noting that the probability is given by $f(x)\,dx$ and $dx > 0$.
   • Eq. 87 is justified by the axioms of probability (see axiom 2 of § 3.1).
   PE: Compare between the properties of density functions (as expressed by Eqs. 86 and 87) and the corresponding properties of mass functions.
2. Make a comparison between mass function and density function.
   Answer: For example:
   • Mass function is discrete, while density function is continuous (and hence the normalization condition, for instance, is given as a sum in the case of mass function as seen in Eq. 67, and is given as an integral in the case of density function as seen in Eq. 87).
   • Mass function is a "probability" function (and hence it gives probability directly), while density function is a "probability rate" function (and hence it gives probability indirectly). This can be seen clearly from the fact that the probability of x_i in the discrete case is given by $P(x_i)$ while the probability of x_i in the continuous case is given by $f(x_i)\,dx$ and not by $f(x_i)$.[108] In fact, this should (partly) explain our use of different symbols for mass and density functions, i.e. we use P for mass function and f for density function.
   • The values of mass function cannot exceed one while the values of density function can exceed one, i.e. it is impossible to have $P(x_i) > 1$ since $P(x_i) \leq 1$ but it is possible to have $f(x_i) > 1$.
   PE: Justify (in technical terms) the last property given in the answer of this Problem (by linking it to the other given properties).
3. Give some examples of functions that can represent probability density (i.e. they meet the conditions of Eqs. 86 and 87).
   Answer: The following functions can represent probability density because they meet the conditions of Eqs. 86 and 87.
   (**a**) $f(x) = \frac{x}{8} \quad (0 \leq x \leq 4), \quad f(x) = 0 \quad (x < 0 \text{ and } x > 4)$.
   (**b**) $f(x) = \frac{\operatorname{sech}(x)}{\pi} \quad (-\infty < x < +\infty)$.
   (**c**) $f(x) = \frac{1}{20} \quad (-10 \leq x \leq 10), \quad f(x) = 0 \quad (x < -10 \text{ and } x > 10)$.
   (**d**) $f(x) = \frac{e^{-x^2}}{\sqrt{\pi}} \quad (-\infty < x < +\infty)$.
   PE: Give more examples like the given ones.
4. Which of the following can be a probability density function:
   (**a**) $f(x) = \frac{3(1-x^2)}{4} \quad (-1 \leq x \leq 1), \quad f(x) = 0 \quad (x < -1 \text{ and } x > 1)$.
   (**b**) $f(x) = \frac{e^{-x^4}}{\pi} \quad (-\infty \leq x \leq +\infty)$.
   (**c**) $f(x) = \frac{\cos(x)}{2} \quad (-\frac{\pi}{2} \leq x \leq \frac{\pi}{2}), \quad f(x) = 0 \quad (x < -\frac{\pi}{2} \text{ and } x > \frac{\pi}{2})$.

[108] The readers are referred to Problem 1 of § 4.3.2 for further clarifications about this issue.

(**d**) $f(x) = \frac{2x}{7} \quad (-3 \le x \le 4), \quad f(x) = 0 \quad (x < -3 \text{ and } x > 4)$.

Answer:
(**a**) It can, because it satisfies the conditions of Eqs. 86 and 87.
(**b**) It cannot, because it does not satisfy the condition of Eq. 87.
(**c**) It can, because it satisfies the conditions of Eqs. 86 and 87.
(**d**) It cannot, because it does not satisfy the condition of Eq. 86.

PE: Verify the answer of this Problem by verifying the conditions of Eqs. 86 and 87 in each case.

5. Find the constant c in the following probability density functions:

(**a**) $f(x) = ce^{-2x} \quad (0 \le x < \infty)$. (**b**) $f(x) = 0.125x + c \quad (1 \le x \le 3)$.

(**c**) $f(x) = cx^2 \quad (-1 \le x \le +1)$. (**d**) $f(x) = c\ln(x) \quad (1 \le x \le 7)$.

Answer: Any probability density function must satisfy Eq. 87 (as well as Eq. 86) and this can be used (i.e. by integrating the function over its entire range and equating the result to 1) to infer the value of c in each case, as follows:

(**a**) $\int_0^{+\infty} ce^{-2x}\,dx = 1 \quad \rightarrow \quad \frac{c}{2} = 1 \quad \rightarrow \quad c = 2.$

(**b**) $\int_1^3 (0.125x + c)\,dx = 1 \quad \rightarrow \quad \frac{4c+1}{2} = 1 \quad \rightarrow \quad c = \frac{1}{4}$

(**c**) $\int_{-1}^{+1} cx^2\,dx = 1 \quad \rightarrow \quad \frac{2c}{3} = 1 \quad \rightarrow \quad c = \frac{3}{2}$

(**d**) $\int_1^7 c\ln(x)\,dx = 1 \quad \rightarrow \quad \big[7\ln(7) - 6\big]c = 1 \quad \rightarrow \quad c = \frac{1}{7\ln(7)-6}$

PE: Find the constant c in the following probability density functions:

(**a**) $f(x) = ce^{-3x} \quad (0 \le x < \infty)$. (**b**) $f(x) = 0.5x + c \quad (1 \le x \le 3)$.

(**c**) $f(x) = cx^4 \quad (-2 \le x \le +2)$. (**d**) $f(x) = c\ln(2x) \quad (1 \le x \le 5)$.

4.2.1 Uniform Distribution

The continuous uniform distribution (which is the simplest density function) is given by:

$$f(x) = \begin{cases} \frac{1}{b-a} & (a \le x \le b) \\ 0 & (\text{otherwise}) \end{cases} \tag{89}$$

where a and b are given real constants. For example, if a particle is moving uniformly on a unit circle then the probability of its position on the perimeter (i.e. the probability of finding the particle in the interval 0 and 2π at a randomly selected instant) is subject to this distribution. For the continuous uniform distribution the mean μ and the variance V are given by (see Problem 10 of § 5.1 and Problem 7 of § 5.2):

$$\mu = \frac{a+b}{2} \qquad V = \frac{(b-a)^2}{12} \tag{90}$$

Problems

1. Give some examples of the continuous uniform distribution.

Answer:
• Having a number between 0 and 2π (representing angle in radian) when spinning a (fair) roulette wheel.
• Getting a number between 4π and 10π (representing perimeter of circle) when randomly plotting a circle of radius $2 \le r \le 5$.

Note: the continuous uniform distribution is rather artificial and hence it is difficult to find real life examples of this distribution that are really and exactly uniform. However, this generally applies to most, if not all, distributions since they are essentially idealizations of real life distributions which

involve many intricate factors and complexities. Also see Problem 3.
PE: Give more examples for the continuous uniform distribution.

2. Show that the continuous uniform distribution is normalized.
Answer: From Eq. 89 we have:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \left[\frac{x}{b-a}\right]_a^b = \frac{b-a}{b-a} = 1$$

PE: Explain and justify all the details of this derivation.

3. Let assume that we have a random number generator that can generate real numbers in the interval $[0, 1]$.[109] What is the probability that the generated numbers are between 0.6 and 0.95?
Answer: Since these numbers are randomly generated, we can assume that they are equally likely and hence the generated numbers are uniformly distributed. Accordingly, the probability that the generated numbers are between 0.6 and 0.95 should be $0.95 - 0.6 = 0.35$.
PE: Create a similar Problem in which coupled (real) random number generators are supposedly used to select points in a rectangle.

4.2.2 Normal Distribution

The normal or Gaussian distribution (which is a function of x and is parameterized by μ and V) is given by:

$$f(x) = f(x, \mu, V) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}} \qquad (-\infty < x < \infty) \tag{91}$$

This distribution is possibly the most important probability distribution (at least for the continuous probability distributions). This is not only because of its beneficial characteristics and its validity as an approximate model for many physical and non-physical phenomena, but also because its ability to represent and replace other distributions (e.g. binomial and Poisson) approximately in many cases and circumstances since it represents (under certain conditions) the continuous limit of these distributions. By definition, the mean of the normal distribution is μ and the variance is V (see Problem 10 of § 5.1 and Problem 7 of § 5.2).

It is important to note the following about the normal distribution:

• The **standard normal distribution** is a normal distribution with $\mu = 0$ and $V = 1$, i.e.

$$f_s(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{92}$$

where the subscript s refers to "standard".[110] This distribution was exceptionally important in the old days where calculations were made with the help of tables in which the values of this standard distribution are listed and used (with translation and scaling) to calculate normal distributions of various shapes and forms. However, with the wide availability of computing equipment and increasing reliance on them these days the importance of this standard distribution is reduced.

• The normal distribution can be used (and is used) to approximate other distributions under certain conditions. In fact,, according to the central limit theorem (refer to § 6.2.3) all well-behaved probability distributions converge to the normal distribution under certain conditions. For example, the normal distribution can be used (and is widely used) as an approximation to the corresponding binomial distribution, i.e. with the mean and variance of the normal distribution being given by those of the binomial distribution, that is: $\mu = np$ and $V = np(1-p)$.[111] In fact, this approximation is reliable in most

[109] In fact, random number generators do not really generate real numbers.

[110] We note that the standard normal distribution is usually given as a function of $z = (x-\mu)/\sqrt{V}$ where μ and V are the mean and variance of the original normal distribution which is standardized through this transformation and x is the random variable of the original distribution.

[111] We note that the normal distribution is a limiting case for the binomial distribution when $n \to \infty$ and p stays finite (so that $np \to \infty$).

cases. Moreover, the approximation becomes more reliable when n is large or/and p is close to 1/2, i.e. it generally improves with increasing n and with p approaching 1/2 (see Problems 7 and 8).

Problems

1. Give some examples of the normal distribution.
   Answer: It is difficult (if not impossible) to find a real life example that is normally distributed exactly. However, many real life random variables are normally distributed approximately, and the approximation in many cases is very good (noting as well that the normal distribution can be used as an approximation to other distributions in certain circumstances, as indicated above). Examples of such variables are the weight of animals of certain species living in a given area, the height of people in a city, the weight of newborn babies (of a given ethnicity and location), and the lifetime of a certain type of electronic components and devices (operated under certain physical conditions). Also see § 6.2.2 and § 6.2.3.
   PE: Investigate applications of normal distribution in physical and social sciences and find more real life examples of this distribution.
2. Calculate the following probability densities of the normal distribution given in the form $f(x, \mu, V)$:
   (**a**) $f(-33.6, -81.9, 3.1)$. (**b**) $f(167.8, 161.8, 11.6)$. (**c**) $f(0.0, 0.0, 0.09)$.
   (**d**) $f(-43.8, 51.6, 0.1)$. (**e**) $f(25.9, -31.8, 43.2)$. (**f**) $f(199.2, -3.8, 13.7)$.
   Answer: From Eq. 91 we have:

$$\begin{aligned}
&(\mathbf{a}) \quad f(-33.6,-81.9,3.1) &&= \frac{1}{\sqrt{6.2\pi}} e^{-\frac{48.3^2}{6.2}} &&\simeq 8.752331 \times 10^{-165}\\
&(\mathbf{b}) \quad f(167.8,161.8,11.6) &&= \frac{1}{\sqrt{23.2\pi}} e^{-\frac{6^2}{23.2}} &&\simeq 2.481852 \times 10^{-2}\\
&(\mathbf{c}) \quad f(0.0,0.0,0.09) &&= \frac{1}{\sqrt{0.18\pi}} e^{0} &&\simeq 1.32980760134\\
&(\mathbf{d}) \quad f(-43.8,51.6,0.1) &&= \frac{1}{\sqrt{0.2\pi}} e^{-\frac{(-95.4)^2}{0.2}} &&\simeq 1.524318 \times 10^{-19763}\\
&(\mathbf{e}) \quad f(25.9,-31.8,43.2) &&= \frac{1}{\sqrt{86.4\pi}} e^{-\frac{57.7^2}{86.4}} &&\simeq 1.117645 \times 10^{-18}\\
&(\mathbf{f}) \quad f(199.2,-3.8,13.7) &&= \frac{1}{\sqrt{27.4\pi}} e^{-\frac{203^2}{27.4}} &&\simeq 7.297248 \times 10^{-655}
\end{aligned}$$

   PE: Repeat the Problem for the following:
   (**a**) $f(52.5, 49.4, 44.7)$. (**b**) $f(-152.5, -177.5, 55.8)$. (**c**) $f(559.6, 638.9, 215.2)$.
   (**d**) $f(-69.2, -69.3, 0.01)$. (**e**) $f(15.9, 78.2, 1.6)$. (**f**) $f(69.3, 0.0, 2.8)$.
3. Outline the main properties of the normal distribution as represented by Eq. 91.
   Answer:
   • It peaks at $x = \mu$.
   • It is symmetric with respect to the vertical line $x = \mu$.
   • It is normalized to unity, i.e. its integral between $-\infty$ and $+\infty$ equals 1.
   • Its graph is bell-shaped.
   • Its mean is μ and its variance is V (see Eq. 91).
   • It is used as an approximation to other distributions (such as binomial and Poisson) in some limiting cases.
   PE: Discuss the significance of the above properties. Also add more properties if you can.
4. Show that the normal distribution is normalized.
   Answer: From Eq. 91 we have:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}}\,dx = \frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{+\infty} e^{-\frac{(x-\mu)^2}{2V}} = \frac{\sqrt{2\pi V}}{\sqrt{2\pi V}} = 1$$

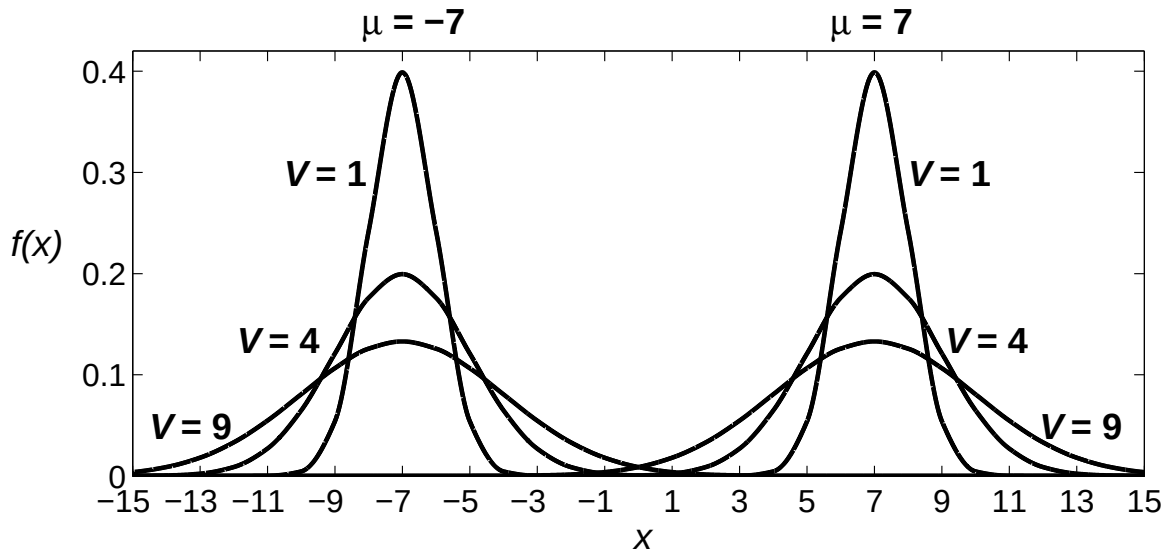


Figure 23: See Problem 5 of § 4.2.2. The numbers on the top (i.e. $\mu = -7$ and $\mu = 7$) belong to the profiles beneath them.

where we used standard integration techniques in the third step.[112]
PE: Explain and justify all the details of this derivation.

5. What is the effect of varying μ and V on the position and shape of (the curve representing) the normal distribution?
Answer: Noting that for the normal distribution μ represents the mean and V represents the variance, increasing μ results in shifting the peak of (the curve representing) the distribution to the right, while increasing V results in flattening (the curve representing) the distribution.[113] See Figure 23 where we plotted the normal distribution for $\mu = -7$ and $\mu = 7$ each with $V = 1, 4, 9$.
PE: Plot the normal distribution for a number of μ's and V's (similar to Figure 23) and hence confirm the observations that we made about the effect of varying μ and V on the position and shape of the distribution.
6. Show that in the normal distribution:
(**a**) About 68% of the values are within $\sqrt{V}$ from the mean.[114]
(**b**) About 95% of the values are within $2\sqrt{V}$ from the mean.
(**c**) About 99.7% of the values are within $3\sqrt{V}$ from the mean.
Answer: All these results can be obtained simply by integration using the standard methods of calculus, that is:

(**a**) $$\int_{\mu-\sqrt{V}}^{\mu+\sqrt{V}} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}} \simeq 0.6827$$

(**b**) $$\int_{\mu-2\sqrt{V}}^{\mu+2\sqrt{V}} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}} \simeq 0.9545$$

(**c**) $$\int_{\mu-3\sqrt{V}}^{\mu+3\sqrt{V}} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}} \simeq 0.9973$$

[112] In fact, we used the error function.
[113] "Flattening" means increasing the width of the profile and hence lowering the height of the peak (due to normalization).
[114] "The values are within" means the probability of getting the random variable within the given interval (which is determined by the limits of the integrals in the upcoming answer).

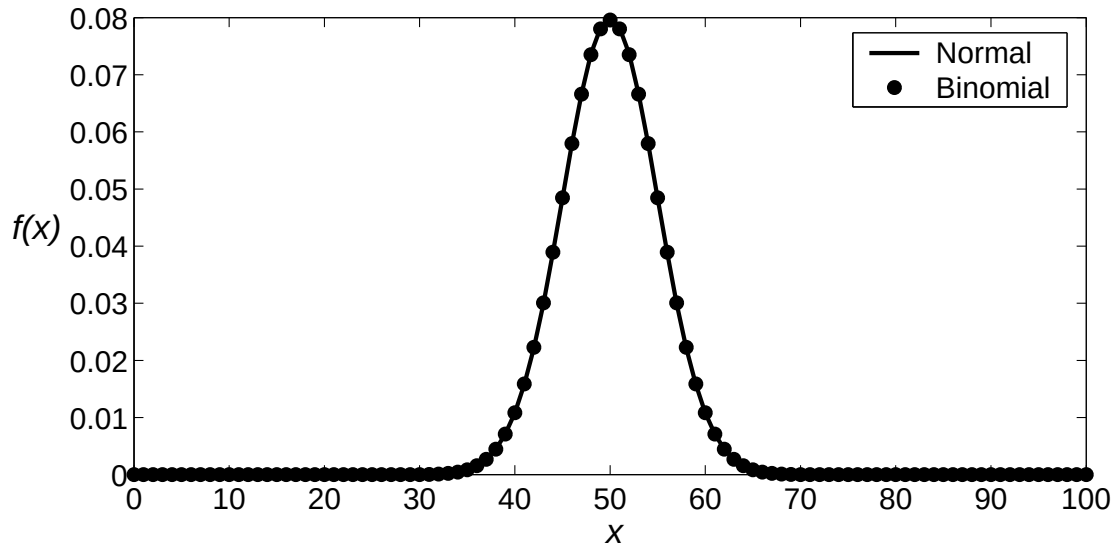


Figure 24: The plot of the normal distribution corresponding to the binomial distribution with $n = 100$ and $p = 1/2$. See Problem 7 of § 4.2.2.

PE: Show more details about these integrations (using calculus books or integral calculators if necessary).

7. Compare the normal distribution to the binomial distribution for the case of normal distribution corresponding to the binomial distribution with $n = 100$ and $p = 1/2$.
   Answer: For the binomial distribution with $n = 100$ and $p = 1/2$ we have $\mu = 50$ and $V = 25$ (see Eq. 72), and hence the normal distribution that corresponds to this binomial distribution is (see Eq. 91):
   $$f(x) = \frac{1}{\sqrt{50\pi}} e^{-\frac{(x-50)^2}{50}}$$
   We plotted this $f(x)$ alongside the corresponding binomial points on the same graph (see Figure 24). As we see, the two are almost identical (noting their different nature as continuous and discrete). Also see Problem 8.
   PE: Explain why for the normal distribution to be more reliable as an approximation to the corresponding binomial distribution we should have n relatively large or/and p close to $1/2$. Try to create plots like Figure 24 (using for example spreadsheets) in which you vary n and p so that you can see graphically the effect of varying these parameters on the quality of approximation.
8. Make plots of the binomial distribution for $n = 140$ with $p = 0.05, 0.20, 0.35, 0.50, 0.70, 0.90$ and their corresponding normal distribution.
   Answer: See Figure 25.
   PE: Do the following:
   • Comment on the quality of agreement between the binomial distribution and the corresponding normal distribution in all cases.
   • Comment on the effect of varying p on the quality of agreement.
   • Investigate the effect of varying n on the quality of agreement.
   • Justify the shift of the peak to the right as p increases.
   • Explain the observed change of the profile (i.e. the variation of the width of the profile and the height of its peak) as a result of varying p (where the width/height increases/decreases first then it decreases/increases afterward).
   • Should we expect the agreement between the binomial distribution and the corresponding normal distribution to deteriorate as $V \to 0$ (see Problem 2 of § 4.2)?

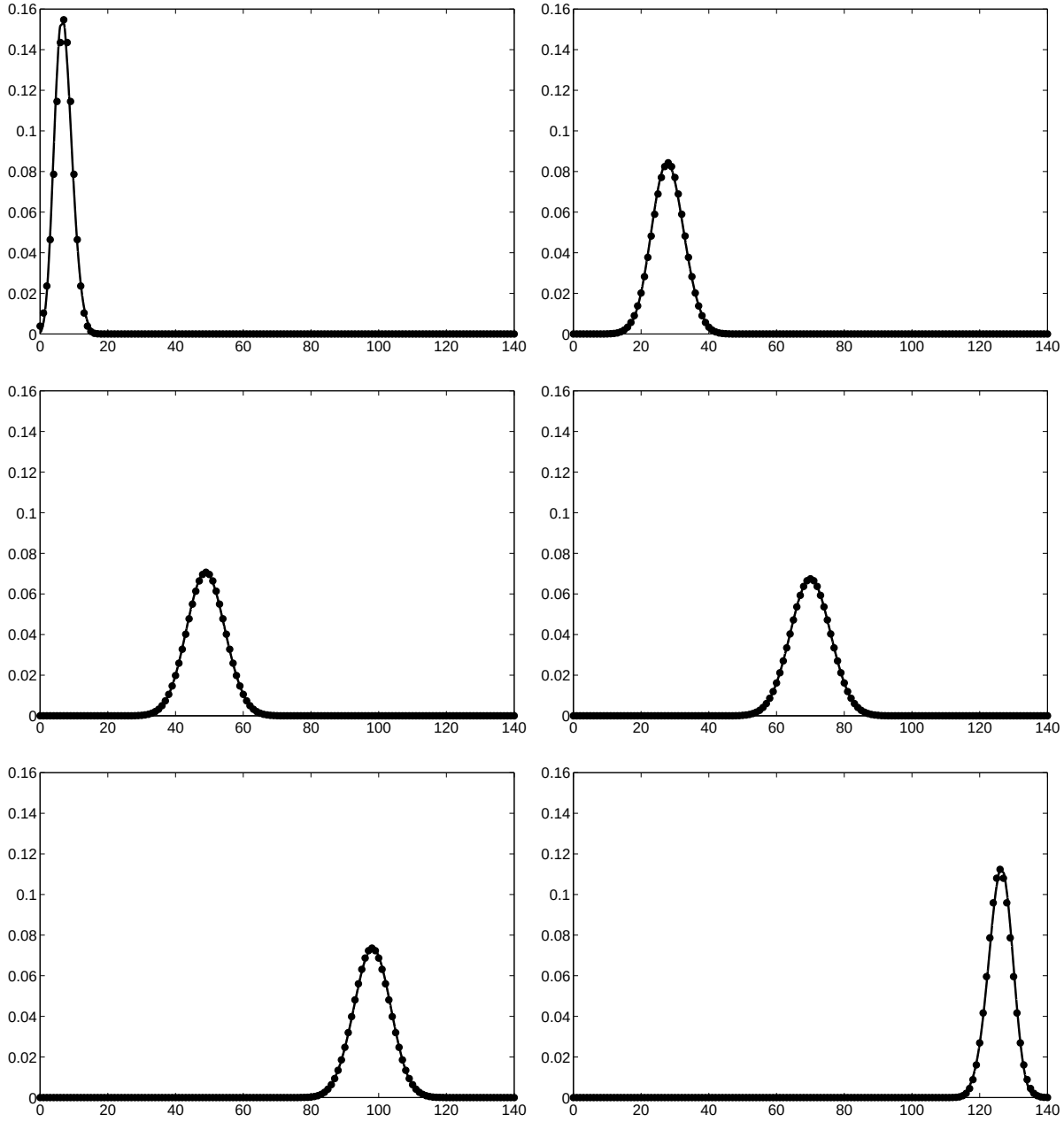


Figure 25: Comparison between the binomial distribution (filled circles) and the corresponding normal distribution (solid curve) for $n = 140$ with $p = 0.05$ (top left), $p = 0.20$ (top right), $p = 0.35$ (middle left), $p = 0.50$ (middle right), $p = 0.70$ (bottom left) and $p = 0.90$ (bottom right). For the corresponding normal distribution we have $\mu = np$ and $V = np(1-p)$ in each case. The horizontal axis in each frame represents x (corresponding to k of the binomial distribution) and the vertical axis represents the probability density $f(x)$ $\big[$corresponding to the probability $P(k)$ of the binomial distribution$\big]$. See Problem 8 of § 4.2.2.

9. Write a simple program that calculates the probability densities of the normal distribution.
   Answer: See the NormalDensity.cpp code which calculates the individual values of probability density of the normal distribution.
   Note: to do extensive calculations of normal distribution (i.e. on discrete points over a certain range of x corresponding for instance to the range of k of a corresponding discrete distribution) conveniently, we wrote another code (see the NormalDistribution.cpp code) in which we put the core of the NormalDensity.cpp code into a k loop and output the results to a file. This code is especially useful for doing extensive calculations in extreme cases (such as those cases involving very large or/and very small numbers).
   PE: Why the NormalDensity.cpp code (as well as the NormalDistribution.cpp code) may be needed to calculate the probability densities of the normal distribution despite the fact that the probability densities of this distribution are easy to calculate by a handheld calculator or a spreadsheet?
10. Compare the binomial distribution (with $n = 1000$ and $p = 0.1, 0.5, 0.9$) to the corresponding Poisson and normal distributions by plotting them on the same figure.
    Answer: We made this comparison in Figure 26.[115] As we see, for this case (in which n is large) the Poisson distribution is fairly close to the binomial distribution at low p (i.e. $p = 0.1$) but it differs considerably when p increases to 0.5 and the difference is exacerbated by increasing p to 0.9.[116] On the other hand, the normal distribution is almost identical to the binomial distribution in all cases (i.e. the cases of $p = 0.1, 0.5, 0.9$) although the agreement between the two distributions is best in the case of $p = 0.5$.
    PE: Repeat the Problem for $n = 2000$ and $p = 0.05, 0.5, 0.95$.

4.2.3 Exponential Distribution

The exponential distribution (which is a function of x and is parameterized by α) is given by:

$$f(x) = f(x, \alpha) = \alpha e^{-\alpha x} \qquad (0 < x < \infty, \alpha > 0) \tag{93}$$

The mean μ and the variance V of the exponential distribution are given by (see Problem 10 of § 5.1 and Problem 7 of § 5.2):

$$\mu = \frac{1}{\alpha} \qquad V = \frac{1}{\alpha^2} \tag{94}$$

It is important to note the following about the exponential distribution:
• The exponential distribution is usually used to model the distribution of waiting time between consecutive events of Poisson type (or lifetime of processes and events of this type).
• The exponential distribution is seen as the continuous limit (or counterpart) of the discrete geometric distribution (see § 4.1.5) and hence they are distinguished by certain properties (such as being memoryless or being models for waiting times).

Problems

1. Give some examples of the exponential distribution.
   Answer:
   • The distribution of waiting time between consecutive tornadoes (where tornadoes usually happen).
   • The distribution of waiting time between consecutive earthquakes (where earthquakes usually happen).
   PE: Give more examples of the exponential distribution (e.g. from atomic transitions and radioactivity).

[115] We note that the calculations are performed using our codes BinomialDistribution.cpp, PoissonDistribution.cpp and NormalDistribution.cpp.

[116] We should remember (see § 4.1.4) that for the Poisson distribution to be a limiting case (and hence a good approximation) to the binomial distribution we should have $p \to 0$ (as well as $n \to \infty$).

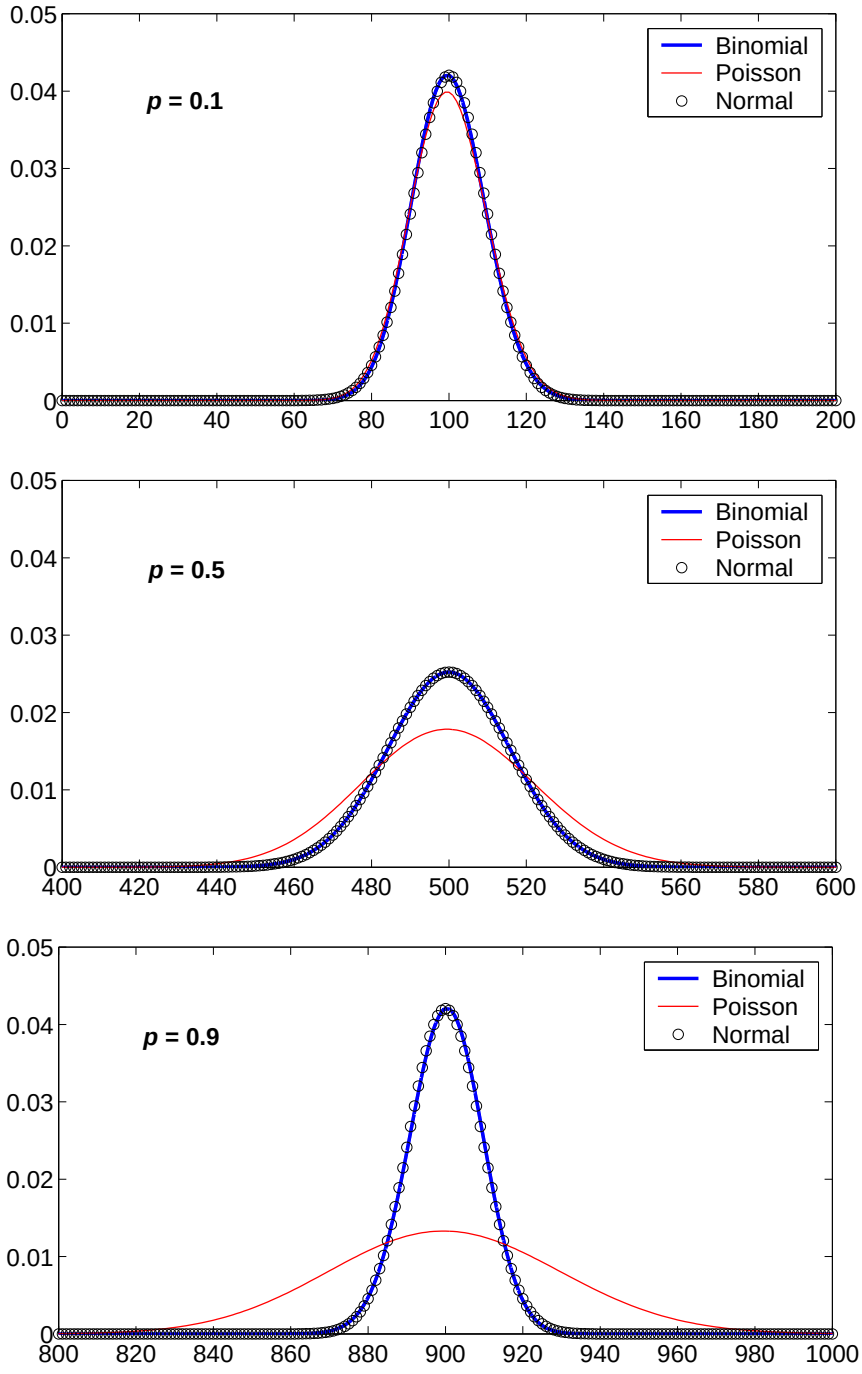


Figure 26: Comparison between the binomial distribution (with $n = 1000$ and $p = 0.1, 0.5, 0.9$) to the corresponding Poisson and normal distributions. For the corresponding Poisson distribution we use $\lambda = np$ and for the corresponding normal distribution we use $\mu = np$ and $V = np(1-p)$. The horizontal axis in each frame represents k (for binomial and Poisson which corresponds to x for normal) and the vertical axis represents the probability $P(k)$ $\big[$or the density $f(x)\big]$. For clarity (as well as other practical reasons), we use continuous curves to represent discrete distributions (and vice versa). See Problem 10 of § 4.2.2.

2. Show that the exponential distribution is normalized.
   Answer: From Eq. 93 we have:
$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{0}^{\infty} \alpha e^{-\alpha x}\,dx = \alpha\left[\frac{e^{-\alpha x}}{-\alpha}\right]_0^{\infty} = \alpha\left[\frac{0}{-\alpha} - \frac{1}{-\alpha}\right]_0^{\infty} = \frac{\alpha}{\alpha} = 1$$
   PE: Explain and justify all the details of this derivation.
3. Justify Eq. 93 (assuming the exponential distribution to be a model for the distribution of time periods between consecutive Poisson-like events).
   Answer: If the average number of events happening in a unit interval is α then in an interval of size x we should have (on average) αx events. Now, if we assume Poisson-like events then we should have (in the corresponding Poisson distribution) $\lambda = \alpha x$ and $k = 0$ (since in the interval x no event happens because x is supposedly an interval between events). Hence, from Eq. 75 we get:
$$P(0) = \frac{(\alpha x)^0\, e^{-\alpha x}}{0!} = e^{-\alpha x}$$
   Now, to get $f(x)$ for the exponential distribution from the $P(0)$ of the corresponding Poisson distribution we argue that the probability of an event occurring in an infinitesimal time interval $d(\alpha x)$ (corresponding to αx in $e^{-\alpha x}$) is:[117]
$$dF = P(0) \times d(\alpha x) = P(0) \times \alpha\, dx = \alpha e^{-\alpha x} dx$$
   Thus:[118]
$$f(x) = \frac{dF}{dx} = \alpha e^{-\alpha x}$$
   PE: Can the (continuous) exponential distribution be considered a limiting case to the (discrete) Poisson distribution (as the continuous normal distribution is considered a limiting case to the discrete Poisson distribution for instance)? Justify your answer.
4. What is the effect of varying α on the shape of the exponential distribution?
   Answer: Noting that $e^0 = 1$ and $\alpha > 0$, we have $f(x = 0) = \alpha$ which means that α is the y-intercept on the (positive) y axis and hence increasing/decreasing α results in raising/lowering the start point of the curve representing this distribution (although this point is not included in the distribution). Also, α is the rate of drop of the exponential function (i.e. how fast it decreases noting that $x > 0$) and hence increasing/decreasing α results in increasing/decreasing this rate of drop. Both these facts are inline with the normalization of the distribution (since the area under the curve should be constant, i.e. equal 1). In Figure 27. we plotted this distribution for a number of α's where these features can be seen clearly.
   PE: It is claimed (as indicated earlier) that the exponential distribution is the continuous counterpart of the discrete geometric distribution. Justify this claim by comparing Eq. 80 to Eq. 93 and comparing Figure 22 to Figure 27. Also explain why $f(x)$ of the exponential distribution can exceed 1 while $P(k)$ of the geometric distribution cannot.

4.2.4 Other Continuous Distributions

There are many other continuous distributions some of which are outlined in the following points:
• **Gamma distribution**: this is a generalization of the exponential distribution which we investigated earlier (see § 4.2.3), and hence it represents the distribution of time intervals (i.e. waiting time) before the r^{th} event in a Poisson-type series of events.[119] The gamma distribution (which is a function of x

[117] We use F here to indicate that this is actually a cumulative probability function (which will be investigated later; see § 4.3).
[118] The reader is referred to Eq. 101 which will be investigated later.
[119] We note that in the exponential distribution the waiting time is before the occurrence of the 1^{st} event. Hence, the gamma distribution (as represented by Eq. 95) reduces to the exponential distribution (as represented by Eq. 93) for $r = 1$. Similarly, Eq. 96 reduces to Eq. 94 for $r = 1$.

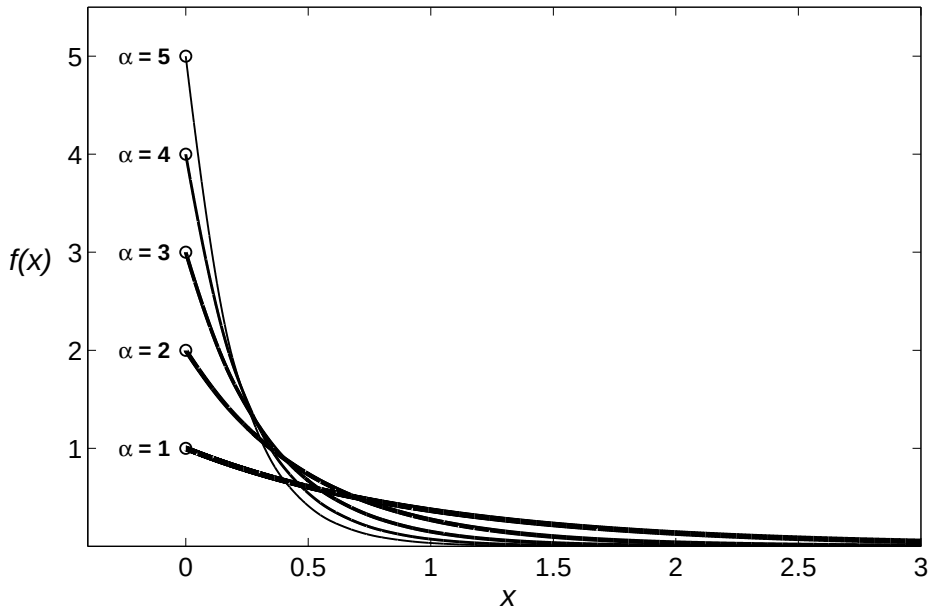


Figure 27: The plot of Problem 4 of § 4.2.3. The circles indicate that $x = 0$ is not included.

and is parameterized by α and r) is given by:[120]

$$f(x) = f(x, \alpha, r) = \frac{\alpha^r\, x^{r-1}}{(r-1)!}\, e^{-\alpha x} \qquad (\alpha > 0, r \in \mathbb{N}, 0 < x < \infty) \tag{95}$$

The mean μ and the variance V of the gamma distribution are given by:

$$\mu = \frac{r}{\alpha} \qquad V = \frac{r}{\alpha^2} \tag{96}$$

• **Cauchy distribution**: this distribution (which may also be called Lorentz distribution or Cauchy-Lorentz distribution) is given by:[121]

$$f(x) = \frac{1}{\pi\,(1 + x^2)} \qquad (-\infty < x < \infty) \tag{97}$$

The mean and variance of this distribution are not defined and hence it is described as pathological. In fact, this distribution is used in the literature as a typical example of pathological distributions. However, in our view it is sensible for the Cauchy distribution (and its alike) to have a mean despite the fact that the integral defining its mean (i.e. Eq. 116) is divergent. This is because the mean can be inferred from the symmetry of the function representing this distribution with respect to $x = 0$ and hence its mean is 0 by the rule of symmetry which will be investigated later (see the notes of § 5.1 as well as Problem 12 of § 5.1). Moreover, we can use the Cauchy principal value to represent the value of the integral defining its

[120] If we note that for all positive integers r we have $\Gamma(r) = (r-1)!$ then this equation can be written as:

$$f(x) = \frac{\alpha^r\, x^{r-1}}{\Gamma(r)}\, e^{-\alpha x}$$

where Γ is the gamma function (and that is why it is called "gamma distribution" of order r). In fact, the gamma distribution can be extended to all positive (real) numbers r (but we do not follow this issue in this book).

[121] In fact, this form is a special case of the more general form of the Cauchy distribution and hence it may be called the standard Cauchy distribution. We also note that the Cauchy distribution is a special case of the Breit-Wigner distribution (which is commonly used in quantum physics).

mean. So we can say: despite the fact that the integral defining the mean of the Cauchy distribution (and its alike) is not defined (and hence the mean of this distribution is not defined formally), this distribution does have a "non-formal" mean which is sensible and useful to accept and use in many theoretical and practical situations.

Problems

1. Give some examples of the gamma distribution.
   Answer: Just generalize the examples of Problem 1 of § 4.2.3 (i.e. the waiting time before the occurrence of the r^{th} tornado or earthquake).
   PE: Why it is "the waiting time before the occurrence of the r^{th}" rather than the $(r-1)^{th}$?
2. Justify Eq. 95.
   Answer: If we follow the method of Problem 3 of § 4.2.3 (noting that the gamma distribution is a generalization of the exponential distribution with $k = r-1$ replacing $k = 0$) then we get:

$$\begin{aligned} P(r-1) &= \frac{(\alpha x)^{r-1}\, e^{-\alpha x}}{(r-1)!} \\ dF &= P(r-1) \times d(\alpha x) = P(r-1) \times \alpha\, dx = \alpha\, \frac{(\alpha x)^{r-1}\, e^{-\alpha x}}{(r-1)!}\, dx = \frac{\alpha^r\, x^{r-1}\, e^{-\alpha x}}{(r-1)!}\, dx \\ f(x) &= \frac{dF}{dx} = \frac{\alpha^r\, x^{r-1}}{(r-1)!}\, e^{-\alpha x} \end{aligned}$$

   PE: Reproduce the above argument independently, i.e. with no consideration of the argument of Problem 3 of § 4.2.3.
3. Show that the gamma distribution is normalized.
   Answer: We have:

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)\, dx &= \int_0^{\infty} \frac{\alpha^r\, x^{r-1}}{(r-1)!}\, e^{-\alpha x}\, dx && \text{(Eq. 95)} \\ &= \frac{\alpha^r}{(r-1)!} \int_0^{\infty} x^{r-1}\, e^{-\alpha x}\, dx && (\alpha \text{ and } r \text{ are constants}) \\ &= \frac{\alpha^r}{(r-1)!} \int_0^{\infty} \left(\frac{y}{\alpha}\right)^{r-1} e^{-y}\, d\left(\frac{y}{\alpha}\right) && \text{(substitution of } y = \alpha x) \\ &= \frac{\alpha^r}{(r-1)!} \left(\frac{1}{\alpha}\right)^{r-1} \left(\frac{1}{\alpha}\right) \int_0^{\infty} y^{r-1}\, e^{-y}\, dy \\ &= \frac{1}{(r-1)!} \int_0^{\infty} y^{r-1}\, e^{-y}\, dy \\ &= \frac{1}{(r-1)!}\, \Gamma(r) && \text{(gamma function identity)} \\ &= \frac{1}{(r-1)!}\, (r-1)! && \left[\Gamma(r) = (r-1)! \text{ since } r \in \mathbb{N}\right] \\ &= 1 \end{aligned}$$

   PE: Why we treated α and r as constants?
4. Show that the Cauchy distribution is normalized.
   Answer: We have (see Eq. 97):

$$\int_{-\infty}^{+\infty} f(x)\, dx = \int_{-\infty}^{+\infty} \frac{1}{\pi\,(1+x^2)}\, dx = \left[\frac{\arctan(x)}{\pi}\right]_{-\infty}^{+\infty} = \frac{(\pi/2) - (-\pi/2)}{\pi} = \frac{\pi}{\pi} = 1$$

   PE: Why a distribution function can be accepted for representing probability even if its mean and variance are not defined, but it cannot be accepted if it is not normalized?

4.3 Cumulative Distribution Functions

Cumulative distribution function was defined descriptively in the preamble of this chapter. Mathematically, it can be defined by the following equation:

$$F(x_i) = P(x \leq x_i) \tag{98}$$

where F is the cumulative distribution function of the random variable x as a function of x_i (which represents a given value of the random variable). Cumulative distribution function is used to obtain the probability of the random variable to be in a given range of values (see Problem 1 of the present section as well as Problem 3 of § 4). As we will see (refer to Eqs. 99 and 100), cumulative distribution function represents the sum of mass function (for discrete random variables) or the integral of density function (for continuous random variables) and hence it is normalized to unity (noting that mass function and density function must be normalized to unity), i.e. we have $F(\infty) = 1$.

Problems

1. Give some examples of how the probability of the random variable to be in a given range can be obtained from the cumulative distribution function.
   Answer: For example (see Eq. 98):

$$\begin{aligned} P(x_i < x \leq x_j) &= F(x_j) - F(x_i) \\ P(x > x_i) &= 1 - F(x_i) \end{aligned}$$

   Similar intuitive relations can be easily obtained.
   PE: Give more examples (like the two given in the answer).[122]
2. On which law of probability the cumulative distribution function is ultimately based?
   Answer: It is the addition law of probability for mutually exclusive events.
   PE: Justify the given answer (considering both cases of discrete and continuous random variables).
3. Outline some of the properties of cumulative distribution function $F(x)$.
   Answer: For example:
   • $F(x)$ is a non-decreasing function of x.
   • $F(x)$ is bounded from both sides, i.e. $0 \leq F(x) \leq 1$ $\big[$where $F(-\infty) = 0$ and $F(\infty) = 1\big]$.
   • $F(x)$ is a sum of the probability mass function and an integral of the probability density function.
   PE: Try to find other properties of cumulative distribution function.

4.3.1 Discrete Random Variables

For discrete random variable, the cumulative distribution function is given by:

$$F(x_i) = \sum_{k \leq i} P(x_k) \tag{99}$$

This equation reflects the relationship between the probability mass function P and the cumulative distribution function F. As we see, we obtain $F(x_i)$ by adding the probabilities of all the values of the random variable less than or equal to x_i.[123]

Problems

1. Referring to Problem 5 of § 3.4, create a table representing the cumulative distribution function that corresponds to Table 4 (which represents a probability mass function).
   Answer: See Table 6.
   PE: Obtain from Table 6 the following probabilities (with x representing sum):

[122] The reader should consider $P(x_i < x < x_j)$, $P(x_i \leq x < x_j)$, $P(x_i \leq x \leq x_j)$ and $P(x \geq x_i)$ considering the discrete case only.

[123] We are assuming $x_k \leq x_i$ when $k \leq i$ (i.e. x is ordered according to its index where the index refers to the sample points).

Table 6: The table of Problem 1 of § 4.3.1.

Sum (x)	2	3	4	5	6	7	8	9	10	11	12
Cumulative (F)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

(**a**) $P(x < 6)$. (**b**) $P(x > 4)$. (**c**) $P(x = 7)$.

(**d**) $P(3 < x \leq 10)$. (**e**) $P(5 \leq x \leq 8)$. (**f**) $P(2 < x < 9)$.

2. Find the following cumulative probabilities of the given mass functions (see Problem 5 of § 4.1):
(**a**) $P(k \leq 24)$ where $P(k) = 2^{-k}$ $(k = 1, 2, \ldots, \infty)$.
(**b**) $P(3 < k \leq 31)$ where $P(k) = (6/\pi^2)\, k^{-2}$ $(k = 1, 2, \ldots, \infty)$.
Answer:
(**a**) We use the geometric series formula, that is:

$$P(k \leq 24) = \sum_{k=1}^{24} 2^{-k} = \frac{1}{2}\left[\frac{1-(1/2)^{24}}{1-(1/2)}\right] = 1 - \left(\frac{1}{2}\right)^{24} = \frac{16777215}{16777216} \simeq 0.9999999404$$

(**b**) We have:

$$P(3 < k \leq 31) = \sum_{k=4}^{31} \frac{6}{\pi^2 k^2} = \frac{6}{\pi^2} \sum_{k=4}^{31} \frac{1}{k^2} \simeq 0.1532460141$$

PE: Find the following cumulative probabilities of the given mass functions:
(**a**) $P(5 < k \leq 17)$ where $P(k) = 2^{-k}$ $(k = 1, 2, \ldots, \infty)$.
(**b**) $P(7 \leq k \leq 25)$ where $P(k) = (6/\pi^2)\, k^{-2}$ $(k = 1, 2, \ldots, \infty)$.

3. Write a simple program that calculates the cumulative probabilities of the binomial distribution.
Answer: See the BinomialCumulative.cpp code.
PE: Explain how the BinomialCumulative.cpp code calculates these probabilities by using the rules of logarithm (despite the fact that the logarithm of a sum is not the sum of its logarithms).

4. Use the BinomialCumulative.cpp code (of Problem 3) to calculate the following cumulative binomial probabilities given in the form $P(n, p, k_1, k_2)$ where both k_1 and k_2 are included:
(**a**) $P(17, 0.14, 0, 11)$. (**b**) $P(59, 0.65, 15, 42)$. (**c**) $P(299, 0.38, 157, 299)$.
(**d**) $P(681, 0.51, 332, 375)$. (**e**) $P(1328, 0.85, 0, 1012)$. (**f**) $P(1569, 0.34, 0, 873)$.
Answer: Using the BinomialCumulative.cpp code we get:
(**a**) $P(17, 0.14, 0, 11) \simeq 0.999999824134$. (**b**) $P(59, 0.65, 15, 42) \simeq 0.872337883771$.
(**c**) $P(299, 0.38, 157, 299) \simeq 2.47781558955 \times 10^{-7}$. (**d**) $P(681, 0.51, 332, 375) \simeq 0.871942528426$.
(**e**) $P(1328, 0.85, 0, 1012) \simeq 2.79892295629 \times 10^{-17}$. (**f**) $P(1569, 0.34, 0, 873) \simeq 1.00000000000$.
PE: Repeat the Problem for the following:
(**a**) $P(46, 0.44, 0, 21)$. (**b**) $P(428, 0.91, 33, 283)$. (**c**) $P(810, 0.26, 268, 810)$.
(**d**) $P(926, 0.57, 28, 718)$. (**e**) $P(2689, 0.05, 194, 1962)$. (**f**) $P(1883, 0.31, 581, 1673)$.

5. Write a simple program that calculates the cumulative probabilities of the Poisson distribution.
Answer: See the PoissonCumulative.cpp code.
PE: Plot a flowchart representing the algorithm of the PoissonCumulative.cpp code.

6. Use the PoissonCumulative.cpp code (of Problem 5) to calculate the following cumulative Poisson probabilities given in the form $P(\lambda, k_1, k_2)$ where both k_1 and k_2 are included or the form $P(\lambda, k > k_2)$:
(**a**) $P(0.95, 0, 5)$. (**b**) $P(215, 14, 59)$. (**c**) $P(33.61, 23, 45)$.
(**d**) $P(6.7, 6, 81)$. (**e**) $P(7.82, k > 17)$. (**f**) $P(2.2, 0, 59)$.

Answer: Using the PoissonCumulative.cpp code we get:
(**a**) $P(0.95, 0, 5) \simeq 0.999544461196$. (**b**) $P(215, 14, 59) \simeq 1.72509161758 \times 10^{-36}$.
(**c**) $P(33.61, 23, 45) \simeq 0.953449107032$. (**d**) $P(6.7, 6, 81) \simeq 0.659350551358$.
(**e**) $P(7.82, k > 17) \simeq 1.24868076898 \times 10^{-3}$. (**f**) $P(2.2, 0, 59) \simeq 1.00000000000$.
PE: Repeat the Problem for the following:
(**a**) $P(53.8, k > 61)$. (**b**) $P(4.8, 0, 6)$. (**c**) $P(231.7, 21, 188)$.
(**d**) $P(19.26, 36, 187)$. (**e**) $P(39.5, 78, 92)$. (**f**) $P(0.18, 4, 9)$.

7. Give formulae for the cumulative distribution of the following discrete probability functions:
(**a**) Uniform. (**b**) Binomial. (**c**) Poisson. (**d**) Geometric. (**e**) Negative binomial.
Answer: We have (noting that $x \in \mathbb{R}$ and n is a given positive integer):
(**a**) From Eqs. 99 and 69 we get :

$$F(x) = \begin{cases} 0 & (x < x_1) \\ k/n & (x_k \le x < x_{k+1},\ k = 1, 2, \ldots, n-1) \\ 1 & (x \ge x_n) \end{cases}$$

(**b**) From Eqs. 99 and 71 we get:

$$F(x) = \begin{cases} 0 & (x < 0) \\ \sum_{i=0}^{k} C_i^n p^i (1-p)^{n-i} & \big[0 \le x < n, k = \text{floor}(x)\big] \\ 1 & (x \ge n) \end{cases}$$

where floor(x) is the greatest integer less than or equal to x.
(**c**) From Eqs. 99 and 75 we get:

$$F(x) = \begin{cases} 0 & (x < 0) \\ \sum_{i=0}^{k} \frac{\lambda^i e^{-\lambda}}{i!} & \big[0 \le x < \infty, k = \text{floor}(x)\big] \end{cases}$$

(**d**) From Eqs. 99 and 80 we get:

$$F(x) = \begin{cases} 0 & (x < 1) \\ 1 - (1-p)^k & \big[1 \le x < \infty, k = \text{floor}(x)\big] \end{cases}$$

We note that $F(x) = 1 - (1-p)^k$ (for $1 \le x < \infty$) because:

$$\begin{aligned} F(x) &= \sum_{i=1}^{k} (1-p)^{i-1} p && \big[k = \text{floor}(x)\big] \\ &= p \frac{1 - (1-p)^k}{1 - (1-p)} && \text{(geometric series)} \\ &= 1 - (1-p)^k \end{aligned}$$

Also see Problem 6 of § 4.1.2.
(**e**) From Eqs. 99 and 82 we get:

$$F(x) = \begin{cases} 0 & (x < 0) \\ \sum_{i=0}^{k} C_i^{i+r-1} p^r (1-p)^i & \big[0 \le x < \infty, k = \text{floor}(x)\big] \end{cases}$$

PE: Explain and justify (in words) each one of the stated cumulative distributions.

4.3.2 Continuous Random Variables

For continuous random variable, the cumulative distribution function is given by:

$$F(x_i) = \int_{-\infty}^{x_i} f(x)\,dx \tag{100}$$

where $f(x)$ is the probability density function. From this equation we conclude:

$$\frac{dF(x)}{dx} = f(x) \tag{101}$$

These equations (i.e. Eqs. 100 and 101) reflect the relationship between the probability density function f and the cumulative distribution function F. We can also conclude easily (using Eq. 100 as well as the properties of integration) that:

$$\begin{aligned} F(x_i \le x \le x_j) &= F(x_j) - F(x_i) = \int_{-\infty}^{x_j} f(x)\,dx - \int_{-\infty}^{x_i} f(x)\,dx \\ &= \int_{-\infty}^{x_j} f(x)\,dx + \int_{x_i}^{-\infty} f(x)\,dx = \int_{x_i}^{x_j} f(x)\,dx \end{aligned} \tag{102}$$

It is worth noting that whether we use strict inequality (i.e. $<$) or non-strict inequality (i.e. $\le$) in determining the limits of the cumulative probability distribution is important in the case of discrete distributions but not in the case of continuous distributions. This is because including and excluding a single point (i.e. a limit point) is important for the discrete distributions but not for the continuous distributions. The reason is that the values of probability in the discrete distributions are attached to individual points and hence a single point has a finite probability value and thus its inclusion and exclusion (i.e. in a sum) affect the cumulative probability. On the other hand, the values of probability in the continuous distributions are attached to areas under a curve (rather than individual points) and hence a single point has no finite probability value and thus its inclusion and exclusion (i.e. in an integral) have no effect on the cumulative probability. For example, $F(x_i < x \le x_j)$ and $F(x_i \le x \le x_j)$ are not equal (in general) if F is a discrete cumulative distribution but they are equal if F is a continuous cumulative distribution. This is because in the former case the probability of x_i is finite and hence the inclusion and exclusion of x_i in the sum of probabilities has an effect on the cumulative probability, while in the latter case the inclusion and exclusion of x_i has no effect on the cumulative probability because regardless of whether we include or exclude x_i the cumulative probability is determined by the same integral, i.e. $\int_{x_i}^{x_j} f(x)\,dx$.

Problems

1. Referring to the above statement: "On the other hand, the values of probability ... have no effect on the cumulative probability", it may be claimed that there is a contradiction between having a finite value of probability density for a given point x_i [since $f(x_i)$ is not zero in general] and the fact that the inclusion and exclusion of that point (i.e. x_i) in the integral have no effect on the cumulative probability. Discuss this issue in detail.
   Answer: We note the following:
   • As indicated above, the nature of discrete and continuous probability (as expressed and represented by the distribution function) is different. This can be seen clearly in the normalization conditions of these probabilities, i.e.

$$\sum_i p_i = 1 \quad \text{(discrete)} \qquad\qquad \int_{-\infty}^{+\infty} f(x)\,dx = 1 \quad \text{(continuous)}$$

   As we see, the continuous probability is not defined as $f(x)$ but as $f(x)\,dx$ (since it is an area under a curve) and that is why the normalization condition is not $\sum_i f(x_i) = 1$ as in the case of discrete

probability. Accordingly, if $dx \to 0$ (i.e. by choosing a single point) then $f(x_i)\,dx \to 0$ even though $f(x_i)$ is finite. So in brief, $f(x_i)$ is not the probability of x_i but it is the probability density (or rate) at x_i from which we obtain the probability of x_i (within an infinitesimal interval dx in the neighborhood of x_i) by multiplying $f(x_i)$ by dx.

• In our view, this can be seen as demonstration of the uncertainty principle (which is commonly associated with quantum physics although it has many applications and instances in many other fields such as mathematics and science in general).[124] This means that obtaining an exact value x_i (e.g. by observation or measurement) with no uncertainty (represented by dx or Δx) is impossible at least practically. However, our view may be challenged by the question: why do we not have such an uncertainty in the discrete case (if uncertainty is supposedly inherent to our observation and measurement)? But we can reply by noting that certain discrete variables do not have this type of uncertainty (e.g. integers representing the number of people or objects). Moreover, having an uncertainty in the discrete case (i.e. in some instances of discrete variables) is less obvious and has almost no practical consequences because the individual values of x_i are usually separated by large gaps (compared to the value of uncertainty dx or Δx) and hence the uncertainty on both sides of x_i (by a tiny margin dx or Δx) can be ignored because this has usually no theoretical or practical effect although we should always keep in mind that there is an uncertainty even in (some instances of) the discrete case (e.g. discrete voltages in digital electronic devices).

PE: Try to expand the above answer by adding more discussion and arguments.

2. Find the following cumulative probabilities of the given density functions (see Problem 5 of § 4.2):
   (**a**) $F(0.52 \le x \le 3.5)$ where $f(x) = 2e^{-2x}$ $(0 \le x < \infty)$.
   (**b**) $F(-0.45 \le x \le 0.69)$ where $f(x) = (3/2)\,x^2$ $(-1 \le x \le +1)$.
   Answer:
   (**a**)
   $$F(0.52 \le x \le 3.5) = \int_{0.52}^{3.5} 2e^{-2x}\,dx = \Big[-e^{-2x}\Big]_{0.52}^{3.5} = -e^{-7} + e^{-1.04} \simeq 0.35254279999$$
   (**b**)
   $$F(-0.45 < x \le 0.69) = \int_{-0.45}^{0.69} \frac{3x^2}{2}\,dx = \Big[\frac{x^3}{2}\Big]_{-0.45}^{+0.69} = \frac{0.328509 + 0.091125}{2} = 0.209817$$
   PE: Find the following cumulative probabilities of the given density functions:
   (**a**) $F(1.37 \le x \le 2.46)$ where $f(x) = 0.125x + 0.25$ $(1 \le x \le 3)$.
   (**b**) $F(3.9 \le x \le 6.8)$ where $f(x) = \frac{\ln(x)}{7\ln(7)-6}$ $(1 \le x \le 7)$.
3. Write a simple program that calculates the cumulative probabilities of the normal distribution.
   Answer: See the NormalCumulative.cpp code.
   PE: Write down the mathematical formulae used in the calculations performed in the NormalCumulative.cpp code.
4. Use the NormalCumulative.cpp code (of Problem 3) to calculate the following cumulative normal probabilities given in the form $F(\mu, V, x_1, x_2)$ where $x_1 < x_2$ (noting that x_1 can be $-\infty$ and x_2 can be $+\infty$):
   (**a**) $F(12.1, 9, -\infty, 17.3)$. (**b**) $F(-11.5, 5.76, -3.7, +\infty)$. (**c**) $F(92.7, 6.6, 33.5, 88.4)$.
   (**d**) $F(0, 39.8, -\infty, 40.1)$. (**e**) $F(19.2, 63, 18.1, 18.3)$. (**f**) $F(217, 71.3, 274, +\infty)$.
   Answer: Using the NormalCumulative.cpp code we get:
   (**a**) $F(12.1, 9, -\infty, 17.3) \simeq 0.95848178031$. (**b**) $F(-11.5, 5.76, -3.7, +\infty) \simeq 0.00057702504$.
   (**c**) $F(92.7, 6.6, 33.5, 88.4) \simeq 0.04708763686$. (**d**) $F(0, 39.8, -\infty, 40.1) \simeq 0.99999999990$.
   (**e**) $F(19.2, 63, 18.1, 18.3) \simeq 0.00997267574$. (**f**) $F(217, 71.3, 274, +\infty) \simeq 7.37143679430 \times 10^{-12}$.

[124] The reader is referred to our book "The Epistemology of Quantum Physics".

PE: Repeat the Problem for the following:
(**a**) $F(1.78, 2.4, -\infty, 0.51)$. (**b**) $F(-14.9, 0.9, -15, -13.2)$. (**c**) $F(37.2, 56.4, 41.4, +\infty)$.
(**d**) $F(67.8, 2.5, 50, 55.3)$. (**e**) $F(-9.3, 3.8, -11.6, +\infty)$. (**f**) $F(123.8, 25, -\infty, 99.2)$.

5. The lengths L of nails produced in a factory are normally distributed with mean $\mu = 3$cm and variance $V = 0.2\text{cm}^2$. Find the percentage of nails whose length is less than 2.7cm or greater than 3.3cm.
Answer: The cumulative probability $F(\mu, V, L_1 \leq L \leq L_2)$ is given by (noting that $\mu = 3$, $V = 0.2$, $L_1 = 2.7$ and $L_2 = 3.3$):

$$F(3, 0.2, 2.7 \leq L \leq 3.3) = \frac{1}{\sqrt{0.4\pi}} \int_{2.7}^{3.3} e^{-\frac{(x-3)^2}{0.4}} dx \simeq 0.497665045639$$

Accordingly, the probability of the length being less than 2.7cm or greater than 3.3cm is:

$$1 - F(3, 0.2, 2.7 \leq L \leq 3.3) \simeq 0.502334954361$$

which means that about 50.2% of the nails produced have lengths less than 2.7cm or greater than 3.3cm.
PE: Find the percentage of nails whose length is greater than 2.8cm.

6. Give an example showing that the normal distribution is generally more convenient in calculation than the corresponding binomial distribution and hence justify the use of the normal as an approximation to the binomial (when the approximation of the binomial by the normal is valid).
Answer: For example, let have a binomial probability problem where we have to calculate the (inner) cumulative probability $P(n, p, k_1 \leq k \leq k_2) = P(692, 0.47, 296 \leq k \leq 317)$. Now, the binomial cumulative probability is:

$$P(692, 0.47, 296 \leq k \leq 317) = \sum_{k=296}^{317} C_k^{692}\, 0.47^k\, 0.53^{692-k} \simeq 0.26631785039$$

On the other hand, the corresponding normal cumulative probability $F(\mu, V, 296 \leq x \leq 317)$ is [noting that $\mu = np = 325.24$ and $V = np(1-p) = 172.3772$]:

$$F(325.24, 172.3772, 296 \leq x \leq 317) = \frac{1}{\sqrt{344.7544\pi}} \int_{296}^{317} e^{-\frac{(x-325.24)^2}{344.7544}} dx \simeq 0.25216025443$$

As we see, the normal cumulative probability requires much less calculations than the corresponding binomial cumulative probability (noting that the binomial is the sum of 22 terms involving many large factorials while the normal is a single simple integral) and the results are reasonably close (i.e. the difference is tolerable for most practical purposes).
PE: Repeat the Problem by giving another example.

7. Give formulae for the cumulative distribution of the following continuous probability functions:
(**a**) Uniform. (**b**) Normal. (**c**) Exponential. (**d**) Gamma. (**e**) Cauchy.
Answer: We have (noting that $x \in \mathbb{R}$):
(**a**) From Eqs. 100 and 89 we get:

$$F(x) = \begin{cases} 0 & (x < a) \\ \frac{x-a}{b-a} & (a \leq x < b) \\ 1 & (x \geq b) \end{cases}$$

(**b**) From Eqs. 100 and 91 we get (noting that erf is the error function):

$$F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2V}}\right)\right] \qquad (-\infty < x < \infty)$$

(**c**) From Eqs. 100 and 93 we get:

$$F(x) = \begin{cases} 0 & (x \leq 0) \\ 1 - e^{-\alpha x} & (0 < x < \infty) \end{cases}$$

(**d**) From Eqs. 100 and 95 we get:

$$F(x) = \begin{cases} 0 & (x \leq 0) \\ \frac{\Gamma(r,\alpha x)}{(r-1)!} & (0 < x < \infty) \end{cases}$$

where Γ here is the lower incomplete gamma function.
(**e**) From Eqs. 100 and 97 we get:

$$F(x) = \frac{\arctan(x)}{\pi} + \frac{1}{2} \qquad (-\infty < x < \infty)$$

PE: Justify each one of the stated cumulative distributions by showing how to obtain the cumulative distribution from Eq. 100 in association with the probability density function of the particular distribution.

4.4 Multivariate Probability Functions

The previous investigations were largely about probability functions of a single random variable. However, two or more associated (or simultaneous) random variables can have a **joint probability function** that represents their interdependent distribution. These random variables could be discrete or continuous or mixed (i.e. some discrete and some continuous) where each one of these variables can be subject to a certain type of discrete or continuous function (such as those investigated earlier) depending on the nature of that variable. Moreover, these variables could be dependent or independent of each other (or they have mixed dependency). We can also have cumulative distribution functions for these multivariate joint probability functions (as well as ordinary distribution functions, i.e. mass and density functions).

Most of the previous probability principles and rules (and even some proofs and arguments) generally apply (with proper adaptations and modifications) to these joint probability functions. For example, a joint probability (mass) function of two discrete random variables $P(x, y)$ $\big[$where $x = x_1, x_2, \ldots, x_i, \ldots, x_m$ and $y = y_1, y_2, \ldots, y_j, \ldots, y_n\big]$ satisfies the following properties:[125]

$$P(x,y) = \begin{cases} P(x_i, y_j) & (x = x_i, y = y_j) \\ 0 & (\text{otherwise}) \end{cases} \quad (\text{definition}) \tag{103}$$

$$P(x,y) \nless 0 \quad (\text{probability } \geq 0) \tag{104}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} P(x_i, y_j) = 1 \quad (\text{normalization}) \tag{105}$$

$$F(x_i, y_j) = \sum_{k=1}^{i}\sum_{l=1}^{j} P(x_k, y_l) \quad (\text{cumulativity}) \tag{106}$$

$$F(a < x \leq b, c < y \leq d) = F(a,c) + F(b,d) - F(a,d) - F(b,c) \quad (\text{inner cumulativity}) \tag{107}$$

$$P(x,y) = P_x(x)\, P_y(y) \quad (\text{if } x, y \text{ are independent}) \tag{108}$$

where P_x and P_y are the probability mass functions of x and y respectively.[126]

[125] Functions of two random variables are labeled as **bivariate probability functions**. We also note that m or/and n can be infinite.

[126] The equation $P(x, y) = P_x(x)\, P_y(y)$ represents the fact that two discrete random variables x, y are independent *iff* their joint probability mass function $P(x, y)$ can be expressed as a product of a probability mass function of x only $P_x(x)$ times a probability mass function of y only $P_y(y)$.

Similarly, a joint probability (density) function of two continuous random variables $f(x, y)$ satisfies the following properties:

$$f(x_i, y_j)\,dx\,dy = P(x_i < x \leq x_i + dx,\, y_j < y \leq y_j + dy) \quad \text{(definition)} \tag{109}$$

$$f(x, y) \not< 0 \quad \text{(probability } \geq 0) \tag{110}$$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x, y)\,dx\,dy = 1 \quad \text{(normalization)} \tag{111}$$

$$F(x_i, y_j) = \int_{-\infty}^{y_j}\int_{-\infty}^{x_i} f(x, y)\,dx\,dy \quad \text{(cumulativity)} \tag{112}$$

$$F(a < x \leq b, c < y \leq d) = \int_{c}^{d}\int_{a}^{b} f(x, y)\,dx\,dy \quad \text{(inner cumulativity)} \tag{113}$$

$$f(x, y) = f_x(x)\,f_y(y) \quad \text{(if } x, y \text{ are independent)} \tag{114}$$

where f_x and f_y are the probability density functions of x and y respectively.[127]

Problems

1. Give some examples of discrete bivariate probability mass functions.
   Answer:
   • Discrete bivariate uniform-uniform:

$$P(x_i, y_j) = \begin{cases} \frac{1}{mn} & (i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n) \\ 0 & \text{(otherwise)} \end{cases}$$

   • The distribution represented by the following table:[128]

		y				
		1	**2**	**3**	**4**	**5**
x	**1**	33/784	21/784	67/784	2/784	78/784
	2	7/784	13/784	29/784	6/784	55/784
	3	42/784	63/784	5/784	21/784	48/784
	4	39/784	19/784	90/784	61/784	85/784

   PE: Show that the given examples are normalized. Also give more examples of discrete bivariate probability mass functions.
2. Give some examples of continuous bivariate probability density functions.
   Answer:
   • Continuous bivariate uniform-uniform:

$$f(x, y) = \begin{cases} \frac{1}{(b-a)\,(d-c)} & (a \leq x \leq b, c \leq y \leq d) \\ 0 & \text{(otherwise)} \end{cases}$$

   • Continuous bivariate normal-normal:

$$f(x, y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2} \qquad (-\infty < x < \infty, -\infty < y < \infty)$$

   PE: Show that the given examples are normalized. Also give more examples of continuous bivariate probability density functions.

[127] The equation $f(x, y) = f_x(x)\,f_y(y)$ represents the fact that two continuous random variables x, y are independent *iff* their joint probability density function $f(x, y)$ can be expressed as a product of a probability density function of x only $f_x(x)$ times a probability density function of y only $f_y(y)$.

[128] We note that x and y in this table represent the random variables while the entries in this table represent the probabilities of the pairs $(x = i, y = j)$ where $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4, 5$.

Chapter 5
Statistical Indicators

In this chapter we present a brief discussion of some well known and widely used statistical indicators which are usually met in the literature and textbooks of probability theory and its applications, and hence awareness and basic understanding of these indicators represent a necessity to those interested in probability theory and its applications (noting that these indicators are also commonly used and met in other branches of mathematics and science and hence their necessity and usefulness are more general). These statistical indicators are used to summarize statistical data sets and present them in a compact form that demonstrates their properties and general features. Hence, these indicators are very useful in the comprehension and interpretation of data sets and the appreciation of their significance.

5.1 Mean

The mean (which may also be called the expectation or expected value although some people distinguish between the two) represents the average value (per trial) expected (for a random variable) in a large number of trials of a given random experiment. For example, if we throw a fair coin many times then we should expect to get H in about half of these throws and T in the other half, and hence if we assign a numerical value of 0 to H and a numerical value of 1 to T then we should expect to get on average a numerical value of $1/2$ per throw.

However, before we go through the details of this investigation we should draw the attention to the notation that we use to represent the mean. In fact, we use μ to represent the mean but in two slightly different forms. So, we use bare μ or with a subscript (e.g. μ_x which means the mean of the random variable x) to represent the mean as a specific number while we use μ with parentheses to represent the operation of taking the mean of the variable inside the parentheses and hence it is like a function. For instance, μ_x means the number 6.25 (as an example) while $\mu(x)$ or $\mu[x]$ means the operation of taking the mean of the variable x. In fact, this notation applies even to the variance (which will be investigated in § 5.2) and hence we use, for instance, V_x and $V(x)$ in the same capacity as μ_x and $\mu(x)$.[129]

For a **discrete** real-valued random variable x (which can take distinct discrete values x_i with corresponding probabilities p_i), the mean is given by the sum:

$$\mu(x) = \frac{1}{n} \sum_i n_i \, x_i = \sum_i \frac{n_i}{n} \, x_i = \sum_i p_i \, x_i \tag{115}$$

where n is the total number of trials, x_i is the i^{th} value of the random variable x, n_i is the number of trials in which x_i is obtained, $p_i \equiv P(x_i)$ is the probability of x_i (i.e. the value of probability mass function corresponding to x_i), and i runs over all possible distinct values of x (noting that $\sum_i n_i = n$ which means that i runs over n_i not over n).

For a **continuous** real-valued random variable x (which can take continuous values between $-\infty < x < \infty$), the mean is given by the integral:

$$\mu(x) = \int_{-\infty}^{+\infty} x \, f(x) \, dx \tag{116}$$

where $f(x)$ is the probability density function of x.

It is worth noting that the mean of a function of x follows the above style, that is:

$$\mu\big[g(x)\big] = \sum_i g(x_i) \, p(x_i) \qquad \text{(discrete variable)} \tag{117}$$

[129] This similarly applies to the standard deviation where we use σ and $\sigma(x)$.

$$\mu\big[g(x)\big] = \int_{-\infty}^{+\infty} g(x)\, f(x)\, dx \qquad \text{(continuous variable)} \tag{118}$$

where g is a function of the random variable x.

The mean satisfies the following properties (with x, y being random variables defined over the same sample space):

$$\mu(C) = C \qquad (C \text{ is constant } \in \mathbb{R}) \tag{119}$$
$$\mu(x+y) = \mu(x) + \mu(y) \tag{120}$$
$$\mu(ax) = a\mu(x) \qquad (a \text{ is constant } \in \mathbb{R}) \tag{121}$$
$$\mu(ax+C) = a\mu(x) + C \qquad (a, C \text{ are constants } \in \mathbb{R}) \tag{122}$$
$$\mu(xy) = \mu(x)\,\mu(y) \qquad (x \text{ and } y \text{ are independent}) \tag{123}$$
$$\mu\big[g(x)\big] = \mu\big[g_1(x)\big] + \mu\big[g_2(x)\big] \qquad \big[g(x) = g_1(x) + g_2(x)\big] \tag{124}$$

These equations can be generalized to more than one variable. For example, Eq. 120 can be extended to more than two variables by repetitive application (see Problem 3). It should be obvious that for a random variable to have a mean, the sum of Eq. 115 (in the case of discrete) and the integral of Eq. 116 (in the case of continuous) should converge.

For a bivariate probability function (see § 4.4) of random variables x and y the mean of x is given by:

$$\mu(x) = \sum_i \sum_j x_i\, P(x_i, y_j) \qquad \text{(discrete variables)} \tag{125}$$

$$\mu(x) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, f(x,y)\, dx\, dy \qquad \text{(continuous variables)} \tag{126}$$

Similar definitions apply to the mean of the variable y, i.e. $\mu(y)$. We may also consider the mean of a bivariate function $g(x, y)$ and hence we have:

$$\mu\big[g(x,y)\big] = \sum_i \sum_j g(x_i, y_j)\, P(x_i, y_j) \qquad \text{(discrete variables)} \tag{127}$$

$$\mu\big[g(x,y)\big] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y)\, f(x,y)\, dx\, dy \qquad \text{(continuous variables)} \tag{128}$$

The **covariance** of two random variables x and y is defined as:

$$\mathrm{Cov}(x,y) = \mu\big[(x-\mu_x)\,(y-\mu_y)\big] \tag{129}$$

where $\mu_x = \mu(x)$ and $\mu_y = \mu(y)$. The **correlation** (or **correlation coefficient**) of two random variables x and y is defined as:

$$\mathrm{Cor}(x,y) = \frac{\mathrm{Cov}(x,y)}{\sqrt{V_x V_y}} \tag{130}$$

where $V_x = V(x) \neq 0$ and $V_y = V(y) \neq 0$. As we see, if $\mathrm{Cov}(x,y) = 0$ then $\mathrm{Cor}(x,y) = 0$ and the random variables x, y are then described as uncorrelated (otherwise they are described as correlated). Noting that the covariance of independent random variables is zero (see Problem 7), independent variables are always uncorrelated.

There are a few important points to note about the mean:

• The above definition of the mean (as given in the first paragraph of this section) is basic and lacks rigor, but it should be sufficient for our purpose. Moreover, conceptually (and even in some applications and according to some conventions) the mean is not the same as the expectation value, although for simplicity we prefer to treat them as identical. The readers are referred to the literature about the details of these issues and the difference in definitions and conventions related to them (noting that our book is introductory and hence these issues are out of scope).

• The mean can be sensibly attributed to the random variable, and hence we talk about the mean of a given random variable x. It can also be sensibly attributed to the probability distribution of the random variable, and hence we can talk about the mean of the mass function $P(x)$ or the density function $f(x)$ of a given random variable x. This should make sense by inspecting Eqs. 115 and 116 which involve both the random variable and its probability distribution. However, the "mean" primarily belongs to the random variable, and this should be evident from the meaning of "mean" as well as from its notation, e.g. $\mu(x)$ or μ_x.
• The mean can be seen as a functional reflecting the characteristics of the distribution function of its random variable. In fact, this should add more justification to our observation in the previous point that the mean can be sensibly attributed to the probability distribution of the random variable (as well as to the random variable itself).
• The purpose of our convention about the notation of mean and variance (i.e. μ or V with and without parentheses) is to avoid awful notations commonly used in probability and statistics which cause confusion and unnecessary complications (as well as extra effort in writing and typing). However, it should be noted that the difference between our two notations is not always clear-cut although this should not affect the rigor or comprehensibility in general.
• "Mean" in this book means **arithmetic mean** (noting that there are other types of mean like **geometric mean** and **harmonic mean**; see Problem 8).
• In the equations that involve more than one random variable (e.g. Eqs. 120 and 127) we generally assume that the random variables are defined over the same sample space.
• The mean of a given probability distribution may not exist and hence the distribution then is classified as pathological (see the Cauchy distribution in § 4.2.4). However, if the mean does exist then it (unlike some other statistical indicators) is unique (i.e. with respect to a given variable).
• The mean of a symmetric distribution with respect to $x = c$ (where x is a random variable and c is a constant) is $\mu(x) = c$ and hence no calculation is needed in this case, i.e. we should inspect the distribution (which we want to find its mean) for a potential symmetry before deciding if we have to go through the calculation of its mean (see Problem 12).
• The mean has the same physical dimensions as its random variable.
• The mean has the same sign as its random variable (if the random variable has a fixed sign). So, if the random variable takes only non-negative/non-positive values then its mean will also be non-negative/non-positive. This can be inferred from Eqs. 115 and 116 noting that p_i (in Eq. 115) and f (in Eq. 116) are non-negative.

Problems

1. Find the mean of the (discrete) data of:
   (**a**) The following list of values: {6, 6, 9.4, 10.7, 10.7, 10.7, 10.7, 15.6, 15.8, 19, 19, 21.7}.
   (**b**) Problem 5 of § 3.4 (see Table 4).
   Answer: From Eq. 115 we have:
   (**a**)

$$\mu(x) = \frac{1}{n}\sum_i n_i\, x_i = \frac{(2\times 6) + 9.4 + (4\times 10.7) + 15.6 + 15.8 + (2\times 19) + 21.7}{12} \simeq 12.94$$

   (**b**)

$$\mu(\text{sum}) = \sum_i p_i\, x_i = \frac{1}{36}\times 2 + \frac{2}{36}\times 3 + \cdots + \frac{1}{36}\times 12 = 7$$

   PE: Find the mean of the following data set: {2.3, 2.5, 5, 4.2, 5, 5, 7.1, 6.3, 7.1, 7.1, 8.4, 7.1, 8.9, 9.6, 11.5, 11.5, 7.6, 11.5, 17.4, 17.9}.
2. Find the mean of the (continuous) data represented by the following probability density functions:
   (**a**) $f(x) = \frac{x}{8}$ $(0 \le x \le 4)$, $f(x) = 0$ (otherwise). (**b**) $f(x) = \frac{\text{sech}(x-\pi)}{\pi}$ $(-\infty < x < +\infty)$.
   Answer: From Eq. 116 we have:

(**a**)

$$\mu(x) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_0^4 x\, \frac{x}{8}\, dx = \frac{8}{3}$$

(**b**)

$$\mu(x) = \int_{-\infty}^{+\infty} x\, f(x)\, dx = \int_{-\infty}^{+\infty} x\, \frac{\operatorname{sech}(x-\pi)}{\pi}\, dx = \pi$$

This result can also be obtained from the symmetry of $\operatorname{sech}(x-\pi)$ with respect to $x = \pi$.

PE: Repeat the Problem for the following probability density functions:

(**a**) $f(x) = \frac{3(2x-x^2)}{4}$ $(0 \le x \le 2)$, $\quad f(x) = 0$ (otherwise). (**b**) $f(x) = \frac{e^{-\frac{(x-2)^2}{18}}}{3\sqrt{2\pi}}$ $\quad (-\infty < x < +\infty)$.

3. Do some generalizations and adaptations to the properties of mean (see Eqs. 119-124).

Answer: For example:

• By repetitive application of Eq. 120 we get:

$$\mu\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} \mu\left(x_i\right)$$

where x_i $(i = 1, 2, \ldots, n)$ are random variables defined over the same sample space. Moreover, if we consider Eq. 121 as well we get (where a_i are constants):

$$\mu\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} \mu(a_i x_i) = \sum_{i=1}^{n} a_i \mu\left(x_i\right) \tag{131}$$

• We can extend Eq. 123 to get:

$$\mu\left(\prod_{i=1}^{n} x_i\right) = \prod_{i=1}^{n} \mu\left(x_i\right)$$

where x_i $(i = 1, 2, \ldots, n)$ are mutually independent random variables defined over the same sample space.

PE: Give more examples of generalizations and adaptations to the properties of mean.

4. Verify Eqs. 119-124 for discrete random variables.

Answer: In the following we mainly use Eqs. 115, 117 and 127.

• Regarding Eq. 119 we have:

$$\mu(C) = \frac{1}{n} \sum_i n_i\, C = \frac{C}{n} \sum_i n_i = \frac{C}{n} \times n = C$$

• Regarding Eq. 120 we have:

$$\begin{aligned}
\mu(x+y) &= \sum_i \sum_j (x_i + y_j)\, P(x_i, y_j) = \left[\sum_i \sum_j x_i\, P(x_i, y_j)\right] + \left[\sum_i \sum_j y_j\, P(x_i, y_j)\right] \\
&= \left[\sum_i x_i \sum_j P(x_i, y_j)\right] + \left[\sum_j y_j \sum_i P(x_i, y_j)\right] = \left[\sum_i x_i\, p(x_i)\right] + \left[\sum_j y_j\, p(y_j)\right] \\
&= \mu(x) + \mu(y)
\end{aligned}$$

• Regarding Eq. 121 we have:

$$\mu(ax) = \sum_i (a x_i)\, p(x_i) = a \sum_i x_i\, p_i = a\mu(x)$$

- Regarding Eq. 122 we have:

$$\mu(ax+C) = \sum_i (ax_i + C)\, p(x_i) = \left(a \sum_i x_i\, p_i \right) + \left(C \sum_i p_i \right) = a\mu(x) + C$$

- Regarding Eq. 123 we have:

$$\mu(xy) = \sum_i \sum_j (x_i y_j)\, P(x_i, y_j) = \sum_i \sum_j x_i\, y_j\, p(x_i)\, p(y_j) = \left(\sum_i p_i\, x_i \right) \left(\sum_j p_j\, y_j \right) = \mu(x)\, \mu(y)$$

- Regarding Eq. 124 we have:

$$\begin{aligned} \mu\big[g(x)\big] &= \sum_i p(x_i)\, g(x_i) = \sum_i p(x_i) \Big[g_1(x_i) + g_2(x_i) \Big] \\ &= \left[\sum_i p(x_i)\, g_1(x_i) \right] + \left[\sum_i p(x_i)\, g_2(x_i) \right] = \mu\big[g_1(x)\big] + \mu\big[g_2(x)\big] \end{aligned}$$

PE: Justify each step of the above verifications.

5. Verify Eqs. 119-124 for continuous random variables.
Answer: The verifications are straightforward by using the properties of integrals as well as the equations given in the preamble of this section (noting that density function is normalized to unity):

$$\begin{aligned} \mu(C) &= \int_{-\infty}^{+\infty} C\, f(x)\, dx = C \int_{-\infty}^{+\infty} f(x)\, dx = C \\ \mu(x+y) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x+y)\, f(x,y)\, dx\, dy \\ &= \left[\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, f(x,y)\, dx\, dy \right] + \left[\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y\, f(x,y)\, dx\, dy \right] \\ &= \left[\int_{-\infty}^{+\infty} x \left(\int_{-\infty}^{+\infty} f(x,y)\, dy \right) dx \right] + \left[\int_{-\infty}^{+\infty} y \left(\int_{-\infty}^{+\infty} f(x,y)\, dx \right) dy \right] \\ &= \left[\int_{-\infty}^{+\infty} x\, f_x(x)\, dx \right] + \left[\int_{-\infty}^{+\infty} y\, f_y(y)\, dy \right] \\ &= \mu(x) + \mu(y) \end{aligned}$$

$$\begin{aligned} \mu(ax) &= \int_{-\infty}^{+\infty} ax\, f(x)\, dx = a \int_{-\infty}^{+\infty} x\, f(x)\, dx = a\mu(x) \\ \mu(ax+C) &= \int_{-\infty}^{+\infty} (ax + C)\, f(x)\, dx = \left[a \int_{-\infty}^{+\infty} x\, f(x)\, dx \right] + \left[C \int_{-\infty}^{+\infty} f(x)\, dx \right] = a\mu(x) + C \\ \mu(xy) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, f(x,y)\, dx\, dy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, f_x(x)\, f_y(y)\, dx\, dy \\ &= \left[\int_{-\infty}^{+\infty} x\, f_x(x)\, dx \right] \left[\int_{-\infty}^{+\infty} y\, f_y(y)\, dy \right] = \mu(x)\, \mu(y) \\ \mu\big[g(x)\big] &= \int_{-\infty}^{+\infty} g(x)\, f(x)\, dx = \int_{-\infty}^{+\infty} \Big[g_1(x) + g_2(x) \Big]\, f(x)\, dx \end{aligned}$$

$$= \left[\int_{-\infty}^{+\infty} g_1(x)\, f(x)\, dx\right] + \left[\int_{-\infty}^{+\infty} g_2(x)\, f(x)\, dx\right] = \mu\big[g_1(x)\big] + \mu\big[g_2(x)\big]$$

PE: Justify each step of the above verifications.

6. Show that:

$$\mathrm{Cov}(x,y) = \mu(xy) - \mu(x)\,\mu(y) \tag{132}$$

Answer: We have:

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \mu\big[(x-\mu_x)\,(y-\mu_y)\big] && \text{(Eq. 129)}\\
&= \mu(xy - \mu_x y - \mu_y x + \mu_x\mu_y) &&\\
&= \mu(xy) - \mu(\mu_x y) - \mu(\mu_y x) + \mu(\mu_x\mu_y) && \text{(Eq. 131)}\\
&= \mu(xy) - \mu_x\mu(y) - \mu_y\mu(x) + \mu_x\mu_y && \text{(Eqs. 119 and 121)}\\
&= \mu(xy) - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y && \text{(definition)}\\
&= \mu(xy) - \mu_x\mu_y &&\\
&= \mu(xy) - \mu(x)\,\mu(y) && \text{(definition)}
\end{aligned}$$

We note that μ_x, μ_y and $\mu_x\mu_y$ are constants.

PE: Explain in words each step of the above derivation.

7. Show that the covariance of two independent random variables x and y is zero.

Answer: From the result of Problem 6 we have:

$$\mathrm{Cov}(x,y) = \mu(xy) - \mu(x)\,\mu(y)$$

Now, if x and y are independent then from Eq. 123 we have $\mu(xy) = \mu(x)\,\mu(y)$ and hence $\mathrm{Cov}(x,y) = 0$.

Note: if we do not assume Eq. 123 then we may prove this result as follows (considering the continuous case):[130]

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \mu\Big[(x-\mu_x)\,(y-\mu_y)\Big] && \text{(Eq. 129)}\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Big[(x-\mu_x)\,(y-\mu_y)\Big]\, f(x,y)\, dx\, dy && \text{(Eq. 128)}\\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Big[(x-\mu_x)\,(y-\mu_y)\Big]\, f_x(x)\, f_y(y)\, dx\, dy && \text{(Eq. 114)}\\
&= \left[\int_{-\infty}^{+\infty}(x-\mu_x)\, f_x(x)\, dx\right]\left[\int_{-\infty}^{+\infty}(y-\mu_y)\, f_y(y)\, dy\right] &&\\
&= \left[\int_{-\infty}^{+\infty} x\, f_x(x)\, dx - \mu_x\int_{-\infty}^{+\infty} f_x(x)\, dx\right]\left[\int_{-\infty}^{+\infty} y\, f_y(y)\, dy - \mu_y\int_{-\infty}^{+\infty} f_y(y)\, dy\right] &&\\
&= \left[\mu_x - \mu_x\int_{-\infty}^{+\infty} f_x(x)\, dx\right]\left[\mu_y - \mu_y\int_{-\infty}^{+\infty} f_y(y)\, dy\right] && \text{(Eq. 116)}\\
&= \Big[\mu_x - \mu_x\Big]\Big[\mu_y - \mu_y\Big] && \text{(Eq. 87)}\\
&= 0 \times 0 = 0 &&
\end{aligned}$$

PE: It is claimed that the converse of this statement is not true in general, i.e. if the covariance is zero then the variables are not necessarily independent. Investigate this issue and determine if this claim

[130] In fact, the purpose of this is to show the derivation in more details; otherwise both methods rest on the same principle (i.e. the joint probability function of two independent random variables, x and y, can be expressed as a product of a probability function of x alone times a probability function of y alone; see Eqs. 108 and 114).

is correct or not.[131]

8. Define arithmetic, geometric and harmonic means and state the relation between them (assuming all the data are positive).
Answer: The arithmetic mean μ is defined above (noting that it is the subject of this section). The geometric mean μ_G is given by:

$$\mu_G = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

The harmonic mean μ_H is given by:[132]

$$\mu_H = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

The relation between them is given by the inequality:

$$\mu_H \leq \mu_G \leq \mu$$

PE: Verify the relation $\mu_H \leq \mu_G \leq \mu$ for the data set of part (a) of Problem 1.

9. Find the mean of the following discrete distributions:
(**a**) Uniform. (**b**) Binomial. (**c**) Multinomial. (**d**) Poisson. (**e**) Geometric.
Answer:
(**a**) **Uniform**: we assume the distribution to be of the form $P(x_i) = P(i) = p_i = 1/n$ (where $i = 1, 2, \ldots, n$). Accordingly:

$$\begin{aligned}
\mu(x) &= \sum_i p_i\, x_i && \text{(Eq. 115)} \\
&= \sum_{i=1}^{n} \frac{1}{n}\, i && \left(p_i = \frac{1}{n} \text{ and } x_i = i\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} i \\
&= \frac{1}{n} \times (1 + 2 + \cdots + n) \\
&= \frac{1}{n} \times \frac{n(1+n)}{2} && \text{(arithmetic series formula)} \\
&= \frac{1+n}{2}
\end{aligned}$$

This result can also be obtained (without calculation) from the fact that the uniform discrete distribution (of the above form) is symmetric with respect to $x = (1+n)/2$ (see Problem 12).
(**b**) **Binomial**:

$$\mu(x) = \sum_i p_i\, x_i \qquad \text{(Eq. 115)}$$

[131] The reader should distinguish between the **converse** and **contrapositive** (i.e. of a conditional statement). For a conditional statement $a \to b$, the converse is $b \to a$ while the contrapositive is $\overline{b} \to \overline{a}$ (where the bar means negation). In general, the truth of contrapositive follows the truth of the statement (i.e. if $a \to b$ is true then $\overline{b} \to \overline{a}$ is also true) but this does not apply to the converse (i.e. if $a \to b$ is true then $b \to a$ is not necessarily true). By the way, the **inverse** of the conditional statement $a \to b$ is $\overline{a} \to \overline{b}$ which (like the converse) does not follow the truth of the statement (i.e. if $a \to b$ is true then $\overline{a} \to \overline{b}$ is not necessarily true).

[132] We note that the subscripts in μ_G and μ_H are specific to the geometric and harmonic means to distinguish them from other types of mean, and hence they are not subject to our convention about the notation of mean (i.e. arithmetic mean) which we outlined earlier in this section.

$$
\begin{aligned}
&= \sum_{i=0}^{n} i\, p_i && \text{(for binomial } x_i = 0, 1, \ldots, n\text{)} \\
&= \sum_{i=1}^{n} i\, p_i && \text{(0 is trivial)} \\
&= \sum_{i=1}^{n} i\, C_i^n p^i (1-p)^{n-i} && \text{(Eq. 71)} \\
&= \sum_{i=1}^{n} i\, \frac{n!}{i!(n-i)!}\, p^i (1-p)^{n-i} && \text{(Eq. 25)} \\
&= \sum_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!}\, p^i (1-p)^{n-i} && \text{(canceling } i\text{)} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-i)!}\, p^{i-1} (1-p)^{n-i} && \text{(taking } np \text{ out)} \\
&= np \sum_{i=1}^{n} \frac{(n-1)!}{(i-1)!(n-1-i+1)!}\, p^{i-1} (1-p)^{n-1-i+1} && \text{(adding and subtracting 1)} \\
&= np \sum_{i=1}^{n} C_{i-1}^{n-1}\, p^{i-1} (1-p)^{(n-1)-(i-1)} && \text{(Eq. 25)} \\
&= np \sum_{i=0}^{n-1} C_i^{n-1}\, p^i (1-p)^{(n-1)-i} && \text{(shifting the index)} \\
&= np(1-p+p)^{n-1} && \text{(Eq. 33)} \\
&= np
\end{aligned}
$$

(**c**) **Multinomial**: referring to Eq. 73, the k possible outcomes (or variables) in the multinomial distribution are mutually exclusive where each one of theses outcomes is subject to the binomial distribution (i.e. relative to the rest of these outcomes). In other words, if outcome i (where $i = 1, 2, \ldots, k$) has a probability p_i then the probability of the other outcomes (which correspond to the non-occurrence of outcome i) is $1 - p_i$ and hence p_i corresponds to p (i.e. the probability of occurrence) in the binomial distribution and $1 - p_i$ corresponds to $1 - p$ (i.e. the probability of non-occurrence) in the binomial distribution.[133] Accordingly, the mean of the x_i outcome in the multinomial distribution must be the same as the corresponding binomial distribution, that is:

$$\mu(x_i) = np_i \tag{133}$$

A simple verification of this (in a special case) is the formula of the mean of the binomial distribution considering that the binomial distribution is a multinomial distribution corresponding to $k = 2$, i.e.

$$P(n_1, n_2; p_1, p_2) = C_{n_1,n_2}^n p_1^{n_1} p_2^{n_2}$$

It is obvious that the mean of this "multinomial distribution" $\big[$which is $\mu(x_1) = np_1\big]$ is the same as the mean of the corresponding binomial distribution $\big[$which is $\mu = np\big]$ noting that $p_1 = p$, $p_2 = 1 - p_1$, $C_{n_1,n_2}^n = C_{n_1}^n$ and $n_1 + n_2 = n$. This should also apply to $\mu(x_2) = np_2$ which is the same as the mean of the non-occurrence in the corresponding binomial distribution $\big[$which is $\mu_{\text{non-occurrence}} = n(1-p)\big]$. So, we conclude that the mean of the variable x_i $(i = 1, 2, \ldots, k)$ in the multinomial distribution is

[133] If we use the terminology of multivariate distributions (by treating the multinomial as multivariate; see the notes of § 4.1.3 as well as § 4.4) then we can say: the *marginal distribution* of each random variable x_i (where $i = 1, 2, \ldots, k$) is a binomial distribution parameterized by n and p_i.

given by Eq. 133.[134]

(**d**) **Poisson**:

$$\begin{aligned}
\mu(x) &= \sum_i p_i\, x_i && \text{(Eq. 115)}\\
&= \sum_{i=0}^{\infty} i\, p_i && \text{(for Poisson } x_i = 0, 1, 2, \ldots\text{)}\\
&= \sum_{i=1}^{\infty} i\, p_i && \text{(0 is trivial)}\\
&= \sum_{i=1}^{\infty} i\, \frac{\lambda^i e^{-\lambda}}{i!} && \text{(Eq. 75)}\\
&= \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!} && \text{(canceling } i\text{)}\\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} && \text{(taking } \lambda e^{-\lambda} \text{ out)}\\
&= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} && \text{(shifting the index)}\\
&= \lambda e^{-\lambda}\, e^{\lambda} && \text{(the exponential series)}\\
&= \lambda
\end{aligned}$$

(**e**) **Geometric** (for brevity and clarity we use $q = 1 - p$):

$$\begin{aligned}
\mu(x) &= \sum_i p_i\, x_i && \text{(Eq. 115)}\\
\mu(x) &= \sum_{i=1}^{\infty} i\, p_i && \text{(for geometric } x_i = 1, 2, 3, \ldots\text{)}\\
\mu(x) &= \sum_{i=1}^{\infty} i\, q^{i-1} p && \text{(Eq. 80 with } q = 1 - p\text{)}\\
\mu(x) &= p + 2qp + 3q^2 p + \cdots && \text{(expanding)}\\
\mu(x) - q\mu(x) &= \left[p + 2qp + 3q^2 p + \cdots\right] - \left[qp + 2q^2 p + 3q^3 p + \cdots\right]\\
\mu(x) - q\mu(x) &= p + qp + q^2 p + q^3 p + \cdots\\
\mu(x) - q\mu(x) &= \frac{p}{1-q} && \text{(geometric series)}\\
\mu(x) &= \frac{p}{(1-q)^2}\\
\mu(x) &= \frac{p}{p^2} && (q = 1 - p)\\
\mu(x) &= \frac{1}{p}
\end{aligned}$$

PE: Investigate other methods for deriving the mean of the above discrete distributions.

10. Find the mean of the following continuous distributions:

(**a**) Uniform. (**b**) Normal. (**c**) Exponential.

[134] In fact, if we treat the multinomial distribution as a multivariate distribution then we can use Eq. 125 (as well as its extension and generalization) to prove Eq. 133.

Answer:
(**a**) **Uniform**:

$$\begin{aligned}
\mu(x) &= \int_{-\infty}^{+\infty} x\, f(x)\, dx && \text{(Eq. 116)} \\
&= \int_{a}^{b} x\, f(x)\, dx && (f = 0 \text{ for } x < a \text{ and } x > b) \\
&= \int_{a}^{b} x\, \frac{1}{b-a}\, dx && \text{(Eq. 89)} \\
&= \frac{1}{b-a} \int_{a}^{b} x\ dx \\
&= \frac{1}{b-a} \left[\frac{x^2}{2}\right]_{a}^{b} \\
&= \frac{1}{b-a} \left[\frac{b^2-a^2}{2}\right] \\
&= \frac{b+a}{2}
\end{aligned}$$

This result can also be obtained (without calculation) from the fact that the uniform continuous distribution is symmetric with respect to $x = (b+a)/2$ (see Problem 12).
(**b**) **Normal**: by definition, the mean of the normal distribution is μ (where μ here means the parameter in Eq. 91). So, all we need to do is to apply the definition of mean (i.e. Eq. 116) to see if this is consistent or not, that is:

$$\begin{aligned}
\mu(x) &= \int_{-\infty}^{+\infty} x\, f(x)\, dx && \text{(Eq. 116)} \\
&= \int_{-\infty}^{+\infty} x\, \frac{1}{\sqrt{2\pi V}}\, e^{-\frac{(x-\mu)^2}{2V}}\, dx && \text{(Eq. 91)} \\
&= \frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{+\infty} x\, e^{-\frac{(x-\mu)^2}{2V}}\, dx \\
&= \frac{1}{\sqrt{2\pi V}} \left(\mu\sqrt{2\pi V}\right) && \text{(by calculus)} \\
&= \mu
\end{aligned}$$

This result can also be obtained (without calculation) from the fact that the normal distribution is symmetric with respect to $x = \mu$ (see Problem 12).
(**c**) **Exponential**:

$$\begin{aligned}
\mu(x) &= \int_{-\infty}^{+\infty} x\, f(x)\, dx && \text{(Eq. 116)} \\
&= \int_{0}^{\infty} x\, f(x)\, dx && (0 < x < \infty) \\
&= \int_{0}^{\infty} x\, \alpha e^{-\alpha x}\, dx && \text{(Eq. 93)} \\
&= \alpha \int_{0}^{\infty} x\, e^{-\alpha x}\, dx \\
&= \alpha \left[-\frac{(1+\alpha x)e^{-\alpha x}}{\alpha^2}\right]_{0}^{\infty}
\end{aligned}$$

$$= \alpha \left[-0 + \frac{1}{\alpha^2} \right]$$
$$= \frac{1}{\alpha}$$

PE: Investigate other methods for deriving the mean of the above continuous distributions.

11. State the law of iterated expectations and prove it.
Answer: The **law of iterated expectations** states that if x and y are random variables defined over the same sample space then the mean of x is equal to the average of the means of x conditioned on y (i.e. for all possible values of y), that is:

$$\mu(x) = \mu\big[\mu(x|y)\big] \tag{134}$$

To prove it (using a discrete approach) we have:[135]

$$\begin{aligned}
\mu\big[\mu(x|y)\big] &= \mu \left[\sum_x x\, P(x|y) \right] && \text{(Eq. 115)} \\
&= \sum_y \left[\sum_x x\, P(x|y) \right] P(y) && \text{(Eq. 117)} \\
&= \sum_y \sum_x x\, P(x|y)\, P(y) \\
&= \sum_y \sum_x x\, P(y|x)\, P(x) && \text{(Eq. 47)} \\
&= \sum_x \sum_y x\, P(x)\, P(y|x) \\
&= \sum_x x\, P(x) \sum_y P(y|x) && \big[\text{no } y \text{ in } x\, P(x)\big] \\
&= \sum_x x\, P(x) && \left[\sum_y P(y|x) = 1 \text{ for all } x \right] \\
&= \mu(x) && \text{(Eq. 115)}
\end{aligned}$$

PE: Justify in words the above derivation.

12. Show that the mean of a probability distribution that is symmetric with respect to $x = c$ is $\mu(x) = c$ (assuming the distribution has a mean).
Answer: We show this for the case of continuous distribution where $f(x)$ is supposedly a symmetric distribution with respect to $x = c$. From Eq. 122 we have (with $a = 1$ and $C = -c$):

$$\mu(x - c) = \mu(x) - c \tag{135}$$

Now, if we use the linear transformation $y = x - c$ then the probability distribution function of y is given by (see Eq. 63 in Problem 2 of § 4):

$$g(y) = f\big[x(y)\big] \frac{dx}{dy} = f(y + c) \times 1 = f(y + c)$$

and thus (see Eq. 116):

$$\mu(x - c) = \mu(y) = \int_{-\infty}^{+\infty} y\, g(y)\, dy = \int_{-\infty}^{+\infty} y\, f(y + c)\, dy \tag{136}$$

[135] From a notational perspective (as well as from other perspectives), this proof is not sufficiently rigorous (noting that the purpose of it is to clarify this law and demonstrate the rationale behind it). Anyway, the law of iterated expectations is investigated and used marginally in this book and hence this proof should be enough for our purpose.

Now, since $f(x)$ is symmetric with respect to $x = c$ then $f(y)$ is symmetric with respect to $y = c$ and hence $f(y + c) = f(y - [-c])$ should be symmetric with respect to $y = 0$. Accordingly, the integral in Eq. 136 must be zero because the integrand is an odd function $\big[$since y is odd and $f(y + c)$ is even$\big]$. Therefore, from Eq. 136 we should have $\mu(x - c) = 0$. Hence, from Eq. 135 we get $\mu(x) - c = 0$ and thus $\mu(x) = c$ as required.
PE: Modify the above argument to fit the case of discrete distribution.

5.2 Variance and Standard Deviation

The variance of a numerical data set is a measure of the spread of the data around its average value. More technically and specifically, it is the mean of the squared deviation of a random variable from its mean.

For a **discrete** real-valued random variable x (which can take distinct discrete values x_i with corresponding probabilities p_i), the variance V is given by the sum:

$$V(x) = \frac{1}{n}\sum_i n_i\,(x_i - \mu_x)^2 = \sum_i \frac{n_i}{n}\,(x_i - \mu_x)^2 = \sum_i (x_i - \mu_x)^2\,p_i = \mu\Big([x - \mu_x]^2\Big) \tag{137}$$

where the symbols are as defined earlier and the last step is based on the definition of mean (see § 5.1 and Eq. 117 in particular).

For a **continuous** real-valued random variable x $\big[$which can take continuous values between $-\infty < x < \infty$ with a probability density function $f(x)\big]$, the variance V is given by the integral:

$$V(x) = \int_{-\infty}^{+\infty} (x - \mu_x)^2\,f(x)\,dx = \mu\Big([x - \mu_x]^2\Big) \tag{138}$$

where the symbols are as defined earlier and the last step is based on the definition of mean (see § 5.1 and Eq. 118 in particular).

As indicated earlier (and expressed explicitly in the last steps of Eqs. 137 and 138), the variance is no more than the mean of the squared deviations from the average (and this should provide a simple way for memorizing and recalling the formula of variance as the "mean square", i.e. the mean of the squared deviations). It should be obvious (from analyzing the above discussion and formulae) that large/small variance means large/small spread of the data around the mean, and this should provide more clarification about our claim earlier that the variance is a measure of the spread of the data around its average value.

The variance satisfies the following properties (with x, y being random variables defined over the same sample space):

$$V(C) = 0 \qquad (C \text{ is constant } \in \mathbb{R}) \tag{139}$$

$$V(x + y) = V(x) + V(y) \qquad (x \text{ and } y \text{ are independent}) \tag{140}$$

$$V(ax + b) = a^2\,V(x) \qquad (a \text{ and } b \text{ are constants } \in \mathbb{R}) \tag{141}$$

$$V(x + C) = V(x) \qquad (C \text{ is constant } \in \mathbb{R}) \tag{142}$$

$$V(x) = \mu\Big[V(x|y)\Big] + V\Big[\mu(x|y)\Big] \tag{143}$$

The last property is known as the **law of total variance** (among many other names). These properties (or some of them at least) can be easily generalized and extended to more variables. For example, Eq. 140 can be extended to more than two (mutually independent) variables by repetitive application of Eq. 140 (also see Problem 4).

For a bivariate probability function (see § 4.4) of random variables x and y the variance of x is given by:

$$V(x) = \sum_i \sum_j (x_i - \mu_x)^2\,P(x_i, y_j) \qquad \text{(discrete variables)} \tag{144}$$

$$V(x) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)^2 \, f(x, y) \, dx \, dy \qquad \text{(continuous variables)} \tag{145}$$

Similar definitions apply to the variance of the variable y, i.e. $V(y)$.

The standard deviation of a random variable is defined as the (positive) square root of its variance, i.e.

$$\sigma(x) = \sqrt{V(x)} \tag{146}$$

Like the variance, the standard deviation is a measure of the spread around the mean. Noting that the standard deviation is the square root of the variance (which is a "mean square"), we can exploit this to memorize and recall the formula of the standard deviation as the "root mean square", i.e. the root of the mean of the squared deviations.

There are a few important points to note about the variance (and standard deviation):

• It should be obvious that for a random variable to have a variance (and standard deviation), the series of Eq. 137 (in the case of discrete) and the integral of Eq. 138 (in the case of continuous) should converge.

• It should be obvious that for a random variable to have a variance (and standard deviation) it should have a mean. However, having a mean does not guarantee having a variance (and standard deviation). In other words, having a mean is a necessary, but not sufficient, condition for having a variance.

• Like the mean, the variance (and standard deviation) can be sensibly attributed to the probability distribution of the random variable as well as to the random variable itself.

• Like the mean, the variance (and standard deviation) can be seen as a functional reflecting the characteristics of the distribution function of its random variable.

• The variance (and thus the standard deviation) of a given probability distribution may not exist and hence the distribution then is classified as pathological (see the Cauchy distribution in § 4.2.4). However, if the variance does exist then it is unique (i.e. with respect to a given variable).

• In the equations that involve more than one random variable (e.g. Eq. 140) we generally assume that the random variables are defined over the same sample space.

• Unlike the mean (which can be positive or negative or zero), the variance and the standard deviation cannot be negative.

• The standard deviation has the same physical dimensions as its random variable, while the variance has the squared of the dimensions of its random variable.

Problems

1. Find the variance and standard deviation of the data sets of parts (a) and (b) of Problem 1 of § 5.1.
   Answer: We use Eqs. 137 and 146.
   (**a**) Noting that $\mu_x \simeq 12.94$ according to part (a) of Problem 1 of 5.1 we have:

$$\begin{aligned} V(x) &= \frac{1}{n} \sum_i n_i \, (x_i - \mu_x)^2 \simeq \frac{2 \times (6 - 12.94)^2 + (9.4 - 12.94)^2 + \cdots + (21.7 - 12.94)^2}{12} \simeq 24.53 \\ \sigma(x) &= \sqrt{V(x)} \simeq 4.95 \end{aligned}$$

   (**b**) Noting that $\mu_{\text{sum}} = 7$ according to part (b) of Problem 1 of 5.1 we have (noting that x represents the sum):

$$\begin{aligned} V(\text{sum}) &= \sum_i (x_i - \mu_x)^2 \, p_i = (2-7)^2 \times \frac{1}{36} + (3-7)^2 \times \frac{2}{36} + \cdots + (12-7)^2 \times \frac{1}{36} = \frac{210}{36} = \frac{35}{6} \\ \sigma(\text{sum}) &= \sqrt{V(\text{sum})} = \sqrt{35/6} \simeq 2.415 \end{aligned}$$

   PE: Find the variance and standard deviation of the data set of the PE of Problem 1 of § 5.1.
2. Find the variance and standard deviation of the density functions of parts (a) and (b) of Problem 2 of § 5.1.

Answer: We use Eqs. 138 and 146.
(**a**) Noting that $\mu_x = 8/3$ according to part (a) of Problem 2 of 5.1 we have:

$$\begin{aligned} V(x) &= \int_{-\infty}^{+\infty} (x-\mu_x)^2 f(x)\,dx = \int_0^4 \left(x - \frac{8}{3}\right)^2 \frac{x}{8}\,dx = \frac{8}{9} \\ \sigma(x) &= \sqrt{V(x)} = \sqrt{8/9} \simeq 0.943 \end{aligned}$$

(**b**) Noting that $\mu_x = \pi$ according to part (b) of Problem 2 of 5.1 we have:

$$\begin{aligned} V(x) &= \int_{-\infty}^{+\infty} (x-\mu_x)^2 f(x)\,dx = \int_{-\infty}^{+\infty} (x-\pi)^2 \frac{\operatorname{sech}(x-\pi)}{\pi}\,dx = \frac{\pi^2}{4} \\ \sigma(x) &= \sqrt{V(x)} = \sqrt{\pi^2/4} = \pi/2 \simeq 1.571 \end{aligned}$$

PE: Find the variance and standard deviation of the density functions of the PE of Problem 2 of § 5.1.

3. Show that the variance can also be given by:

$$V(x) = \mu(x^2) - [\mu(x)]^2 = \mu_{x^2} - \mu_x^2 \tag{147}$$

Answer: We have:

$$\begin{aligned} V(x) &= \mu\Big([x-\mu_x]^2\Big) && \text{(Eqs. 137 and 138)} \\ &= \mu(x^2 - 2x\,\mu_x + \mu_x^2) \\ &= \mu(x^2) - \mu\,(2x\,\mu_x) + \mu(\mu_x^2) && \text{(extension of Eq. 124)} \\ &= \mu(x^2) - 2\mu_x\mu(x) + \mu_x^2 && \text{(Eqs. 119 and 121)} \\ &= \mu(x^2) - 2\mu_x^2 + \mu_x^2 && \text{(definition of } \mu_x) \\ &= \mu(x^2) - \mu_x^2 \\ &= \mu(x^2) - [\mu(x)]^2 && \text{(definition of } \mu_x) \end{aligned}$$

Note: this relation is commonly written as $V(x) = \langle x^2 \rangle - \langle x \rangle^2$ where the triangular brackets $\langle\,\rangle$ symbolize mean. This form may be easier to remember (i.e. as square-in minus square-out).
PE: Investigate the possible advantages and disadvantages of using Eq. 147 instead of Eqs. 137 and 138. Can we conclude from Eq. 147 that for a random variable to have a variance it should have a mean $\mu(x)$ and a mean of its square $\mu(x^2)$, i.e. having a mean $\mu(x)$ is not sufficient for having a variance?

4. Give some examples of generalizations and adaptations that can be made to the properties of variance (see Eqs. 139-143).
Answer: For example, if $x_1, x_2, \ldots, x_n$ are mutually independent random variables defined over the same sample space then we can generalize Eq. 140 to:

$$V\left(\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} V\,(x_i)$$

Moreover, if we consider Eq. 141 as well then we get (where a_i are constants):

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

PE: Investigate other potential generalizations and adaptations to the properties of variance.

5. Verify Eqs. 139-143.
Answer:
• Regarding Eq. 139 we have (considering discrete and continuous cases together):

$$V(C) = \mu(C^2) - [\mu(C)]^2 = C^2 - [C]^2 = C^2 - C^2 = 0$$

where we used Eq. 147 in step 1 and Eq. 119 in step 2.
We may also consider the discrete case specifically:

$$V(C) = \frac{1}{n}\sum_i n_i\,(C - \mu[C])^2 = \frac{1}{n}\sum_i n_i\,(C - C)^2 = \frac{1}{n}\sum_i 0 = 0$$

where we used Eq. 137 in step 1 and Eq. 119 in step 2.
We may also consider the continuous case specifically:

$$V(C) = \int_{-\infty}^{+\infty} (C - \mu[C])^2\, f(x)\, dx = \int_{-\infty}^{+\infty} (C - C)^2\, f(x)\, dx = 0$$

where we used Eq. 138 in step 1 and Eq. 119 in step 2.
• Regarding Eq. 140 we have (considering discrete and continuous cases together):

$$\begin{aligned}
V(x+y) &= \mu\big[(x+y)^2\big] - \big[\mu(x+y)\big]^2 && \text{(Eq. 147)}\\
&= \mu\big[x^2 + 2xy + y^2\big] - \big[\mu(x) + \mu(y)\big]^2 && \text{(Eq. 120)}\\
&= \mu(x^2) + 2\mu(xy) + \mu(y^2) - \big[\mu(x)\big]^2 - 2\mu(x)\mu(y) - \big[\mu(y)\big]^2 && \text{(Eq. 131)}\\
&= \Big\{\mu(x^2) - \big[\mu(x)\big]^2\Big\} + \Big\{\mu(y^2) - \big[\mu(y)\big]^2\Big\} + \Big\{2\mu(xy) - 2\mu(x)\mu(y)\Big\}\\
&= V(x) + V(y) + 2\Big\{\mu(xy) - \mu(x)\mu(y)\Big\} && \text{(Eq. 147)}\\
&= V(x) + V(y) + 2\,\mathrm{Cov}(x,y) && \text{(Eq. 132)}
\end{aligned}$$

Now, since x and y are independent then we can use the result of Problem 7 of § 5.1 $\big[$i.e. $\mathrm{Cov}(x,y) = 0\big]$ and hence we conclude that $V(x+y) = V(x) + V(y)$. It is worth noting that if we discard the condition that x and y are independent then from the last line we have:

$$V(x+y) = V(x) + V(y) + 2\,\mathrm{Cov}(x,y)$$

So, Eq. 140 is a special case of this general equation.
• Regarding Eq. 141 we have (considering discrete and continuous cases together):

$$\begin{aligned}
V(ax+b) &= \mu\big[(ax+b)^2\big] - \big[\mu(ax+b)\big]^2 && \text{(Eq. 147)}\\
&= \mu\big[a^2x^2 + 2abx + b^2\big] - \big[a\mu(x) + b\big]^2 && \text{(Eq. 122)}\\
&= a^2\mu(x^2) + 2ab\,\mu(x) + \mu(b^2) - a^2\big[\mu(x)\big]^2 - 2ab\,\mu(x) - b^2 && \text{(Eqs. 121 and 124)}\\
&= a^2\mu(x^2) + 2ab\,\mu(x) + b^2 - a^2\big[\mu(x)\big]^2 - 2ab\,\mu(x) - b^2 && \text{(Eq. 119)}\\
&= a^2\mu(x^2) - a^2\big[\mu(x)\big]^2\\
&= a^2\Big\{\mu(x^2) - \big[\mu(x)\big]^2\Big\}\\
&= a^2\,V(x) && \text{(Eq. 147)}
\end{aligned}$$

We may also consider the discrete case specifically (see Eq. 137 as well as Eqs. 119-124):[136]

$$V(ax+b) \quad = \quad \frac{1}{n}\sum_i n_i\Big[(ax_i + b) - \mu(ax+b)\Big]^2 = \frac{1}{n}\sum_i n_i\Big[(ax_i + b) - a\mu(x) - \mu(b)\Big]^2$$

[136] Because the variance is a mean then the variance of a function of x follows the style of the variance of x.

$$\begin{aligned} &= \frac{1}{n}\sum_i n_i \Big[(ax_i+b)-a\mu_x-b\Big]^2 = \frac{1}{n}\sum_i n_i \Big[ax_i-a\mu_x\Big]^2 \\ &= a^2\,\frac{1}{n}\sum_i n_i\,(x_i-\mu_x)^2 = a^2\,V(x) \end{aligned}$$

We may also consider the continuous case specifically (see Eq. 138 as well as Eqs. 119-124):

$$\begin{aligned} V(ax+b) &= \int_{-\infty}^{+\infty}\Big[(ax+b)-\mu(ax+b)\Big]^2 f(x)\,dx = \int_{-\infty}^{+\infty}\Big[(ax+b)-(a\mu_x+b)\Big]^2 f(x)\,dx \\ &= \int_{-\infty}^{+\infty}\Big[a(x-\mu_x)\Big]^2 f(x)\,dx = a^2\int_{-\infty}^{+\infty}(x-\mu_x)^2 f(x)\,dx = a^2\,V(x) \end{aligned}$$

- Regarding Eq. 142, it is an instance of Eq. 141 with $a=1$ and $b=C$.
- Regarding Eq. 143 we have (considering discrete and continuous cases together):

$$\begin{aligned} V(x) &= \mu(x^2)-\Big[\mu(x)\Big]^2 && \text{(Eq. 147)} \\ &= \mu\Big[\mu(x^2|y)\Big]-\Big[\mu(x)\Big]^2 && \text{(Eq. 134)} \\ &= \mu\Big[V(x|y)+[\mu(x|y)]^2\Big]-\Big[\mu(x)\Big]^2 && \text{(Eq. 147)} \\ &= \mu\Big[V(x|y)\Big]+\mu\Big[[\mu(x|y)]^2\Big]-\Big[\mu(x)\Big]^2 && \text{(Eq. 124)} \\ &= \mu\Big[V(x|y)\Big]+\mu\Big[[\mu(x|y)]^2\Big]-\Big[\mu\big[\mu(x|y)\big]\Big]^2 && \text{(Eq. 134)} \\ &= \mu\Big[V(x|y)\Big]+V\Big[\mu(x|y)\Big] && \text{(Eq. 147)} \end{aligned}$$

PE: Justify the steps that we did not justify in the above verifications.

6. Find the variance of the following discrete distributions:
(**a**) Uniform. (**b**) Binomial. (**c**) Multinomial. (**d**) Poisson. (**e**) Geometric.

Answer:
(**a**) **Uniform**: we assume the distribution to be of the form $P(x_i)=P(i)=p_i=1/n$ (where $i=1,2,\ldots,n$). Accordingly:

$$\begin{aligned} V(x) &= \mu(x^2)-[\mu(x)]^2 && \text{(Eq. 147)} \\ &= \mu(x^2)-\left(\frac{1+n}{2}\right)^2 && \text{(Eq. 70)} \\ &= \left(\sum_i p_i\,x_i^2\right)-\left(\frac{1+n}{2}\right)^2 && \text{(Eq. 117)} \\ &= \left(\sum_{i=1}^{n}\frac{1}{n}\,i^2\right)-\left(\frac{1+n}{2}\right)^2 && \left(p_i=\frac{1}{n}\text{ and }x_i=i\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} i^2\right)-\left(\frac{1+n}{2}\right)^2 && \\ &= \frac{1}{n}\times\frac{n(n+1)(2n+1)}{6}-\left(\frac{1+n}{2}\right)^2 && \left(\text{formula of }\sum_{i=1}^{n} i^2\right) \\ &= \frac{(n+1)(2n+1)}{6}-\left(\frac{1+n}{2}\right)^2 && \end{aligned}$$

$$= (n+1)\left[\frac{2n+1}{6} - \frac{1+n}{4}\right]$$
$$= (n+1)\left[\frac{n-1}{12}\right]$$
$$= \frac{n^2-1}{12}$$

(**b**) **Binomial**:

$$\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)}\\
&= \mu(x^2) - (np)^2 && \text{(Eq. 72)}\\
&= \mu(x^2 - x + x) - (np)^2 && (-\text{ and } +\, x)\\
&= \mu\Big(x[x-1] + x\Big) - (np)^2 &&\\
&= \mu\Big(x[x-1]\Big) + \mu(x) - (np)^2 && \text{(Eq. 124)}\\
&= \mu\Big(x[x-1]\Big) + np - (np)^2 && \text{(Eq. 72)}\\
&= np - (np)^2 + \mu\Big(x[x-1]\Big) &&\\
&= np - (np)^2 + \sum_i x_i(x_i-1)\, p_i && \text{(Eq. 117)}\\
&= np - (np)^2 + \sum_{i=0}^{n} i(i-1)\, C_i^n p^i (1-p)^{n-i} && \text{(Eq. 71)}\\
&= np - (np)^2 + \sum_{i=2}^{n} i(i-1)\, C_i^n p^i (1-p)^{n-i} && \text{(0 is trivial)}\\
&= np - (np)^2 + \sum_{i=2}^{n} i(i-1)\, \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} && \text{(Eq. 25)}\\
&= np - (np)^2 + \sum_{i=2}^{n} \frac{n!}{(i-2)!(n-i)!} p^i (1-p)^{n-i} && \Big[\text{cancel } i(i-1)\Big]\\
&= np - (np)^2 + n(n-1)\sum_{i=2}^{n} \frac{(n-2)!}{(i-2)!(n-i)!} p^i (1-p)^{n-i} && \Big[\text{take } n(n-1) \text{ out}\Big]\\
&= np - (np)^2 + n(n-1)p^2 \sum_{i=2}^{n} \frac{(n-2)!}{(i-2)!(n-i)!} p^{i-2} (1-p)^{n-i} && (\text{take } p^2 \text{ out})\\
&= np - (np)^2 + n(n-1)p^2 \sum_{i=2}^{n} \frac{(n-2)!}{(i-2)!(n-2-i+2)!} p^{i-2} (1-p)^{n-2-i+2} && (-\text{ and } +\, 2)\\
&= np - (np)^2 + n(n-1)p^2 \sum_{i=2}^{n} C_{i-2}^{n-2} p^{i-2} (1-p)^{(n-2)-(i-2)} && \text{(Eq. 25)}\\
&= np - (np)^2 + n(n-1)p^2 \sum_{i=0}^{n-2} C_i^{n-2} p^i (1-p)^{n-2-i} && (\text{shif the index})\\
&= np - (np)^2 + n(n-1)p^2 (1-p+p)^{n-2} && \text{(Eq. 33)}\\
&= np - (np)^2 + n(n-1)p^2 &&\\
&= np - (np)^2 + n^2p^2 - np^2 &&\\
&= np(1-p) &&
\end{aligned}$$

(**c**) **Multinomial**: we repeat our argument in part (c) of Problem 9 of § 5.1 and hence we conclude that the variance of the variable x_i $(i = 1, 2, \ldots, k)$ in the multinomial distribution is given by (see Eq. 72):

$$V(x_i) = np_i(1 - p_i)$$

(**d**) **Poisson**:

$$\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)}\\
&= \mu(x^2) - \lambda^2 && \text{(Eq. 76)}\\
&= -\lambda^2 + \mu(x^2)\\
&= -\lambda^2 + \sum_i x_i^2\, p_i && \text{(Eq. 117)}\\
&= -\lambda^2 + \sum_{i=0}^{\infty} i^2\, \frac{\lambda^i e^{-\lambda}}{i!} && \text{(Eq. 75)}\\
&= -\lambda^2 + \sum_{i=1}^{\infty} i^2\, \frac{\lambda^i e^{-\lambda}}{i!} && \text{(0 is trivial)}\\
&= -\lambda^2 + \sum_{i=1}^{\infty} i\, \frac{\lambda^i e^{-\lambda}}{(i-1)!} && \text{(cancel } i\text{)}\\
&= -\lambda^2 + \sum_{i=1}^{\infty} (i - 1 + 1)\, \frac{\lambda^i e^{-\lambda}}{(i-1)!} && (-\text{ and } +\ 1)\\
&= -\lambda^2 + \left[\sum_{i=1}^{\infty} (i-1)\, \frac{\lambda^i e^{-\lambda}}{(i-1)!}\right] + \left[\sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!}\right] && \text{(distribution)}\\
&= -\lambda^2 + \left[\sum_{i=2}^{\infty} (i-1)\, \frac{\lambda^i e^{-\lambda}}{(i-1)!}\right] + \left[\sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!}\right] && \text{(0 is trivial)}\\
&= -\lambda^2 + \left[\sum_{i=2}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-2)!}\right] + \left[\sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!}\right] && \text{(cancel } i-1\text{)}\\
&= -\lambda^2 + \left[\lambda^2 e^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!}\right] + \left[\lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!}\right] && \text{(factorize)}\\
&= -\lambda^2 + \left[\lambda^2 e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\right] + \left[\lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}\right] && \text{(shift the indices)}\\
&= -\lambda^2 + \lambda^2 e^{-\lambda}\, e^{\lambda} + \lambda e^{-\lambda}\, e^{\lambda} && \text{(the exponential series)}\\
&= -\lambda^2 + \lambda^2 + \lambda\\
&= \lambda
\end{aligned}$$

(**e**) **Geometric** (for brevity and clarity we use $q = 1 - p$):

$$\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)}\\
&= \mu(x^2) - \mu(x) + \mu(x) - [\mu(x)]^2 && \big[-\text{ and } +\ \mu(\text{x})\big]\\
&= \mu(x^2) - \mu(x) + \frac{1}{p} - \frac{1}{p^2} && \text{(Eq. 81)}\\
&= \mu(x^2 - x) + \frac{1}{p} - \frac{1}{p^2} && \text{(Eq. 124)}
\end{aligned} \tag{148}$$

So, what we need now to find $V(x)$ is to find $\mu(x^2 - x)$, that is:

$$\begin{aligned}
\mu(x^2 - x) &= \mu\big[x(x-1)\big] \\
&= \sum_i x_i(x_i - 1)\, p_i && \text{(Eq. 117)} \\
&= \sum_{i=1}^{\infty} i(i-1)\, q^{i-1} p && \text{(Eq. 80 with } q = 1-p\text{)} \\
&= p \sum_{i=1}^{\infty} i(i-1)\, q^{i-1} && \text{(factorize)} \\
&= p \frac{d}{dq} \left[\sum_{i=1}^{\infty} (i-1)\, q^i \right] && \text{(calculus)} \\
&= p \frac{d}{dq} \left[\sum_{i=2}^{\infty} (i-1)\, q^i \right] && \text{(0 is trivial)} \\
&= p \frac{d}{dq} \left[q^2 \sum_{i=2}^{\infty} (i-1)\, q^{i-2} \right] && \text{(factorize)} \\
&= p \frac{d}{dq} \left[q^2 \frac{d}{dq} \left\{ \sum_{i=2}^{\infty} q^{i-1} \right\} \right] && \text{(calculus)} \\
&= p \frac{d}{dq} \left[q^2 \frac{d}{dq} \left\{ \sum_{i=1}^{\infty} q^{i} \right\} \right] && \text{(shift index)} \\
&= p \frac{d}{dq} \left[q^2 \frac{d}{dq} \left\{ \frac{q}{1-q} \right\} \right] && \text{(geometric series)} \\
&= p \frac{d}{dq} \left[q^2 \frac{1}{(1-q)^2} \right] && \text{(calculus)} \\
&= p \left[\frac{2q}{(1-q)^3} \right] && \text{(calculus)} \\
&= (1-q) \left[\frac{2q}{(1-q)^3} \right] && (p = 1-q) \\
&= \frac{2q}{(1-q)^2} \\
&= \frac{2(1-p)}{p^2} && (q = 1-p)
\end{aligned}$$

On substituting from the last equation into Eq. 148 we get:

$$V(x) = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2 - 2p + p - 1}{p^2} = \frac{1-p}{p^2}$$

PE: Investigate other methods for deriving the variance of the above discrete distributions.

7. Find the variance of the following continuous distributions:

(**a**) Uniform. (**b**) Normal. (**c**) Exponential.

Answer:

(**a**) **Uniform**:

$$\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)} \\
&= \mu(x^2) - \left(\frac{a+b}{2} \right)^2 && \text{(Eq. 90)}
\end{aligned}$$

$$
\begin{aligned}
&= -\left(\frac{a+b}{2}\right)^2 + \mu(x^2) \\
&= -\left(\frac{a+b}{2}\right)^2 + \int_{-\infty}^{+\infty} x^2\, f(x)\, dx && \text{(Eq. 118)} \\
&= -\left(\frac{a+b}{2}\right)^2 + \int_{a}^{b} x^2\, \frac{1}{b-a}\, dx && \text{(Eq. 89)} \\
&= -\left(\frac{a+b}{2}\right)^2 + \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b \\
&= -\left(\frac{a+b}{2}\right)^2 + \frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right] \\
&= -\frac{a^2+2ab+b^2}{4} + \frac{b^2+ab+a^2}{3} \\
&= \frac{-3a^2-6ab-3b^2+4b^2+4ab+4a^2}{12} \\
&= \frac{a^2-2ab+b^2}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}
$$

(**b**) **Normal**: by definition, the variance of the normal distribution is V (where V here means the parameter in Eq. 91). So, all we need to do is to apply the equation of variance (i.e. Eq. 147) to see if this is consistent or not, that is:

$$
\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)} \\
&= \mu(x^2) - \mu^2 && \text{(see part b of Problem 10 of § 5.1)} \\
&= -\mu^2 + \mu(x^2) \\
&= -\mu^2 + \int_{-\infty}^{+\infty} x^2\, f(x)\, dx && \text{(Eq. 118)} \\
&= -\mu^2 + \int_{-\infty}^{+\infty} x^2\, \frac{1}{\sqrt{2\pi V}} e^{-\frac{(x-\mu)^2}{2V}}\, dx && \text{(Eq. 91)} \\
&= -\mu^2 + \frac{1}{\sqrt{2\pi V}} \int_{-\infty}^{+\infty} x^2\, e^{-\frac{(x-\mu)^2}{2V}}\, dx \\
&= -\mu^2 + \frac{1}{\sqrt{2\pi V}} \sqrt{2\pi V}\left(\mu^2 + V\right) && \text{(from calculus)} \\
&= -\mu^2 + \mu^2 + V \\
&= V
\end{aligned}
$$

(**c**) **Exponential**:

$$
\begin{aligned}
V(x) &= \mu(x^2) - [\mu(x)]^2 && \text{(Eq. 147)} \\
&= \mu(x^2) - \frac{1}{\alpha^2} && \text{(Eq. 94)} \\
&= -\frac{1}{\alpha^2} + \mu(x^2) \\
&= -\frac{1}{\alpha^2} + \int_{-\infty}^{+\infty} x^2\, f(x)\, dx && \text{(Eq. 118)} \\
&= -\frac{1}{\alpha^2} + \int_{0}^{\infty} x^2\, \alpha e^{-\alpha x}\, dx && \text{(Eq. 93)}
\end{aligned}
$$

$$= -\frac{1}{\alpha^2} + \frac{2}{\alpha^2} \qquad \text{(from calculus)}$$
$$= \frac{1}{\alpha^2}$$

PE: Investigate other methods for deriving the variance of the above continuous distributions.

8. An experimentally-collected data set is given by the list of values $\{0.12, 1.23, 0.89, 5.60, 3.22\}$ with corresponding probabilities $\{0.15, 0.13, 0.21, 0.33, 0.18\}$. If the random variable x represents the values in the list while the random variable y represents the square root of these values, find the covariance and correlation of x and y.

Answer: We have:

$$\mu(x) = \sum_{i=1}^{5} p_i\, x_i = 2.7924 \qquad \text{(Eq. 115)}$$

$$\mu(y) = \sum_{i=1}^{5} p_i\, \sqrt{x_i} \simeq 1.498173 \qquad \text{(Eq. 117)}$$

$$\mu(xy) = \sum_{i=1}^{5} p_i\, x_i^{3/2} \simeq 5.773115 \qquad \text{(Eq. 117)}$$

$$V(x) = \sum_{i=1}^{5} (x_i - \mu_x)^2\, p_i \simeq 4.782792 \qquad \text{(Eq. 137)}$$

$$V(y) = \sum_{i=1}^{5} (y_i - \mu_y)^2\, p_i \simeq 0.547877 \qquad \text{(Eq. 137)}$$

$$\text{Cov}(x, y) = \mu(xy) - \mu(x)\mu(y) \simeq 1.589617 \qquad \text{(Eq. 132)}$$

$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V_x V_y}} \simeq 0.981997 \qquad \text{(Eq. 130)}$$

Note: the covariance can also be calculated from its definition (i.e. by using Eq. 129), that is:

$$\text{Cov}(x, y) = \mu\big[(x - \mu_x)\,(y - \mu_y)\big] = \sum_{i=1}^{5} (x_i - \mu_x)\,(y_i - \mu_y)\, p_i \simeq 1.589617$$

PE: Repeat the Problem where y now represents the cubic root of these values.

9. Verify the following (noting that x, y, x_1, x_2, y_1, y_2 are random variables and a_1, a_2, b_1, b_2 are constants):

(**a**) $\text{Cov}(x, x) = V(x)$. (**b**) $\text{Cov}(x, y) = \text{Cov}(y, x)$.

(**c**) If $y_1 = a_1 x_1 + b_1$ and $y_2 = a_2 x_2 + b_2$ then $\text{Cov}(y_1, y_2) = a_1 a_2\, \text{Cov}(x_1, x_2)$.

Answer:

(**a**) We have:

$$\text{Cov}(x, x) = \mu\Big[(x - \mu_x)\,(x - \mu_x)\Big] \qquad \text{(Eq. 129)}$$
$$= \mu\Big[(x - \mu_x)^2\Big]$$
$$= V(x) \qquad \text{(Eqs. 137 and 138)}$$

(**b**) We have:

$$\text{Cov}(x, y) = \mu\Big[(x - \mu_x)\,(y - \mu_y)\Big] \qquad \text{(Eq. 129)}$$
$$= \mu\Big[(y - \mu_y)\,(x - \mu_x)\Big]$$

$$= \mathrm{Cov}(y, x) \qquad \text{(Eq. 129)}$$

(**c**) We have:

$$\begin{aligned}
\mathrm{Cov}(y_1, y_2) &= \mu\Big[(y_1 - \mu_{y_1})\,(y_2 - \mu_{y_2})\Big] && \text{(Eq. 129)}\\
&= \mu\Big[\Big(\{a_1x_1 + b_1\} - \{a_1\mu(x_1) + b_1\}\Big)\ \Big(\{a_2x_2 + b_2\} - \{a_2\mu(x_2) + b_2\}\Big)\Big] && \text{(Eq. 122)}\\
&= \mu\Big[\Big(a_1x_1 - a_1\mu(x_1)\Big)\ \Big(a_2x_2 - a_2\mu(x_2)\Big)\Big]\\
&= \mu\Big[a_1a_2\Big(x_1 - \mu(x_1)\Big)\ \Big(x_2 - \mu(x_2)\Big)\Big]\\
&= a_1a_2\,\mu\Big[\Big(x_1 - \mu(x_1)\Big)\ \Big(x_2 - \mu(x_2)\Big)\Big] && \text{(Eq. 121)}\\
&= a_1a_2\,\mu\Big[\Big(x_1 - \mu_{x_1}\Big)\ \Big(x_2 - \mu_{x_2}\Big)\Big]\\
&= a_1a_2\,\mathrm{Cov}(x_1, x_2) && \text{(Eq. 129)}
\end{aligned}$$

Chapter 6
Useful Theorems

In this chapter we investigate some useful theorems (or laws or results or ...) in the probability theory which are commonly met in the literature of probability, and hence the reader should have some awareness and understanding of their content, significance and application. In fact, some of these theorems (such as the Bayes theorem) are fundamental and central to the probability theory at all levels and for very wide range of topics and applications, and hence detailed awareness and deep understanding of them are essential for a comprehensive investigation of the probability theory.

6.1 Bayes Theorem

The Bayes theorem (which is about conditional probability and may also be called Bayes rule among other names and labels) is given by:

$$P(A\,|B) = \frac{P(A)\,P(B\,|A)}{P(B)} \qquad \Big[P(B) \neq 0\Big] \tag{149}$$

where $P(A)$ is the **prior** probability of event A (i.e. the probability of event A before the occurrence of event B), $P(A\,|B)$ is the **posterior** probability of event A (i.e. the probability of event A after the occurrence of event B), $P(B)$ is the probability of event B and $P(B\,|A)$ is the conditional probability of event B. Bayes rule (i.e. Eq. 149) can be simply obtained from Eq. 47 (or rather it is a manipulated form of Eq. 47).

Problems

1. Prove the following equation:
$$P(B) = \sum_{i=1}^{n} P(A_i)\,P(B\,|A_i) \tag{150}$$
where A_i $(i = 1, \ldots, n)$ are mutually exclusive random events whose union is the entire sample space, and B is a random event in this sample space.
Answer: We proved this earlier (see part b of Problem 14 of § 3.3). In brief:
$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(A_i)\,P(B\,|A_i)$$
where step 1 is because A_i's are mutually exclusive and their union represents the entire sample space (and hence the intersections $B \cap A_i$ are mutually exclusive and represent all parts of B; see Eqs. 7 and 53), while the second step is from Eq. 46.
PE: It is claimed that Eq. 150 is an instance of the Bayes rule (i.e. Eq. 149) despite the fact that Eq. 150 is about unconditional probability $\big[$i.e. $P(B)\big]$ while Eq. 149 is about conditional probability $\big[$i.e. $P(A\,|B)\big]$. Try to justify this claim.
2. Show that the Bayes theorem can be written as:
$$P(A\,|B) = \frac{P(A)\,P(B\,|A)}{\sum_{i=1}^{n} P(A_i)\,P(B\,|A_i)} \tag{151}$$
$$P(A\,|B) = \frac{P(A)\,P(B\,|A)}{P(A)\,P(B\,|A) + P(\overline{A})\,P(B\,|\overline{A})} \tag{152}$$

$$\frac{P(A\,|B)}{P(C\,|B)} = \frac{P(A)\,P(B\,|A)}{P(C)\,P(B\,|C)} \tag{153}$$

$$\frac{P(B\,|A)}{P(B\,|C)} = \frac{P(C)\,P(A\,|B)}{P(A)\,P(C\,|B)} \tag{154}$$

where A, B, C, D are events in the sample space and A_i's (in Eq. 151) are mutually exclusive events whose union is the sample space.[137]

Answer:

• Eq. 151 is the same as Eq. 149 with $P(B)$ given by Eq. 150.

• Eq. 152 is a special case of Eq. 151 (where the sample space is divided into A and $\overline{A}$).

• Eq. 153 is obtained by dividing $P(A\,|B)$ by $P(C\,|B)$ where these probabilities are obtained from Eq. 149.

• Eq. 154 is obtained by multiplying the two sides of Eq. 153 by $P(C)/P(A)$.

Note: each one of the above forms of the Bayes theorem (as well as the original form of Eq. 149) has its own uses and applications where a given form is either necessary or more appropriate to use depending on the cases and circumstances (which are determined, for instance, by the available data and information or by convenience).

PE: Give some examples where Eqs. 151-154 are necessary or convenient to use instead of Eq. 149.

3. The Monty Hall problem (which is controversial and may be classified as a paradox) is regarded as an instance for the application of Bayes rule. The problem is stated as follows (which we quote):
Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice? (End of quote)
Solve the Monty Hall problem by applying the Bayes theorem.
Answer: Let adopt the following:
• A is the event that "the contestant has chosen the car door".
• B is the event that "the host reveals a goat door".
• $P(A\,|B)$ is the (posterior) probability that the contestant has chosen the car door given that the host reveals a goat door. This is the probability that we want to find.
• $P(A)$ is the (prior) probability that the contestant has chosen the car door. This probability is obviously 1/3 since all three doors are equally likely to be the car door.
• $P(B\,|A)$ is the (conditional) probability that the host reveals a goat door given that the contestant has chosen the car door. This probability is obviously 1 since by the rules of the game the host cannot reveal the car door in any circumstance and under any condition during the game.
• $P(B)$ is the probability that the host reveals a goat door. Again, this probability is obviously 1 since the host cannot reveal the car door.
So, from Eq. 149 we get:

$$P(A\,|B) = \frac{P(A)\,P(B\,|A)}{P(B)} = \frac{(1/3)\times 1}{1} = \frac{1}{3}$$

This means that the contestant has a 1/3 chance of winning the car if he keeps his choice, and hence he has a 2/3 chance of winning the car if he switches his choice (noting that after the host reveals a goat door "keeping his choice" and "switching his choice" are complementary). So, it is to the advantage of the contestant to switch his choice.
PE: Although the theory of probability is generally based on common sense, the result of the Monty Hall problem may not seem (to some) consistent with common sense.[138] Try to explain and justify this by investigating what is weird about this result.

[137] We note that A in Eq. 151 is not the union of A_i's.

[138] In my view, the result if not intuitive is at least not counter-intuitive (although this judgment may be based on experience rather than instinctive intuition and common sense).

Table 7: The table of Problem 4 of § 6.1. DoC means "Door of Car" and CoC means "Choice of Contestant".

CoC \ DoC	**1**	**2**	**3**
1	2,3	3	2
2	3	1,3	1
3	2	1	1,2

CoC \ DoC	**1**	**2**	**3**
1	$\mathcal{X}$	✓	✓
2	✓	$\mathcal{X}$	✓
3	✓	✓	$\mathcal{X}$

4. Solve the Monty Hall problem (see Problem 3) by simple count.
   Answer: In Table 7 we have two sub-tables. The columns in both these sub-tables represent the number of the door behind which the car is located, while the rows in these sub-tables represent the choice of the contestant about the number of the door.
   The entries inside the cells of the left sub-table represent the choice of the host (i.e. the door which he opens after the initial choice of the contestant). For example, the cell in column 1 and row 1 means that if the car is behind door 1 and the contestant chooses door 1 then the host will open either door 2 or door 3, while the cell in column 3 and row 2 means that if the car is behind door 3 and the contestant chooses door 2 then the host will open door 1.
   The entries inside the cells of the right sub-table represent the result of the game show if the contestant switches his choice where ✓ means win (for the contestant) and $\mathcal{X}$ means loss. As we see, there are 6 possibilities (out of 9) for the win and hence we can conclude that the contestant has a 2/3 chance of win if he changes his initial choice. So, it is to the advantage of the contestant to switch his choice.
   PE: Demonstrate and justify the patterns in Table 7.
5. Solve the Monty Hall problem (see Problem 3) by simulation.
   Answer: We solved the Monty Hall problem by simulation using C++ programming language (see MontyHall.cpp code). The results are similar to the results of the previous Problems. The simulation result improves as the number of runs increases and it converges to the theoretical result of Problem 3 as the number of runs becomes large.
   PE: Plot a flowchart representing the algorithm of the MontyHall.cpp code.
6. Make an argument in support of the result obtained in the previous Problems about the Monty Hall problem.
   Answer: We present in the following an argument in support of the result obtained in the previous Problems:
   Let identify the three doors as α, β and γ. When the contestant chooses a particular door (say door α) he has a 1/3 chance of winning. This means that there is a 2/3 chance that the car is behind β or γ. Now, the host cannot choose door α (because it is already chosen by the contestant) and hence the host has no access to this door. This means that the host has no ability to change its initial probability of 1/3 because the probability of door α can be changed only if it is accessible by the intervention of the host and can be affected by it. So, when the host chooses one of the two remaining doors (i.e. β or γ) he will certainly not choose the door of the car and he cannot choose door α. Accordingly, when the host opens a door (which is certainly a goat door) the entire 2/3 probability of the two remaining doors will transfer to the door that is not chosen by the contestant or by the host. This means that the door chosen neither by the contestant nor by the host has a 2/3 probability of being the car door, and therefore it is to the advantage of the contestant to switch his choice.
   To sum up, the prior probability $P(A)$ is 1/3 and there is no reason for this probability to change after the host reveals a goat door, i.e. $P(A|B) = P(A)$. This means that the contestant has a 1/3 chance of winning if he keeps his original choice and a 2/3 chance of winning if he changes his original choice and hence it is to his advantage to switch his choice.
   PE: Assess the argument given in the answer of this Problem, pointing out to any potential challenges to this argument. Also, make (if you can) another argument in support of the result obtained in the

previous Problems or (at least) improve the argument given in the answer of this Problem.

7. Let have 3 urns u_1, u_2, u_3 each of which contains 5 balls. The balls of u_1 are all white, the balls of u_2 are 4 white and 1 black, and the balls of u_3 are 3 white and 2 black. One of these urns is selected randomly (without identifying which urn is) and 3 balls from it are drawn where they are found to be all white. What is the probability that the selected urn is u_1?
Answer: This Problem can obviously be solved by Bayes theorem. So, let adopt the following:
• A_1 is the event that "the selected urn is u_1".
• B is the event that "the 3 drawn balls are all white".
• $P(A_1\,|B)$ is the (posterior) probability that the selected urn is u_1 given that the 3 drawn balls are all white. This is the probability that we want to find.
• $P(A_1)$ is the (prior) probability that the selected urn is u_1. This probability is obviously 1/3 since all three urns are equally likely to be selected (noting that the selection is random).
• $P(B\,|A_1)$ is the (conditional) probability that the 3 drawn balls are all white given that the selected urn is u_1. This probability is obviously 1 since u_1 contains only white balls.
• $P(B)$ is the probability that the 3 drawn balls are all white. This probability is not obvious and hence we need to calculate it. However, it should be obvious that $P(B)$ represents the number of possibilities of "drawing 3 white balls" from all the three urns divided by the total number of possibilities of "drawing 3 balls" from all the three urns. Now, the number of possibilities of "drawing 3 white balls" from all the three urns is (noting that C_3^5, C_3^4, C_3^3 correspond respectively to u_1, u_2, u_3):

$$C_3^5 + C_3^4 + C_3^3 = 10 + 4 + 1 = 15$$

while the total number of possibilities of "drawing 3 balls" from all the three urns is:

$$C_3^5 + C_3^5 + C_3^5 = 10 + 10 + 10 = 30$$

Accordingly, $P(B) = 15/30 = 1/2$.
So, from Eq. 149 (with A_1 here corresponding to A in Eq. 149) we get:

$$P(A_1\,|B) = \frac{P(A_1)\,P(B\,|A_1)}{P(B)} = \frac{(1/3)\times 1}{1/2} = \frac{2}{3}$$

PE: Repeat the Problem assuming this time that each of u_1, u_2, u_3 contains 6 balls where the balls of u_1 are all white, the balls of u_2 are 5 white and 1 black, and the balls of u_3 are 4 white and 2 black.

8. Referring to Problem 7, what is the probability that the selected urn is (**a**) u_2 (**b**) u_3?
Answer:
(**a**) If we repeat the argument of Problem 7 $\big[$noting that in this case we use A_2 to represent the event that "the selected urn is u_2" and hence $P(B\,|A_2) = C_3^4/C_3^5 = 4/10\big]$, then from Eq. 149 (with A_2 here corresponding to A in Eq. 149) we get:

$$P(A_2\,|B) = \frac{P(A_2)\,P(B\,|A_2)}{P(B)} = \frac{(1/3)\times(4/10)}{1/2} = \frac{4}{15}$$

(**b**) If we repeat the argument of Problem 7 $\big[$noting that in this case we use A_3 to represent the event that "the selected urn is u_3" and hence $P(B\,|A_3) = C_3^3/C_3^5 = 1/10\big]$, then from Eq. 149 (with A_3 here corresponding to A in Eq. 149) we get:

$$P(A_3\,|B) = \frac{P(A_3)\,P(B\,|A_3)}{P(B)} = \frac{(1/3)\times(1/10)}{1/2} = \frac{1}{15}$$

We may also obtain this probability more simply from the previous probabilities, i.e. $1-(2/3)-(4/15) = 1/15$ $\big[$noting that $(A_1\,|B), (A_2\,|B), (A_3\,|B)$ are complementary events$\big]$.
PE: Repeat the Problem for the PE of Problem 7.

9. Confirm the results of Problems 7 and 8 by simulation.
Answer: We simulated those Problems using C++ programming language (see Urn3W1.cpp code for Problem 7 and Urn3W2.cpp and Urn3W3.cpp codes for Problem 8). The results are similar to the results of those Problems. The simulation results improve as the number of runs increases and they converge to the theoretical results of Problems 7 and 8 as the number of runs becomes large.
PE: Try to modify the Urn3W1.cpp, Urn3W2.cpp and Urn3W3.cpp codes for the PE of Problems 7 and 8.
10. Referring to Problem 7, if we draw a fourth ball (following the drawing of 3 white balls) what is the probability that this ball is (**a**) white (**b**) black?
Answer:
(**a**) Let adopt the following:
• A_{u_i} $(i = 1, 2, 3)$ is the event that "the fourth ball drawn from urn u_i is white" (i.e. following the drawing of 3 white balls from the selected urn).
• B_{u_i} $(i = 1, 2, 3)$ is the event that "the selected urn is u_i" (i.e. following the drawing of 3 white balls from the selected urn).
The required probability P_w is the *sum* (over $i = 1, 2, 3$) of the (exhaustive) probabilities of "getting a fourth white ball from urn u_i given that 3 white balls were drawn from the selected urn". The *sum* is justified by the fact that these events (i.e. "getting a fourth ... selected urn") are members of a union of disjoint events (because they correspond to different urns) and hence their probabilities are subject to the addition law of probability for mutually exclusive events (see Eq. 53).
Now, each one of these three probabilities is the *product* of the probability of "the fourth ball drawn from urn u_i is white given that the selected urn is u_i" times the probability that "the selected urn is u_i". The *product* is justified by the fact that we are looking for the probability of intersection of A_{u_i} and B_{u_i} (since we are looking for "A_{u_i} AND B_{u_i}") and hence it is subject to the multiplication law for associated events (see Eq. 46). Accordingly, we get:[139]

$$\begin{aligned} P_w &= \sum_{i=1}^{3} P(A_{u_i} \cap B_{u_i}) = \sum_{i=1}^{3} P(A_{u_i} | B_{u_i})\, P(B_{u_i}) \\ &= P(A_{u_1} | B_{u_1})\, P(B_{u_1}) + P(A_{u_2} | B_{u_2})\, P(B_{u_2}) + P(A_{u_3} | B_{u_3})\, P(B_{u_3}) \\ &= \left(1 \times \frac{2}{3}\right) + \left(\frac{1}{2} \times \frac{4}{15}\right) + \left(0 \times \frac{1}{15}\right) = \frac{4}{5} \end{aligned}$$

(**b**) The events of "getting white" and "getting black" in the fourth draw are complementary and hence the required probability is $P_b = 1 - P_w = 1 - (4/5) = 1/5$. To check, we follow the method of part a (with replacement of "white" by "black"), that is:

$$P_b = \left(0 \times \frac{2}{3}\right) + \left(\frac{1}{2} \times \frac{4}{15}\right) + \left(1 \times \frac{1}{15}\right) = \frac{1}{5}$$

PE: Repeat the Problem for the PE of Problem 7.
11. Resolve Problem 10 by using Eq. 150.
Answer:
(**a**) Let the symbols of Eq. 150 represent the following:
• A_i $(i = 1, 2, 3)$ is the event that "the selected urn is u_i" (i.e. following the drawing of 3 white balls from the selected urn).
• B is the event that "the fourth ball is white" (i.e. following the drawing of 3 white balls from the selected urn).

[139] We note that $P(A_{u_1} | B_{u_1}) = 1$ because all the balls of u_1 are white, $P(A_{u_2} | B_{u_2}) = 1/2$ because after drawing three white balls u_2 contains only one black ball and one white ball, and $P(A_{u_3} | B_{u_3}) = 0$ because after drawing three white balls u_3 contains only two black balls. Regarding $P(B_{u_i})$ $(i = 1, 2, 3)$ they were calculated in Problems 7 and 8, i.e. $P(B_{u_i}) = P(A_i | B)$.

Now, from the results of Problems 7 and 8 we have:

$$P(A_1) = \frac{2}{3} \qquad P(A_2) = \frac{4}{15} \qquad P(A_3) = \frac{1}{15}$$

Moreover, from the discussion of Problem 10 we have:

$$P(B\,|A_1) = 1 \qquad P(B\,|A_2) = \frac{1}{2} \qquad P(B\,|A_3) = 0$$

Hence, from Eq. 150 we get:

$$P(B) = \sum_{i=1}^{3} P(A_i)\,P(B\,|A_i) = \left(\frac{2}{3}\times 1\right) + \left(\frac{4}{15}\times\frac{1}{2}\right) + \left(\frac{1}{15}\times 0\right) = \frac{24}{30} = \frac{4}{5}$$

(**b**) If $\overline{B}$ means "the fourth ball is black" (i.e. following the drawing of 3 white balls from the selected urn) then $P(\overline{B}) = 1 - P(B) = 1 - (4/5) = 1/5$. Alternatively:

$$P(\overline{B}) = \sum_{i=1}^{3} P(A_i)\,P(\overline{B}\,|A_i) = \left(\frac{2}{3}\times 0\right) + \left(\frac{4}{15}\times\frac{1}{2}\right) + \left(\frac{1}{15}\times 1\right) = \frac{6}{30} = \frac{1}{5}$$

PE: Compare the method of solution of the present Problem to the method of solution of Problem 10 and comment.

12. Confirm the results of Problem 10 by simulation.
Answer: We simulated that Problem using C++ programming language (see Urn4thW.cpp and Urn4thB.cpp codes). The results are similar to the results of Problem 10. The simulation results improve as the number of runs increases and they converge to the theoretical results of Problem 10 as the number of runs becomes large.
PE: Try to modify the Urn4thW.cpp and Urn4thB.cpp codes for the PE of Problem 10.
13. We have two boxes: b_1 and b_2 where b_1 contains r_1 red cards and g_1 green cards while b_2 contains r_2 red cards and g_2 green cards. We draw a card randomly from b_1 and put it in b_2. We then draw a card randomly from b_2. What is the probability that the card drawn is red?
Answer: This Problem can obviously be solved by Bayes theorem. In fact, we will use Eq. 150. So, let adopt the following: A_1, A_2, R are the events (correspondingly) of drawing a red card from b_1, drawing a green card from b_1, and drawing a red card from b_2. So, what we want to find is $P(R)$. Now:

$$P(A_1) = \frac{r_1}{r_1+g_1} \qquad P(A_2) = \frac{g_1}{r_1+g_1} \qquad P(R\,|A_1) = \frac{r_2+1}{r_2+g_2+1} \qquad P(R\,|A_2) = \frac{r_2}{r_2+g_2+1}$$

Hence, from Eq. 150 we get:

$$\begin{aligned} P(R) &= \sum_{i=1}^{2} P(A_i)\,P(R\,|A_i) = P(A_1)\,P(R\,|A_1) + P(A_2)\,P(R\,|A_2) \\ &= \left(\frac{r_1}{r_1+g_1}\right)\left(\frac{r_2+1}{r_2+g_2+1}\right) + \left(\frac{g_1}{r_1+g_1}\right)\left(\frac{r_2}{r_2+g_2+1}\right) = \frac{r_1+r_1r_2+g_1r_2}{(r_1+g_1)\,(r_2+g_2+1)} \end{aligned}$$

PE: Justify the use of Eq. 150.

6.2 Limit Theorems

There are many limit theorems related to probability and probability distributions (as well as associated random variables and their parameters) which are frequently used in derivations and calculations in the probability theory and related subjects. Some of these theorems have been met earlier in this book. In the following subsections we briefly discuss some of the most common of these theorems.

6.2.1 Stirling Formula Theorem

This is related to the approximation of the factorial by the Stirling formula which we discussed earlier (see § 2.3.2). As a limit theorem, this formula can be written as:

$$\lim_{n\to\infty} n! = \sqrt{2\pi n}\, n^n\, e^{-n} \tag{155}$$

6.2.2 Theorems about Convergence of Distributions

There are many theorems related to the convergence of some probability distributions to other probability distributions under certain conditions and in special circumstances (where the converging and converged-to distributions could be both discrete or both continuous or mixed). These limit theorems include for instance:

• The convergence of the binomial distribution to the Poisson distribution under certain conditions (see Problem 1).

• The convergence of the binomial distribution to the normal distribution under certain conditions (see Problem 1).

• The convergence of the Poisson distribution to the normal distribution under certain conditions (see Problem 1).

• The convergence of the hypergeometric distribution to the binomial distribution under certain conditions (see Problem 5 of § 4.1.6).

These theorems (and similar theorems) were investigated or indicated earlier in various places of chapter 4 and hence they do not require further investigation. In fact, there are other limit theorems about the convergence of some distributions to other distributions which we did not mention or investigate (e.g. the convergence of the negative binomial and hypergeometric distributions to the Poisson distribution under certain conditions). We may also include (loosely and in a rather different sense) in this type of limit theorems the propositions about the special cases (e.g. "The Bernoulli distribution is a special case of the binomial distribution corresponding to $n = 1$" or "The geometric distribution is a special case of the negative binomial distribution corresponding to $r = 1$").

Problems

1. Outline the "convergence theorems"[140] of the binomial, Poisson and normal probability distributions.
   Answer:
   • **Binomial to Poisson**: as n tends to infinity and p tends to zero (with $\mu = np$ remaining finite and constant), the binomial distribution converges to the corresponding Poisson distribution (with $\mu = \lambda$). Also see Problems 2 and 3.
   • **Binomial to normal**: as n tends to infinity and p remains finite (so $\mu = np$ becomes very large), the binomial distribution converges to the corresponding normal distribution $\big[$with $x = k$, $\mu = np$ and $V = np(1-p)\big]$.
   • **Poisson to normal**: for large k and λ the Poisson distribution converges to the corresponding normal distribution (with $x = k$ and $\mu = V = \lambda$).
   Note: the last statement (i.e. about the convergence of Poisson to normal) is what we found in the literature (noting that some may not impose the condition of large λ). However, from the comparison of Problem 10 of § 4.2.2 we can see that these conditions are not sufficient (and may not even be necessary in some cases). In our view, if the Poisson distribution should converge to the normal distribution (considering that both can be seen as limits to the binomial, as outlined in the first two theorems, and hence we can take the binomial as a reference for their convergence) then we should have $\mu = \lambda \simeq np$ and $V = \lambda \simeq np(1-p)$ which on comparison should lead to $1 - p \simeq 1$ which means that p must be small. In fact, the result of of Problem 10 of § 4.2.2 (as seen in Figure 26) should support this condition. We should also remember (see Problem 1 and § 4.1.4) that for the Poisson

[140] The "convergence theorems" is a term that we use to label these theorems (which may also be labeled by some as "the laws of large numbers"). So, this is not a standard term and hence it should not be confused with similar terms found in the literature.

distribution to be a limiting case (and hence a good approximation) to the binomial distribution we should have $p \to 0$ (as well as $n \to \infty$) and hence the condition of small p can also be obtained from the condition of $p \to 0$. Also see Problems 3 and 4.

Anyway, the reader should be aware that there is some mess and lack of clarity (as well as potential contradiction and lack of sufficient details) about some of the limit theorems related to the convergence of some distributions to other distributions, and hence the reader should be generally cautious about this issue. In fact, from our personal experience we found many cases in which some of these theorems fail in certain circumstances, and this should indicate that the given conditions are inaccurate or insufficient or not general.

PE: What can you conclude from the above convergence theorems?

2. Show that the binomial distribution converges to the Poisson distribution when the number of trials n becomes large and hence the probability of occurrence p becomes small (assuming $\lambda = np$ remains constant).

Answer: For the binomial distribution $\mu = np$ and for the Poisson distribution $\mu = \lambda$. Hence, if we have to compare these distributions sensibly (by assuming that they give similar results according to the requirement of convergence) then we should take $\lambda = np$ and hence $p = \lambda/n$.[141] Now, if we write Eq. 71 (of binomial) in terms of this expression of p then we have:

$$\begin{aligned}
P(k) &= C_k^n \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n \times (n-1) \times \cdots \times (n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n}{n} \times \frac{(n-1)}{n} \times \cdots \times \frac{(n-k+1)}{n} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \left[\frac{n}{n} \times \frac{(n-1)}{n} \times \cdots \times \frac{(n-k+1)}{n}\right] \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-k}
\end{aligned}$$

Now, if n becomes large (with k fixed) then all the fractions inside the square brackets (in the last line) tend to unity, and this also applies to the last factor (noting that λ is fixed).[142] Accordingly, we get:

$$P(k) \simeq \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^n \simeq \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{(large } n\text{)}$$

where the second step is based on a well known result from calculus.[143] As we see, this is the same as Eq. 75.

Note: the statement in this Problem may be given more rigorously as follows: when n approaches infinity and p approaches 0 such that np remains finite and constant, the binomial distribution (parameterized by n and p) converges to the corresponding Poisson distribution (parameterized by $\lambda \equiv \mu = np$). In the following we prove this statement (more rigorously) using the technique of limits.

From calculus we have (with k kept finite):

$$\lim_{n\to\infty} \left[\frac{n!}{(n-k)!\, n^k}\right] = \lim_{n\to\infty} \left[\frac{n \times (n-1) \times \cdots \times (n-k+1)}{n^k}\right]$$

[141] In fact, we should also consider the condition $\lambda = np(1-p)$ which we indicated earlier. However, the condition $p \to 0$ should ensure that $(1-p)$ tends to unity (although this applies to the limit but not necessarily to cases of approximation).

[142] We note that the condition "the probability of occurrence p becomes small" (which we stated above) is considered implicitly in $p = \lambda/n$ noting that λ is supposedly fixed.

[143] We refer to the following relation:

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

Table 8: The table of Problem 3 of § 6.2.2.

n	k	p	λ	Binomial	Poisson	% Difference
50	1	0.02	1	0.3716	0.3679	1.0
100	2	0.01	1	0.1849	0.1839	0.5
150	6	0.02	3	0.0499	0.0504	1.0
200	7	0.05	10	0.0896	0.0901	0.6
300	5	0.02	6	0.1617	0.1606	0.7
1000	14	0.01	10	0.0520	0.0521	0.1

$$= \lim_{n\to\infty}\left(\frac{n}{n}\right) \times \lim_{n\to\infty}\left(\frac{n-1}{n}\right) \times \cdots \times \lim_{n\to\infty}\left(\frac{n-k+1}{n}\right) = 1 \times 1 \times \cdots \times 1 = 1$$

Moreover, if we note that $p = \lambda/n$ (since $\lambda = np$) and k is finite then we have (using calculus again):

$$\begin{aligned}
\lim_{n\to\infty}(1-p)^{n-k} &= \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n-k} = \lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^{n}\left(1-\frac{\lambda}{n}\right)^{-k}\right] \\
&= \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} \times \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} = \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} \times (1-0)^{-k} \\
&= \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n} \times 1 = e^{-\lambda}
\end{aligned}$$

Now, if we substitute these results into the equation of the binomial distribution (see Eq. 71) considering the above assumptions we get:

$$\begin{aligned}
P(k) &= C_k^n p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \frac{n!\,n^k}{k!(n-k)!\,n^k}p^k(1-p)^{n-k} \\
&= \left(\frac{n!}{(n-k)!\,n^k}\right)\frac{n^k}{k!}\,p^k(1-p)^{n-k} = 1 \times \frac{n^k}{k!}\,p^k(1-p)^{n-k} = \frac{n^k}{k!}\left(\frac{\lambda}{n}\right)^k e^{-\lambda} = \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

which is the same as the Poisson distribution (see Eq. 75).

PE: Explain and justify each step in the above derivations.

3. Calculate the Poisson probability corresponding to a number of binomial cases (where p of the binomial is small considering a number of values of n) and hence compare the two distributions showing that the Poisson distribution is a good approximation to the binomial distribution in these cases.

   Answer: We made such a comparison in Table 8.

   Note: the results of this Problem indicate that the approximation is good when p is small even when n is relatively small, and this is consistent in part with our observation in the note of Problem 1.

   PE: Do more comparisons using different values of n and larger values of k and p. Try to make some observations on how these changes affect the results.

4. Make plots of the binomial distribution for $n = 140$ with $p = 0.05, 0.20, 0.40, 0.60, 0.80, 0.95$ and their corresponding Poisson distribution.

   Answer: See Figure 28.

   PE: Do the following:

   • Comment on the effect of varying p on the quality of agreement between the binomial distribution and the corresponding Poisson distribution (considering the notes of Problems 1 and 3).

   • Investigate the effect of varying p on properties other than the quality of agreement between the binomial distribution and the corresponding Poisson distribution.

   • Try to make similar plots for different n and hence investigate the effect of varying n on the quality of agreement between the binomial distribution and the corresponding Poisson distribution (as well as other potential effects).

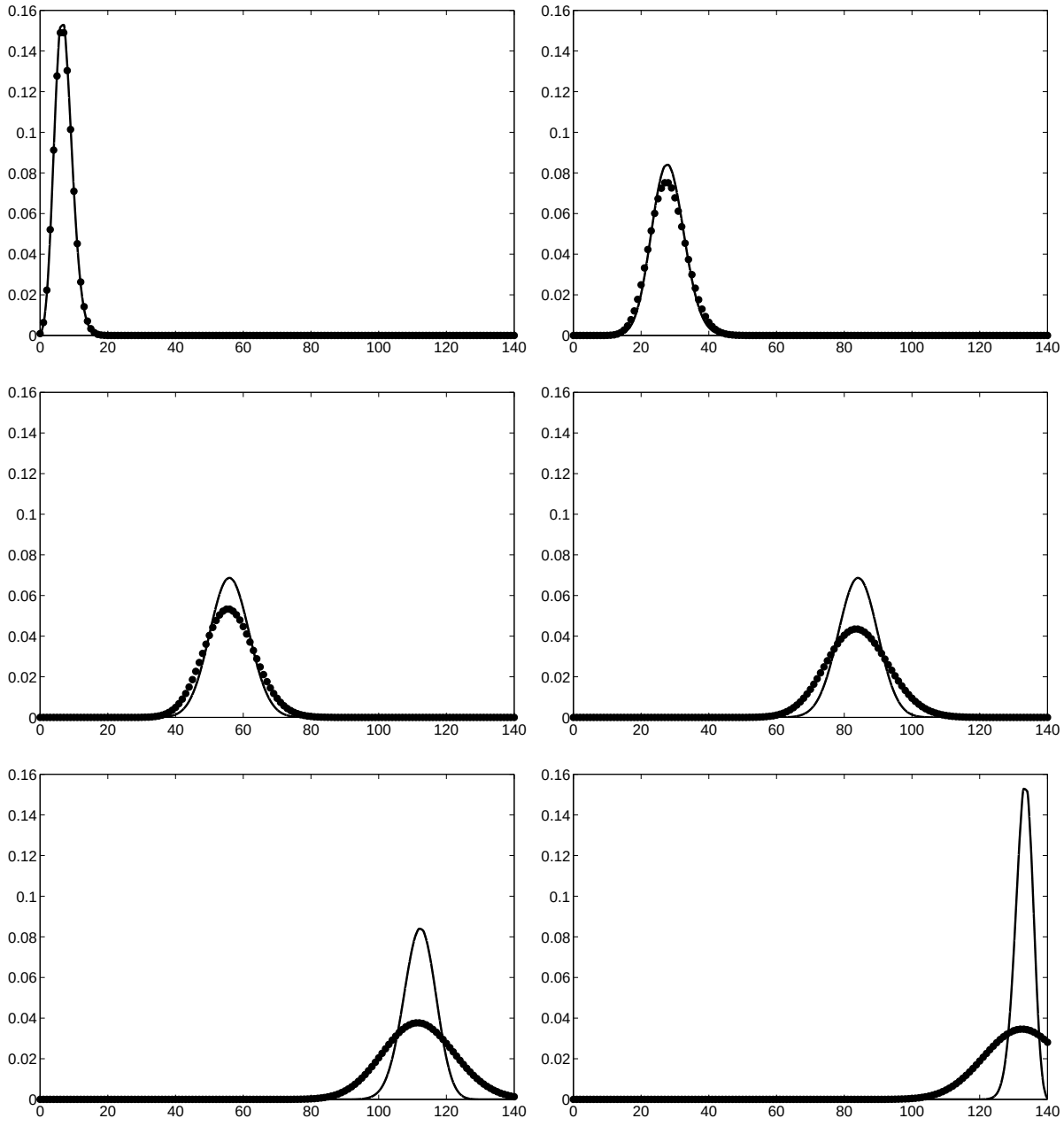


Figure 28: Comparison between the binomial distribution (solid curve) and the corresponding Poisson distribution (filled circles) for $n = 140$ with $p = 0.05$ (top left), $p = 0.20$ (top right), $p = 0.40$ (middle left), $p = 0.60$ (middle right), $p = 0.80$ (bottom left) and $p = 0.95$ (bottom right). For the corresponding Poisson distribution we have $\lambda = np$ in each case. The horizontal axis in each frame represents k and the vertical axis represents the probability of the distribution P. See Problem 4 of § 6.2.2.

6.2.3 Central Limit Theorem

The central limit theorem can be seen as a generalization of other (more special) limit theorems (such as some of those indicated in § 6.2.2). According to this theorem if $X_1, X_2, \ldots, X_n$ are independent random variables represented by probability functions $f_1, f_2, \ldots, f_n$ with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $V_1, V_2, \ldots, V_n$ then the "mean" Z of these random variables, i.e.

$$Z = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{156}$$

possesses the following properties:
(**a**) The mean of Z is the average value of $\mu_1, \mu_2, \ldots, \mu_n$, that is:

$$\mu(Z) = \frac{1}{n}\sum_{i=1}^{n} \mu_i \tag{157}$$

(**b**) The variance of Z is the average value of $V_1, V_2, \ldots, V_n$ divided by n, that is:

$$V(Z) = \frac{1}{n}\sum_{i=1}^{n} \frac{V_i}{n} = \frac{1}{n^2}\sum_{i=1}^{n} V_i \tag{158}$$

(**c**) As n goes to infinity, the probability distribution function of Z $\big[$i.e. $g(Z)\big]$ tends to a normal distribution with mean $\mu_Z = \mu(Z)$ and variance $V_Z = V(Z)$, that is:

$$\lim_{n\to\infty} g(Z) = \frac{1}{\sqrt{2\pi V_Z}} e^{-\frac{(x-\mu_Z)^2}{2V_Z}} \tag{159}$$

In the following points we discuss briefly some issues related to the central limit theorem:
• The central limit theorem is stated in the literature in various shapes and forms and with different flavors and terminologies (e.g. some represent probability approach and others represent statistics approach, and some are more general than others). So, the above statement is just one of these variants. In fact, there may even be differences in significance and content (i.e. we have central limit *theorems*). Therefore, those who have special interest in this theorem should determine the context and purpose of this theorem which they want so that they choose the statement that is more suitable to their objectives.
• It is obvious that if the central limit theorem should apply then the probability functions $f_1, f_2, \ldots, f_n$ must have means and variances (and hence pathological distributions, such as the Cauchy distribution, should be excluded from the domain of this theorem).
• The central limit theorem should explain (in part) what may have been noticed (or guessed) previously that probability distributions (such as binomial) generally converge to the normal distribution when n becomes large (within certain conditions). The details of this should be sought in the literature.

6.2.4 Large Numbers Theorem

The essence of this theorem (which is commonly known as the **law of large numbers**) is that as the sample size of a random variable increases, its mean becomes closer to the mean of the entire population.[144] This should sound reasonable because as the sample increases in size it becomes more representative of the population and hence its mean approaches the mean of the population. It is noteworthy that the **Bernoulli theorem** is a special case of the law of large numbers (see Problem 1). In fact, there are many details and theorems related to the law of large numbers,[145] so the interested reader should consult the literature about this issue.

[144] We are assuming the sample is not biased.

[145] Accordingly, we can say we have "large numbers theorems", i.e. plural. In fact, even the theorem of Stirling (see § 6.2.1) and the theorems of convergence (see § 6.2.2) and their alike may be classified in this category and hence they are labeled as "large numbers theorems". However, the reader should be aware of the difference in meaning between these labels (as well as the difference in significance and essence between the intended theorems) to avoid confusion.

Problems

1. Outline the Bernoulli theorem.
   Answer: According to the **Bernoulli theorem**, the relative frequency of a given outcome (say success) in a number of Bernoulli trials tends to the probability of that outcome (i.e. success) as the number of trials tends to infinity. For example, in a sequence of (fair) coin-tossing trials the probability of H (i.e. getting head) is $1/2$, but in a limited number of trials the relative frequency does not necessarily reflect this probability. So, if we toss the coin 4 times we may get H three times (representing a relative frequency of 0.75) rather than the "expected" two times (which is the frequency of occurrence that reflects the probability of $1/2$ for getting H). However, as we increase the number of trials we will see that the relative frequency fluctuates around this probability and converges towards it. So, in a 100 trials we may get 45 H's (with a relative frequency of 0.45 which is closer to $1/2$ than 0.75), and in a 1000 trials we may get 511 H's (with a relative frequency of 0.511 which is closer to $1/2$ than 0.75 and 0.45). As indicated above, the Bernoulli theorem is a special case (or an instance or an application) of the law of large numbers (depending on how it is stated).
   PE: The Bernoulli theorem may be stated in terms of the average of the outcomes (rather than the relative frequency as we did above):
   (**a**) State this theorem in terms of the average.
   (**b**) Express your statement (or our statement) of the theorem mathematically as a limit relation, i.e. $\lim_{n\to\infty}(\cdots) = (\cdots)$.

6.3 Inequality Theorems

These are probability theorems in the form of inequalities. Most of these theorems are widely used and have many applications in the probability theory and related fields of mathematics and science. In fact, there are many inequality theorems (noting that some of these theorems are not restricted to probability but they have versions or instances related to probability or tailored for it). In the following subsections we present a few of these inequality theorems.

6.3.1 Markov Inequality

According to this theorem, if x is a random variable that takes only non-negative values and has a mean $\mu(x)$, then for any real number $c > 0$ we have:

$$P(x \geq c) \leq \frac{\mu(x)}{c} \tag{160}$$

For example, if x is a random variable that meets the above conditions then we know from this theorem without doing any substantial calculations (i.e. by just setting $c = 3\mu$ in the above equation) that the probability of x being not less than three times its mean is not larger than $1/3$. In fact, this theorem is very handy in many practical situations where quick estimates of probability and the limits on its value are required. The theorem is also useful in many theoretical arguments and analytical derivations related to probability (see for instance Problem 1 of § 6.3.3).

Problems

1. Prove the Markov inequality (considering the discrete case).
   Answer: We have:

$$\begin{aligned}
\mu(x) &= \sum_i p_i\, x_i && \text{(Eq. 115)} \\
&= \sum_{x_i<c} p_i\, x_i + \sum_{x_i\geq c} p_i\, x_i \\
&\geq \sum_{x_i\geq c} p_i\, x_i && (1^{st} \text{ sum } \geq 0 \text{ since } x_i \text{ are non-negative})
\end{aligned}$$

$$\begin{aligned} &\geq \sum_{x_i \geq c} c\, p_i && (x_i \geq c \text{ in this sum}) \\ &= c \sum_{x_i \geq c} p_i && (c \text{ is constant}) \\ &= c\, P(x \geq c) && (\text{addition rule for disjoint events}) \\ \frac{\mu(x)}{c} &\geq P(x \geq c) && (c > 0) \end{aligned}$$

PE: Prove the Markov inequality (considering the continuous case).

2. Show that the result of part (a) of Problem 2 of § 5.1 is consistent with the Markov inequality.
Answer: From the left hand side of Eq. 160 we have (with reference to Problem 2 of § 5.1):

$$P(x \geq c) = \int_c^4 \frac{x}{8}\, dx = \left[\frac{x^2}{16}\right]_c^4 = \frac{16 - c^2}{16} \qquad (0 < c \leq 4)$$

while from the right hand side of Eq. 160 we have (with reference to Problem 2 of § 5.1):

$$\frac{\mu(x)}{c} = \frac{8}{3c} \qquad (0 < c \leq 4)$$

Now, it is straightforward to show (e.g. by calculus or by a plot using for instance a spreadsheet) that $\frac{16-c^2}{16} - \frac{8}{3c}$ is non-positive for $0 < c \leq 4$ and hence $P(x \geq c) \leq \frac{\mu(x)}{c}$.
PE: Repeat the Problem for part (a) of the PE of Problem 2 of § 5.1.

6.3.2 Chebyshev Inequality

The essence of the Chebyshev inequality is that if x is a random variable with mean μ and standard deviation σ and k is a positive real number (> 1)[146] then the probability that x differs from μ by more than k standard deviations is $\leq 1/k^2$, that is:

$$P\Big(\,|x - \mu| \geq k\sigma\Big) \leq \frac{1}{k^2} \tag{161}$$

For example, the probability that x deviates from μ by more than 2σ is $\leq 1/4$ (corresponding to $k = 2$).

Problems

1. Prove the Chebyshev inequality (considering the discrete case).
Answer: From Eq. 137 (noting that $V = \sigma^2$ and $\mu \equiv \mu_x$) we have:

$$\sigma^2 = \sum_i (x_i - \mu)^2\, p_i \tag{162}$$

For the values of x for which $|x - \mu| \geq k\sigma$ the sum in the last equation should exclude (in general) some terms and hence we can write:

$$\sigma^2 \geq \sum_j (x_j - \mu)^2\, p_j$$

where j refers only to the terms (of Eq. 162) for which $|x - \mu| \geq k\sigma$. Now, if we replace $(x_j - \mu)^2$ in the last equation by $(k\sigma)^2$ (noting that the last equation includes only the terms for which $|x - \mu| \geq k\sigma$ and hence the inequality is not affected by this replacement) then we can write:

$$\sigma^2 \quad \geq \quad \sum_j (k\sigma)^2\, p_j$$

[146] It is obvious that the Chebyshev inequality is true trivially for $0 < k \leq 1$ (since any probability must be ≤ 1).

$$\begin{aligned}\sigma^2 &\geq (k\sigma)^2 \sum_j p_j \\ \frac{1}{k^2} &\geq \sum_j p_j\end{aligned}$$

Noting that p_j in the last equation represent the probabilities corresponding to $|x-\mu| \geq k\sigma$ we can rewrite the last equation as:

$$\frac{1}{k^2} \geq P\Big(|x-\mu| \geq k\sigma\Big)$$

which is the Chebyshev inequality.

PE: Prove the Chebyshev inequality (considering the continuous case).

2. Show that the results of Problem 6 of § 4.2.2 are consistent with the Chebyshev inequality.
Answer:
(**a**) For $k=1$ we have[147] (according to the Chebyshev inequality) $P\Big(|x-\mu| \geq \sigma\Big) \leq 1$, and from part (a) of Problem 6 of § 4.2.2 we have $P\Big(|x-\mu| \geq \sigma\Big) \simeq 1 - 0.6827 = 0.3173$, and hence the two results are consistent.
(**b**) For $k=2$ we have (according to the Chebyshev inequality) $P\Big(|x-\mu| \geq 2\sigma\Big) \leq 0.25$, and from part (b) of Problem 6 of § 4.2.2 we have $P\Big(|x-\mu| \geq 2\sigma\Big) \simeq 1 - 0.9545 = 0.0455$, and hence the two results are consistent.
(**c**) For $k=3$ we have (according to the Chebyshev inequality) $P\Big(|x-\mu| \geq 3\sigma\Big) \leq 1/9 \simeq 0.11$, and from part (c) of Problem 6 of § 4.2.2 we have $P\Big(|x-\mu| \geq 3\sigma\Big) \simeq 1 - 0.9973 = 0.0027$, and hence the two results are consistent.
PE: Explain (in words) the logic of our arguments above, e.g. why $P\Big(|x-\mu| \geq \sigma\Big) \simeq 1 - 0.6827$.

3. Show that the result of part (a) of Problem 2 of § 5.2 is consistent with the Chebyshev inequality.
Answer: From the left hand side of Eq. 161 we have (with reference to Problem 2 of § 5.2 as well as Problem 2 of § 5.1):[148]

$$\begin{aligned}P\Big(|x-\mu| \geq k\sigma\Big) &= 1 - P\Big(-k\sigma \leq x-\mu \leq k\sigma\Big) = 1 - P\Big(\mu - k\sigma \leq x \leq \mu + k\sigma\Big) \\ &= 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} \frac{x}{8}\,dx = 1 - \int_{(8/3)-k\sqrt{8/9}}^{(8/3)+k\sqrt{8/9}} \frac{x}{8}\,dx = \frac{9-4\sqrt{2}k}{9}\end{aligned}$$

It is straightforward to show (with no need for going through the details about the range of k) that $\frac{9-4\sqrt{2}k}{9} - \frac{1}{k^2}$ is non-positive for all positive k and hence we conclude that $P\Big(|x-\mu| \geq k\sigma\Big) \leq \frac{1}{k^2}$.
PE: Repeat the Problem for part (a) of the PE of Problem 2 of § 5.2 (with reference to Problem 2 of § 5.1).

6.3.3 One-Sided Chebyshev Inequality

The one-sided Chebyshev inequality is also known as the Cantelli inequality and the Chebyshev-Cantelli inequality. According to this theorem, if x is a random variable with mean μ and variance V and a is a positive real number then:

$$P(x \geq \mu + a) \leq \frac{V}{a^2 + V} \tag{163}$$

Problems

[147] Noting the condition $k > 1$, we do this case for demonstration.
[148] We also refer the reader to Problem 2 of § 4.

1. Prove the one-sided Chebyshev inequality (considering the continuous case).
Answer: If t is a real parameter such that $a+t>0$ then we have:

$$\begin{aligned}
P(x\geq\mu+a) &= P(x-\mu\geq a) && \text{(see Problem 2 of § 4)}\\
&= P(x-\mu+t\geq a+t) && \text{(see Problem 2 of § 4)}\\
&= P\left(\frac{x-\mu+t}{a+t}\geq 1\right) && \text{(see Problem 2 of § 4)}\\
&\leq P\left(\left[\frac{x-\mu+t}{a+t}\right]^2\geq 1\right) && \text{(see upcoming note)}\\
&\leq \mu\left(\left[\frac{x-\mu+t}{a+t}\right]^2\right) && \text{(Eq. 160)}\\
&= \mu\left(\frac{(x-\mu+t)^2}{(a+t)^2}\right)\\
&= \frac{\mu(x-\mu+t)^2}{(a+t)^2} && \text{(Eq. 121)}\\
&= \frac{\mu\left[x^2+2(-\mu+t)x+(-\mu+t)^2\right]}{(a+t)^2}\\
&= \frac{\mu(x^2)+2(-\mu+t)\mu(x)+(-\mu+t)^2}{(a+t)^2} && \text{(Eqs. 119-124)}\\
&= \frac{\mu(x^2)-2\mu^2+2t\mu+\mu^2-2\mu t+t^2}{(a+t)^2}\\
&= \frac{\mu(x^2)-\mu^2+t^2}{(a+t)^2}\\
&= \frac{V+t^2}{(a+t)^2} && \text{(Eq. 147)}
\end{aligned}$$

Now, the inequality $P(x\geq\mu+a)\leq\frac{V+t^2}{(a+t)^2}$ which we already obtained should be valid for any t as long as $a+t>0$. However, as a minimization inequality (i.e. $\leq$) it is appropriate to consider the extreme case of its validity which is when $\frac{V+t^2}{(a+t)^2}$ is minimum. To find the minimum of $\frac{V+t^2}{(a+t)^2}$ we treat t as a variable and differentiate $\frac{V+t^2}{(a+t)^2}$, that is:

$$\frac{d}{dt}\left[\frac{V+t^2}{(a+t)^2}\right]=\frac{2t(a+t)^2-2(a+t)(V+t^2)}{(a+t)^4}=\frac{2t(a+t)-2(V+t^2)}{(a+t)^3}=\frac{2(at-V)}{(a+t)^3}$$

As we see, this derivative is zero (and thus $\frac{V+t^2}{(a+t)^2}$ is minimum considering other tests and conditions)[149] when $(at-V)=0$, i.e. when $t=V/a$. On substituting this value of t in $\frac{V+t^2}{(a+t)^2}$ we get:

$$\left.\frac{V+t^2}{(a+t)^2}\right|_{\min}=\frac{V+(V/a)^2}{[a+(V/a)]^2}=\frac{a^2V+V^2}{(a^2+V)^2}=\frac{V\left(a^2+V\right)}{(a^2+V)^2}=\frac{V}{a^2+V}$$

So, finally we get $P(x\geq\mu+a)\leq\frac{V}{a^2+V}$ which is the one-sided Chebyshev inequality.

Note: let $z=\frac{x-\mu+t}{a+t}$ have the density function $f(z)$ and $w(z)=z^2=\left(\frac{x-\mu+t}{a+t}\right)^2$ have the density function $g(w)$. Referring to Eq. 63 in Problem 2 of § 4 we have:

$$g(w)=f\big[z(w)\big]\,\frac{dz}{dw}=f\big(\sqrt{w}\big)\times\frac{1}{\sqrt{4w}}=\frac{1}{\sqrt{4w}}\,f\big(\sqrt{w}\big) \tag{164}$$

[149] We are referring to the test of second derivative which is positive at $t=V/a$ and hence $t=V/a$ is a point of local minimum.

Now, for any $1 \leq z < z'$ we have (see § 4.3.2 and Eq. 102 in particular):

$$P\left(1 \leq z < z'\right) = \int_1^{z'} f(z)\, dz \tag{165}$$

Similarly $\left[\text{where } w' = w(z')\right]$:

$$\begin{aligned} P\left(1 \leq w < w'\right) &= \int_1^{w'} g(w)\, dw && \text{(Eq. 102)} \\ &= \int_1^{w'} \frac{1}{\sqrt{4w}}\, f\left(\sqrt{w}\right)\, dw && \text{(Eq. 164)} \\ &= \int_1^{z'^2} \frac{1}{2z}\, f(z)\; 2z\, dz && (w = z^2) \\ &= \int_1^{z'^2} f(z)\; dz && (166) \end{aligned}$$

Now, for $z' \geq 1$ we have $z' \leq z'^2$ and hence on comparing Eq. 165 and Eq. 166 we conclude:

$$\begin{aligned} P\left(z \geq 1\right) &\leq P\left(w \geq 1\right) \\ P\left(z \geq 1\right) &\leq P\left(z^2 \geq 1\right) \\ \text{that is:} \qquad P\left(\frac{x-\mu+t}{a+t} \geq 1\right) &\leq P\left(\left[\frac{x-\mu+t}{a+t}\right]^2 \geq 1\right) \end{aligned}$$

as required.

PE: Explain and justify the use of Problem 2 of § 4 and Eq. 63 in the above proof.

2. Show that the one-sided Chebyshev inequality is valid for the exponential distribution (see § 4.2.3).
Answer: From the left hand side of the one-sided Chebyshev inequality (i.e. Eq. 163) we have (with reference to Eqs. 94, 102 and 93):

$$\begin{aligned} P(x \geq \mu + a) &= P\left(x \geq \frac{1}{\alpha} + a\right) = \int_{\frac{1}{\alpha}+a}^{\infty} f(x)\, dx = \int_{\frac{1}{\alpha}+a}^{\infty} \alpha e^{-\alpha x}\, dx \\ &= \left[-e^{-\alpha x}\right]_{\frac{1}{\alpha}+a}^{\infty} = 0 + e^{-\alpha\left(\frac{1}{\alpha}+a\right)} = e^{-1-a\alpha} = \frac{1}{e^{1+a\alpha}} \end{aligned} \tag{167}$$

Also, from the right hand side of the one-sided Chebyshev inequality we have (with reference to Eq. 94):

$$\frac{V}{a^2+V} = \frac{\frac{1}{\alpha^2}}{a^2 + \frac{1}{\alpha^2}} = \frac{1}{a^2\alpha^2 + 1} \tag{168}$$

Now, let assume (tentatively) the validity of the one-sided Chebyshev inequality to see if it will lead to a correct result (and hence it is valid) or it will lead to a wrong result (and hence it is invalid), that is:

$$\begin{aligned} P(x \geq \mu + a) &\overset{?}{\leq} \frac{V}{a^2+V} && \text{(Eq. 163)} \\ \frac{1}{e^{1+a\alpha}} &\overset{?}{\leq} \frac{1}{a^2\alpha^2+1} && \text{(Eqs. 167 and 168)} \\ e^{1+a\alpha} &\overset{?}{\geq} a^2\alpha^2 + 1 \\ e^{1+z} &\overset{?}{\geq} z^2 + 1 && (z = a\alpha) \end{aligned}$$

$$e^{1+z} - z^2 - 1 \overset{?}{\geq} 0 \tag{169}$$

Now, $a > 0$ and $\alpha > 0$ and hence $z > 0$. So, the function $(e^{1+z} - z^2 - 1)$ is increasing because its derivative is positive for all $z > 0$. Moreover, at $z = 0$ we have $(e^{1+z} - z^2 - 1) > 0$. Accordingly, $(e^{1+z} - z^2 - 1) > 0$ for all $z > 0$ and hence the inequality of Eq. 169 (which is obtained tentatively from the one-sided Chebyshev inequality) is correct. So, we can conclude that the one-sided Chebyshev inequality is valid for the exponential distribution, as required.[150]

PE: Can we use the above argument (possibly with some modifications) to show the validity of the one-sided Chebyshev inequality for the geometric distribution (see § 4.1.5)?

6.3.4 Cauchy-Schwarz Inequality

According to this theorem, if x and y are random variables for which $\mu(xy)$, $\mu(x^2)$ and $\mu(y^2)$ do exist then:

$$\big[\mu(xy)\big]^2 \leq \mu(x^2)\,\mu(y^2) \tag{170}$$

Problems

1. Prove the Cauchy-Schwarz inequality.
   Answer: We have (assuming $c \in \mathbb{R}$):

$$\begin{aligned}
\mu\left[(x-cy)^2\right] &\geq 0 && \left[(x-cy)^2 \geq 0,\ \text{see notes of § 5.1}\right]\\
\mu(x^2 - 2cxy + c^2y^2) &\geq 0 &&\\
\mu(x^2) - 2c\,\mu(xy) + c^2\mu(y^2) &\geq 0 && \text{(Eqs. 119-124)}\\
\mu(x^2) - 2\left[\frac{\mu(xy)}{\mu(y^2)}\right]\mu(xy) + \left[\frac{\mu(xy)}{\mu(y^2)}\right]^2\mu(y^2) &\geq 0 && \left(c = \left[\frac{\mu(xy)}{\mu(y^2)}\right] \in \mathbb{R},\ \mu(y^2) > 0\right)\\
\mu(x^2) - 2\frac{\big[\mu(xy)\big]^2}{\mu(y^2)} + \frac{\big[\mu(xy)\big]^2}{\mu(y^2)} &\geq 0 &&\\
\mu(x^2) - \frac{\big[\mu(xy)\big]^2}{\mu(y^2)} &\geq 0 &&\\
\mu(x^2)\,\mu(y^2) - \big[\mu(xy)\big]^2 &\geq 0 && \left[\mu(y^2) > 0,\ \text{see notes of § 5.1}\right]\\
\mu(x^2)\,\mu(y^2) &\geq \big[\mu(xy)\big]^2 &&
\end{aligned}$$

   We note that $\mu(y^2) > 0$ (i.e. in lines 4 and 7) is justified (as indicated) by the fact that y^2 is positive (assuming the random variable y is non-trivial, i.e. it is not identically zero) and hence its mean must be positive (see the last point in the preamble of § 5.1).

   PE: Fully explain and justify (in words) each step of the above proof.

6.4 Iterated Expectations Theorem

This theorem (which is also known as the **law of iterated expectations**) was investigated briefly in Problem 11 of § 5.1 (and this should be enough for what we need in this book).

[150] This kind of arguments may not look straightforward, however we can reverse the above derivation.

Chapter 7
Applications

In this chapter we present a tiny sample of the applications of probability and probability theory (as well as related subjects like counting) in a number of branches and disciplines (e.g. mathematics, physics, biology, etc.). In the sections of this chapter, we will investigate these applications mainly in the form of solved Problems. However, we should note that these Problems do not necessarily represent real life issues and case studies, and hence some (and possibly most) of them are based on hypothetical situations that mimic real life issues and situations. Our purpose, after all, is the investigation of probability rather than the investigation of these specific branches (like physics or biology). Anyway, even real life case studies are usually based on (or associated with) some modeling simplifications, idealistic hypotheses and unrealistic assumptions and hence they are, to some degree, idealized and hypothesized.

7.1 Calculus

It is well known that simulated probabilistic processes can be used to integrate definite integrals numerically (or rather computationally and stochastically). This sort of stochastic integration is especially useful when integration by analytical methods is difficult or impossible. The idea of this type of numerical integration is simply based on the fact that definite integral (in one variable) represents the area under a curve (representing the integrand)[151] between the two limits of the integral. Hence, if we randomly select points in a given area (say a rectangle A) that contains the area of the definite integral (say B) then the probability P of the points being inside B is equal to the number of points inside B divided by the total number of points generated, i.e. inside A. Therefore, the area B (which is the same as the value of the definite integral) is equal to A times P, i.e. $B = A \times P$.[152] This sort of probabilistic (or stochastic) numerical integration in one variable (or 1D) space can be easily generalized and extended to multi-variable spaces (also see Problem 1).

In the following points we draw the attention to some useful remarks:

• Large number of randomly selected points are usually needed to get satisfactory results (and this may require considerable computing time although this time is usually trivial on modern computers). In fact, the number of required points for a given problem[153] depends on the required level of accuracy as well as the dimensionality (e.g. being in one variable or in two variables) of the problem in question.

• The random selection of points is done by using random number generators. The number of required random number generators is proportional to the dimensionality of the problem. These random number generators are coupled to produce random point generators (see the note of Problem 1 for more details).

• The advantages of this method of stochastic integration include: ease of implementation, flexibility in application, and possibility of being the only possible or practicable method of integration (and hence it becomes a necessity rather than a choice). The disadvantages of this method include: being an approximate method, requirement of computational equipment and resources, requirement of programming skills and resources, possibility of taking considerable computational time (although on modern computing equipment this is true only in exceptional cases and circumstances).

Problems

[151] To be more accurate we should say: the area between a curve and an axis (usually the x axis).

[152] For simplicity, we use A and B to refer to the geometric entities as well as their areas (i.e. "area" is used in two meanings).

[153] We note that the type of problem (and its size in particular) is the main factor for determining the number of required points.

1. Explain in more details the stochastic (or probabilistic) methods used to evaluate definite integrals.
   Answer: There are two main methods:
   (**a**) The method that we outlined above (for the case of 1D) where we randomly sample points inside a given rectangle and hence calculate the proportion of the area under the curve of the integrand to the total area of the rectangle (from the proportion of the number of points under the curve to the total number of points inside the rectangle). The extension of this method to higher dimensions is straightforward. This method is used in Problem 2 (for 1D) and Problem 3 (for 2D).
   (**b**) The method (for the case of 1D) of randomly sampling the integrand (by sampling the interval of integration which represents the independent variable) and hence estimating its average value where this average can be used in conjunction with the mean value theorem (of calculus) to calculate the area under the curve of the integrand which is the same as the area of the rectangle whose base is the interval of the integral and whose height is the average value of the integrand (that we estimated probabilistically). The extension of this method to higher dimensions is straightforward. This method is used in Problem 4 (for 1D) and Problem 5 (for 2D). However, we should note that this method is virtually redundant and does not seem to offer a tangible advantage because we can sample the interval (i.e. the x-interval in 1D or the xy area in 2D) systematically rather than stochastically. Nevertheless, we included this method to demonstrate the usability of stochastic processes in such applications.[154] Moreover, the coding of the stochastic process could be simpler than the coding of systematic (or deterministic) sampling.
   Note: for the method of part (a) the required number of random number generators is $n+1$ where n represents the number of variables. So, for 1-variable problems we need to couple two random number generators (to select points in an area, one of its dimensions represents the independent variable while its other dimension represents the dependent variable), for 2-variable problems we need to couple three random number generators, and so on. For the method of part (b) the required number of random number generators is equal to the number of variables (and hence n random number generators are required for n-variable problems).
   PE: Describe in sufficient details the application of the above two methods in 2D and 3D problems.
2. Write computer codes to integrate the following 1-variable definite integrals numerically using the stochastic method of part (a) of Problem 1:

   (**a**) $\int_{x_1}^{x_2} \ln x \, dx \quad (1 \le x_1 < x_2 \le 100)$. (**b**) $\int_{x_1}^{x_2} e^{x/4} \, dx \quad (-10 \le x_1 < x_2 \le 10)$.

   Answer: We used C++ programming language to do these definite integrals. The results are similar to the results of the analytical solutions. The numerical results generally improve as the number of random points increases.
   (**a**) See Integrate1DM1Log.cpp file. (**b**) See Integrate1DM1Exp.cpp file.
   PE: Do the following:
   (**a**) Comment the Integrate1DM1Log.cpp and Integrate1DM1Exp.cpp codes explaining what each line is supposed to do.
   (**b**) Calculate stochastically the value of π by calculating the area of a circle of unit radius inscribed inside a square of area 4 (noting that this is not a calculus problem but it can be solved by a similar method to the method of stochastic integration which we described already in part a of Problem 1).
3. Write computer codes to integrate the following 2-variable definite integrals numerically using the stochastic method of part (a) of Problem 1:

   (**a**) $\int_{y_1}^{y_2} \int_{x_1}^{x_2} xy^2 \, dx \, dy \quad (0 \le x_1 < x_2 \le 5 \quad \text{and} \quad 0 \le y_1 < y_2 \le 5)$.

   (**b**) $\int_{y_1}^{y_2} \int_{x_1}^{x_2} \sin(x)\cosh(y) \, dx \, dy \quad (0 \le x_1 < x_2 \le \pi/2 \quad \text{and} \quad 0 \le y_1 < y_2 \le \pi/2)$.

   Answer: We used C++ programming language to do these definite integrals. The results are similar to the results of the analytical solutions. The numerical results generally improve as the number of random points increases.

[154] In fact, the main purpose of this method is to show the possibility of achieving some deterministic processes (which is integration in this case) stochastically by using probabilistic approaches and techniques.

(**a**) See Integrate2DM1XYY.cpp file. (**b**) See Integrate2DM1sinXcoshY.cpp file.

PE: Repeat the Problem for the following integrals:[155]

(**a**) $\int_{y_1}^{y_2}\int_{x_1}^{x_2} x^3 y\,dx\,dy \qquad (0 \le x_1 < x_2 \le 5 \quad \text{and} \quad 0 \le y_1 < y_2 \le 5)$.

(**b**) $\int_{y_1}^{y_2}\int_{x_1}^{x_2} \cos(x)\sinh(y)\,dx\,dy \qquad (0 \le x_1 < x_2 \le \pi/2 \quad \text{and} \quad 0 \le y_1 < y_2 \le \pi/2)$.

4. Repeat Problem 2 using this time the stochastic method of part (b) of Problem 1.
   Answer: We used C++ programming language to do these definite integrals. The results are similar to the results of the analytical solutions. The numerical results generally improve as the number of random points increases.
   (**a**) See Integrate1DM2Log.cpp file. (**b**) See Integrate1DM2Exp.cpp file.
   PE: Comment the Integrate1DM2Log.cpp and Integrate1DM2Exp.cpp codes explaining what each line is supposed to do.
5. Repeat Problem 3 using this time the stochastic method of part (b) of Problem 1.
   Answer: We used C++ programming language to do these definite integrals. The results are similar to the results of the analytical solutions. The numerical results generally improve as the number of random points increases.
   (**a**) See Integrate2DM2XYY.cpp file. (**b**) See Integrate2DM2sinXcoshY.cpp file.
   PE: Comment the Integrate2DM2XYY.cpp and Integrate2DM2sinXcoshY.cpp codes explaining what each line is supposed to do.
6. Outline a possible use of stochastic processes in differential calculus.
   Answer: Due to the close relation between integral calculus and differential calculus, some of the stochastic methods of integration may be used to solve initial-value differential equations. The most direct and "intuitive" method of such use is outlined in the following points (noting that this method exploits the integration method of part b of Problem 1):
   • Let assume we want to solve a differential equation of the form $dy/dx = f(x)$ with the initial condition $y_0 = y(x_0)$ over the interval $[x_0, x_n]$.
   • We divide the interval $[x_0, x_n]$ to n sub-intervals.
   • We integrate stochastically over the first interval (using the method of part b of Problem 1) and add the value of this integral (say δy_1) to the initial value (i.e. y_0), and hence we obtain the solution at x_1, i.e. $y_1(x_1) = y_0 + \delta y_1$.
   • We take y_1 as the new initial value and integrate over the second interval and hence we obtain the solution at x_2, i.e. $y_2(x_2) = y_1 + \delta y_2$.
   • We continue this process until we obtain the solution at x_n, i.e. $y_n(x_n) = y_{n-1} + \delta y_n$.
   However, we note on this method the following:
   ★ The differential equations that lend themselves to solution by this method are usually of very simple form (normally linear first order). However, the method may be elaborated to tackle differential equations of more difficult and elaborate forms. Moreover, it can be useful (like other numerical methods) for solving differential equations which are difficult or impossible to solve analytically (even though they may be of simple form).
   ★ Because this method uses the method of integration of part (b) of Problem 1, it faces the same criticism as the criticism indicated in that Problem, i.e. it is virtually redundant and does not seem to offer a tangible advantage over systematic sampling of points or over other numerical methods. However, it is generally simpler in coding in comparison to comparable numerical methods (e.g. finite difference) and possibly even to systematic sampling, and this could be an advantage that can justify its use even in simple cases where other methods are available and viable.
   Anyway, we demonstrate the application of this method in the next Problem (mainly for the purpose of demonstrating the use and applicability of stochastic methods in differential calculus) despite its triviality.
   PE: Can we use the method of integration of part (a) of Problem 1 to solve this type of differential

[155] This kind of PE can be easily done by modifying the provided codes.

equations?

7. Solve the following initial-value differential equation using the method of Problem 6:

$$\frac{dy}{dx} = x^2 e^{-x/2} \qquad \text{with } y(x=0) = -1 \qquad (\text{for } 0 \leq x \leq 10)$$

Answer: See StochasticDifferential.cpp file.
PE: Comment the StochasticDifferential.cpp code explaining what each line is supposed to do.

7.2 Physics

Random events and processes are everywhere in Nature and hence the probability theory has many applications in physics (and physical sciences in general). In fact, there are certain branches of physics that are fundamentally based on probability such as statistical mechanics and quantum mechanics (noting the different nature of reliance on the probability theory in these two branches).

Problems

1. What "phase space" in statistical mechanics means?
   Answer: It means (in the terminology of probability theory) sample space.
   PE: Mention other concepts of probabilistic nature in statistical mechanics.
2. How probabilities in quantum mechanics are expressed and quantified?
   Answer:[156] If a quantum object (say an electron) is in a state represented by the (normalized) wavefunction $\psi(\mathbf{r}, t)$ then $|\psi|^2 \, d^3\mathbf{r}$ represents the probability of finding the object in the infinitesimal volume element $d^3\mathbf{r}$ around the position $\mathbf{r}$ at time t. Accordingly, if we integrate the probability density $|\psi|^2$ over a patch of space we get the probability of finding the object in that patch at time t (and hence the integral should equal unity if the patch represents the entire space).
   PE: Compare the probability in quantum mechanics with the probability in statistical mechanics.
3. Give an example of a continuous probability function (i.e. density function) that is commonly used in physics to model the distribution of certain properties of gases.
   Answer: It is the Maxwell-Boltzmann distribution which is given by:

$$f(v_x) = \frac{2m}{kT}\sqrt{\frac{m}{2\pi kT}}\, v_x^2 \, e^{-\frac{m v_x^2}{2kT}} \qquad (0 \leq v_x < \infty)$$

   where v_x is the speed of the gas molecules in the x direction, m is the mass of the gas molecules, T is the temperature of the gas and k is the Boltzmann constant.
   PE: Show that the Maxwell-Boltzmann distribution (as given by the above equation) satisfies the conditions of probability density functions, i.e. Eqs. 86 and 87.
4. Referring to Problem 34 of § 2.2, find the probabilities of occupancy in the three cases (i.e. what is the probability of any given possible occupancy in each one of the three cases).
   Answer:
   (**a**) For **Maxwell-Boltzmann** statistics, the probability of each possible occupancy is $1/k^n$ because we have k^n equally likely possibilities for occupancy (see part a of Problem 34 of § 2.2).
   (**b**) For **Fermi-Dirac** statistics, the probability of each possible occupancy is $1/C_n^k$ because we have C_n^k equally likely possibilities for occupancy (see part b of Problem 34 of § 2.2).
   (**c**) For **Bose-Einstein** statistics, the probability of each possible occupancy is $1/C_n^{n+k-1}$ because we have C_n^{n+k-1} equally likely possibilities for occupancy (see part c of Problem 34 of § 2.2).
   PE: Can we order these probabilities (i.e. by using inequalities)? If so, order them increasingly.
5. In quantum mechanics, the degeneracy in 3D simple harmonic oscillator requires calculating the number of ordered triplets of non-negative integers n_1, n_2, n_3 restricted by the condition that $n_1 + n_2 + n_3 = n$ with n being a positive integer. Find a formula for the number of such triplets and give some examples of such triplets.

[156] This answer represents just an example.

Answer: Referring to part (c) of Problem 34 of § 2.2, we can consider "ordered triplet" as 3 states and consider n as the number of indistinguishable particles and hence we can use the Bose-Einstein statistics with $k = 3$. Hence, the number of such triplets is $C_n^{n+k-1} = C_n^{n+3-1} = C_n^{n+2}$. For example:
• If $n = 1$ then we have $C_n^{n+2} = C_1^3 = 3$ triplets which are:

$(1, 0, 0)$ $(0, 1, 0)$ $(0, 0, 1)$

• If $n = 2$ then we have $C_n^{n+2} = C_2^4 = 6$ triplets which are:

$(1, 1, 0)$ $(1, 0, 1)$ $(0, 1, 1)$ $(2, 0, 0)$ $(0, 2, 0)$ $(0, 0, 2)$

• If $n = 3$ then we have $C_n^{n+2} = C_3^5 = 10$ triplets which are:

$(1, 1, 1)$ $(2, 1, 0)$ $(2, 0, 1)$ $(1, 2, 0)$ $(0, 2, 1)$

$(1, 0, 2)$ $(0, 1, 2)$ $(3, 0, 0)$ $(0, 3, 0)$ $(0, 0, 3)$

• If $n = 4$ then we have $C_n^{n+2} = C_4^6 = 15$ triplets which are:

$(2, 1, 1)$ $(1, 2, 1)$ $(1, 1, 2)$ $(2, 2, 0)$ $(2, 0, 2)$ $(0, 2, 2)$ $(3, 1, 0)$ $(3, 0, 1)$

$(1, 3, 0)$ $(0, 3, 1)$ $(1, 0, 3)$ $(0, 1, 3)$ $(4, 0, 0)$ $(0, 4, 0)$ $(0, 0, 4)$

PE: Find all the triplets for $n = 5$.

6. In a quantum physics (or particle physics) experiment a source of emission is placed at S (see Figure 29) where at any non-terminal node (or point or junction) the emitted particles reaching that node (including S where the particles are emitted) can go randomly (with equal probability) in any one of the available one-way guided tacks that branch from that node (as shown in Figure 29). What is the probability that an emitted particle can reach the terminal (or destination) points $D_1, D_2, D_3, D_4, D_5, D_6, D_7$?
Answer: If $P(s_1)$ symbolizes the probability of reaching node s_1 and $P(D_1|s_1)$ symbolizes the probability of reaching node D_1 given that the particle reached point s_1 (and similar symbols apply to the other s and D nodes) then we have (noting that all other probabilities are zero):

(**a**) $P(s_1) = 1/4$. (**b**) $P(s_2) = 1/4$. (**c**) $P(s_3) = 1/4$. (**d**) $P(s_4) = 1/4$.

(**e**) $P(D_1|s_1) = 1$. (**f**) $P(D_2|s_2) = 1/3$. (**g**) $P(D_3|s_2) = 1/3$. (**h**) $P(D_4|s_2) = 1/3$.

(**i**) $P(D_4|s_3) = 1/2$. (**j**) $P(D_5|s_3) = 1/2$. (**k**) $P(D_6|s_4) = 1/2$. (**l**) $P(D_7|s_4) = 1/2$.

Now, if we note that reaching s_1, s_2, s_3, s_4 are mutually exclusive events whose union represents the entire sample space (and hence we can use Eq. 150) then we have:[157]

$$\begin{aligned}
P(D_1) &= \sum_{i=1}^{4} P(s_i)\, P(D_1\,|s_i) = \left(\frac{1}{4} \times 1\right) + 0 + 0 + 0 = \frac{1}{4} \\
P(D_2) &= \sum_{i=1}^{4} P(s_i)\, P(D_2\,|s_i) = 0 + \left(\frac{1}{4} \times \frac{1}{3}\right) + 0 + 0 = \frac{1}{12} \\
P(D_3) &= \sum_{i=1}^{4} P(s_i)\, P(D_3\,|s_i) = 0 + \left(\frac{1}{4} \times \frac{1}{3}\right) + 0 + 0 = \frac{1}{12} \\
P(D_4) &= \sum_{i=1}^{4} P(s_i)\, P(D_4\,|s_i) = 0 + \left(\frac{1}{4} \times \frac{1}{3}\right) + \left(\frac{1}{4} \times \frac{1}{2}\right) + 0 = \frac{5}{24} \\
P(D_5) &= \sum_{i=1}^{4} P(s_i)\, P(D_5\,|s_i) = 0 + 0 + \left(\frac{1}{4} \times \frac{1}{2}\right) + 0 = \frac{1}{8} \\
P(D_6) &= \sum_{i=1}^{4} P(s_i)\, P(D_6\,|s_i) = 0 + 0 + 0 + \left(\frac{1}{4} \times \frac{1}{2}\right) = \frac{1}{8}
\end{aligned}$$

[157] As before, $P(D_j)$ is the probability of reaching the terminal node D_j where $j = 1, 2, \ldots, 7$.

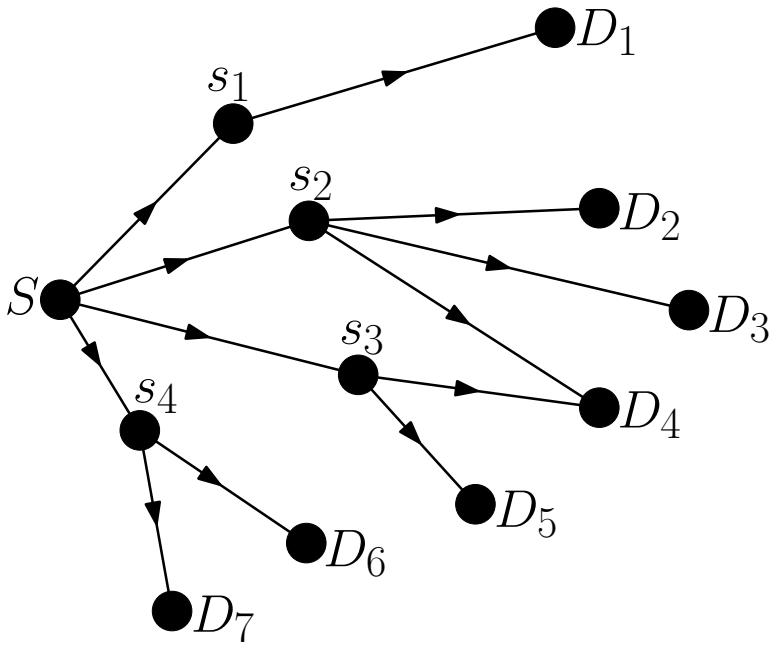


Figure 29: The schematic of Problem 6 of § 7.2. The filled circles represent nodes (or junctions) while the directed lines represent the one way tracks (noting that at any non-terminal node the probability of going through any available one-way track originating from that node is the same).

$$P(D_7) \quad = \quad \sum_{i=1}^{4} P(s_i)\, P(D_7\,|s_i) = 0 + 0 + 0 + \left(\frac{1}{4} \times \frac{1}{2}\right) = \frac{1}{8}$$

PE: Should we have $\sum_{j=1}^{7} P(D_j) = 1$? Can we consider this as a (partial) check for the validity of the obtained results?

7. In quantum physics, the wavefunction ψ of an electron in the 1s orbital in a hydrogen atom is given by $\psi = Ae^{-r/a_0}$ where A is a constant, r is the radial distance from the center of the atom and a_0 is the Bohr radius. Find the following (for this 1s electron):

   (**a**) The constant A. (**b**) Its mean distance from the center. (**c**) The variance of its distance.

   Answer:

   (**a**) The 1s orbital is spherically symmetric and hence it has only radial dependency. In quantum mechanics, the probability density function is given by $|\psi|^2 = \psi^*\psi$. By the normalization condition of probability, the integral of the probability density function over the entire space should be unity, that is (where $\mathrm{d}^3\mathbf{r}$ is an infinitesimal volume element):

$$\begin{aligned}
\int_{\text{all space}} |\psi|^2 \, \mathrm{d}^3\mathbf{r} &= 1 \\
\int_{\text{all space}} A^2 e^{-2r/a_0} \, \mathrm{d}^3\mathbf{r} &= 1 \\
\int_{r=0}^{+\infty} A^2 \, e^{-2r/a_0} \, 4\pi r^2 \, dr &= 1 \\
4\pi A^2 \int_0^{\infty} e^{-2r/a_0} \, r^2 \, dr &= 1 \\
4\pi A^2 \times \frac{a_0^3}{4} &= 1
\end{aligned}$$

$$A = \frac{1}{a_0^{3/2}\sqrt{\pi}}$$

(**b**) The mean distance is given by (see Eq. 116):[158]

$$\mu(r) = \int_0^\infty r\, f(r)\, dr = \int_0^\infty r\, \frac{4r^2}{a_0^3} e^{-2r/a_0}\, dr = \frac{4}{a_0^3}\int_0^\infty r^3\, e^{-2r/a_0}\, dr = \frac{4}{a_0^3} \times \frac{3a_0^4}{8} = \frac{3a_0}{2}$$

(**c**) The variance is given by (see Eq. 138):

$$\begin{aligned} V(r) &= \int_0^\infty (r-\mu_r)^2\, \frac{4r^2}{a_0^3} e^{-2r/a_0}\, dr = \frac{4}{a_0^3}\int_0^\infty (r^4 - 2\mu_r r^3 + \mu_r^2 r^2)\, e^{-2r/a_0}\, dr \\ &= \frac{4}{a_0^3}\int_0^\infty \left(r^4 - 3a_0 r^3 + \frac{9a_0^2}{4} r^2\right) e^{-2r/a_0}\, dr = \frac{4}{a_0^3} \times \frac{3a_0^5}{16} = \frac{3a_0^2}{4} \end{aligned}$$

PE: Do the following:
(**a**) Justify each step of the above derivations.
(**b**) Calculate the numeric value of A, $\mu(r)$ and $V(r)$.
(**c**) Find $\sigma(r)$ and calculate its numeric value.

7.3 Biology

One of the best known examples of the use of probability theory in biological sciences is in genetics. Probability theory (and related branches) is also essential in the quantitative investigation of infectious diseases, epidemics, pandemics and so on. In fact, it underlies almost all the quantitative investigations related to health and medical issues noting that these types of investigation generally rely on statistical data and they lack reliable analytical models (unlike physical sciences, to some degree, for instance) and hence they heavily rely on probability and statistics.

Problems

1. Give an example of the "use of probability" by animals.
   Answer: The principle of "safety in numbers" which many social animals use to avoid predators (or rather reduce their chance of being caught by predators) is an example of the "use of probability" by animals.
   PE: Give other examples of the "use of probability" by animals (e.g. in mating and breeding habits of some animals or living beings in general). Also give some examples of the use of probability (consciously and unconsciously) in our daily life.
2. A virus test produces a positive result in 98.23% of the cases of infection (i.e. the result is correct). It also produces a positive result in 0.09% of the cases of non-infection (i.e. the result is incorrect).[159] If this test is conducted on a person picked randomly from a population with 0.013% infection rate and it tests positive, what is the chance of this person being really infected?
   Answer: Let adopt the following:
   • A is the event that "the person is really infected".
   • B is the event that "the test is positive".
   The required probability is $P(A\,|B)$, i.e. the probability that the person is really infected given that the test is positive. From the given information we have:
   $P(A) = 0.00013$ $\qquad$ $P(\overline{A}) = 0.99987$ $\qquad$ $P(B\,|A) = 0.9823$ $\qquad$ $P(B\,|\overline{A}) = 0.0009$

[158] It should be noted that $|\psi|^2$ is the probability density function with regard to space (i.e. corresponding to $\mathrm{d}^3\mathbf{r}$) and $\frac{4r^2}{a_0^3}e^{-2r/a_0}$ is the probability density function with regard to radial distance (i.e. corresponding to dr).

[159] In brief, if we assume that the test produces only positive and negative results (i.e. there is no possibility of indeterminate results) then we can say: in the cases of infection it produces 98.23% correct positive results and 1.77% incorrect negative results, while in the cases of non-infection it produces 99.91% correct negative results and 0.09% incorrect positive results.

On substituting these values in Eq. 152 we get:

$$P(A\,|B) = \frac{P(A)\,P(B\,|A)}{P(A)\,P(B\,|A) + P(\overline{A})\,P(B\,|\overline{A})} = \frac{0.00013 \times 0.9823}{0.00013 \times 0.9823 + 0.99987 \times 0.0009} \simeq 0.124271348$$

This result may look odd because we have 98.23% success rate (i.e. when the test is positive). However, this oddity should disappear if we notice that the actual infection rate is 0.013% which is very small and this should drive the probability down to this low level. In loose terms, if the infection rate of population is very small then the probability (before test) of this person being infected is very small and hence even if the probability of correctness of test is very high the final probability [i.e. the probability (after positive test) of this person being really infected given that the infection rate is very small] should be relatively small because this final probability is the result of coupling these two probabilities where the high probability of correctness of test is moderated by the low probability of this person (i.e. prior to test and as a member of population) being infected.
Anyway, let verify (partially) this result. We have four possible cases for the result of the test:
• The test is positive and the person is really infected. The probability in this case is $P(A\,|B)$.
• The test is positive and the person is not really infected. The probability in this case is $P(\overline{A}\,|B)$.
• The test is negative and the person is really infected. The probability in this case is $P(A\,|\overline{B})$.
• The test is negative and the person is not really infected. The probability in this case is $P(\overline{A}\,|\overline{B})$.
Now, the probabilities of the first two cases should add up to unity (because if the test is positive then either the person is infected or not). Similarly, the probabilities of the last two cases should add up to unity (because if the test is negative then either the person is infected or not). If we calculate the probabilities of the last three cases as we did already for the first case (using Eq. 152) then we get [noting that $P(\overline{B}\,|A) = 0.0177$ and $P(\overline{B}\,|\overline{A}) = 0.9991$]:

$$\begin{aligned}
P(\overline{A}\,|B) &= \frac{P(\overline{A})\,P(B\,|\overline{A})}{P(\overline{A})\,P(B\,|\overline{A}) + P(A)\,P(B\,|A)} = \frac{0.99987 \times 0.0009}{0.99987 \times 0.0009 + 0.00013 \times 0.9823} \simeq 0.875728652 \\
P(A\,|\overline{B}) &= \frac{P(A)\,P(\overline{B}\,|A)}{P(A)\,P(\overline{B}\,|A) + P(\overline{A})\,P(\overline{B}\,|\overline{A})} = \frac{0.00013 \times 0.0177}{0.00013 \times 0.0177 + 0.99987 \times 0.9991} \simeq 0.000002303 \\
P(\overline{A}\,|\overline{B}) &= \frac{P(\overline{A})\,P(\overline{B}\,|\overline{A})}{P(\overline{A})\,P(\overline{B}\,|\overline{A}) + P(A)\,P(\overline{B}\,|A)} = \frac{0.99987 \times 0.9991}{0.99987 \times 0.9991 + 0.00013 \times 0.0177} \simeq 0.999997697
\end{aligned}$$

As we see, $P(A\,|B) + P(\overline{A}\,|B) = 1$ and $P(A\,|\overline{B}) + P(\overline{A}\,|\overline{B}) = 1$.
Note: the very low $P(A\,|\overline{B})$ should endorse and clarify the rationale of our argument (which we gave in the answer) further because the very low probability of being infected is pushed down further by the highly reliable negative test which implies that the chance of being infected is very low. Similarly, the very high $P(\overline{A}\,|\overline{B})$ should endorse and clarify the rationale of our argument further because the very high probability of being non-infected is pushed up further by the highly reliable negative test which implies that the chance of being non-infected is very high.
PE: Why we used the Bayes theorem in the form of Eq. 152 instead of the form of Eq. 149?

7.4 Gambling

Gambling is entirely based on the theory of probability. Historically, the emergence of this theory (in its mathematical form) was largely as a response to the demand of gamblers to make good predictions and decisions (and hence increase their chances of winning). Therefore, we can consider gambling as the birthplace of the probability theory.

Problems

1. A gambler participated in a game of lotto where the player chooses five numbers from the numbers $1, 2, \ldots, 30$ and he wins if his numbers hit the jackpot. If the price of ticket is \$1 and the prize money is \$100000, is he a winner or a looser (i.e. statistically)? Assume in your answer that only one player

can participate at any game.

Answer: The number of combinations for choosing 5 numbers out of 30 is $C_5^{30} = 142506$. This means that on a statistical (or probabilistic) basis he needs to play 142506 times (i.e. by playing all the possible combinations) to win once. In other words, he needs to spend \$142506 (by buying 142506 tickets) to win a \$100000 (i.e. the prize money). So, on a statistical basis he is a looser because he invests \$142506 to get a return of \$100000.

Warning to gamblers: it is noteworthy that for the "statistical basis" to have an effect the gambler should play many times (say tens of thousands) which is usually not achievable. So, if he plays one time (or a few times) in each lottery draw he is generally a looser even if the prize money exceeds \$142506.[160] Yes, if he plays many times (i.e. tens of thousands by buying tickets of different numbers) in each draw then he should generally be a winner if the prize money exceeds \$142506. However, no lotto offers such a high prize money (because the lotto organizer will loose). Moreover, buying too many tickets in each draw is not a practical possibility (at least for the overwhelming majority of gamblers). In fact, we also ignored another important factor that we indicated in the question, i.e. in an ordinary lottery many people participate and hence the jackpot can be won by more than one participant and hence any winner will take only part of the prize money. So, our advice is to avoid all kinds of gambling even when they look profitable.[161]

PE: Find the break-even prize money for a problem similar to the present Problem but assume this time that six numbers are selected from the numbers $1, 2, \ldots, 40$ and the price of the ticket is \$2.

2. The participant in the game of lotto in the United Kingdom chooses 6 numbers from 1-59. Winning and losing (and hence prizes) are determined by the number of winning numbers that he gets in his choice (out of 6 subsequently-drawn winning numbers). What are the probabilities of getting $0, 1, 2, 3, 4, 5, 6$ winning numbers in his choice?

   Answer: This is an example of the hypergeometric distribution (see § 4.1.6) where $N = 59$, $n = 6$, $R = 6$ and $r = 0, 1, 2, 3, 4, 5, 6$. Accordingly (see Eq. 84):

$$\begin{aligned}
P(59, 6, 6, 0) &= C_0^6 C_{6-0}^{59-6} / C_6^{59} \simeq 0.509515469 \\
P(59, 6, 6, 1) &= C_1^6 C_{6-1}^{59-6} / C_6^{59} \simeq 0.382136602 \\
P(59, 6, 6, 2) &= C_2^6 C_{6-2}^{59-6} / C_6^{59} \simeq 0.097483827 \\
P(59, 6, 6, 3) &= C_3^6 C_{6-3}^{59-6} / C_6^{59} \simeq 0.010398275 \\
P(59, 6, 6, 4) &= C_4^6 C_{6-4}^{59-6} / C_6^{59} \simeq 0.000458747 \\
P(59, 6, 6, 5) &= C_5^6 C_{6-5}^{59-6} / C_6^{59} \simeq 7.05765 \times 10^{-6} \\
P(59, 6, 6, 6) &= C_6^6 C_{6-6}^{59-6} / C_6^{59} \simeq 2.21939 \times 10^{-8}
\end{aligned}$$

   As we see, these probabilities add up to 1 as it should be (and hence this is a partial check).

   PE: Repeat the Problem by assuming $N = 55$, $n = 5$, $R = 5$ and $r = 0, 1, 2, 3, 4, 5$.

7.5 Business

Probability theory plays a central role in many types of business activities. For example, probability theory is essential in the insurance industry to make reliable predictions (i.e. of statistical nature) about damage or destruction or loss or death, for instance, and hence assess the chance of making profit or loss. This assessment is essential to the insurers for creating and tailoring their insurance policies and coverage packages and any type of product they offer to their clients. Also, entrepreneurs should consider many probabilistic factors in the business models of their projects and in the setting and arrangement of their novel enterprises and adventures.

[160] Some may disagree with this, but we think it is logical (considering the rationale of statistics).

[161] This advice is not only because of potential moral considerations but also because of certain pragmatic considerations. Gambling destroys families and brings sorrow, poverty and disasters.

7.6 Finance

The stock and foreign exchange markets are strongly affected by trends and factors of probabilistic nature and hence probability (and its theory) has a strong presence in the considerations of (and the decisions made by) banks, companies, brokers and individual investors. In fact, there are many examples about the role and significance of probability and probability theory in most types of financial activities.

7.7 Public Opinion and Trends

Probability theory is central to the disciplines and activities related to gauging public opinions and trends (e.g. in politics or marketing or advertising) to make reliable predictions and accurate projections. Accordingly, probability theory is important to political parties (or rather their strategists), pollsters and advertisers among many other professionals and professions of these types.

7.8 Meteorology

The discipline of meteorology is based on many physical factors some of which are deterministic while others are probabilistic or subject to fluctuations and uncertainties. The weather forecast, for instance, is fundamentally based on many factors and effects of stochastic nature and hence probability theory is used (partly) to make reliable predictions about the weather conditions and determine uncertainties and margins of error.

7.9 Social and Political Sciences

As social and political sciences are generally related to human behavior and activities (which are not totally deterministic), probability and its mathematical theory play an important role in these disciplines. For example, patterns of migration, trends of cultural development, political instabilities, social upheavals and wars can be investigated quantitatively (and partly) by probability theory.

7.10 Economics

Economy is based on many stochastic factors which determine for instance growth, decline, inflation, trends of markets (locally, nationally and globally), competitions, etc. Therefore, economists (especially those whose decisions have an impact on national and international levels) must take many probabilistic factors and considerations in their models, decisions, judgments, forecasts, etc.

7.11 Industry

Many industrial processes and activities are subject to fluctuations and uncertainties and hence probability is essential in the industrial modeling and assessment as well as the expectation of yield and return (e.g. whether or not a certain project is profitable considering the availability of raw materials, the demand for the final product, the proportion of defective units produced, the risks in transportation of product to customers and consumers, etc.).

References

G.B. Arfken; H.J. Weber; F.E. Harris. *Mathematical Methods for Physicists A Comprehensive Guide.* Elsevier Academic Press, seventh edition, 2013.

M.L. Boas. *Mathematical Methods in the Physical Sciences.* John Wiley & Sons Inc., third edition, 2006.

T.L. Chow. *Mathematical Methods for Physicists: A Concise Introduction.* Cambridge University Press, first edition, 2003.

W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1.* John Wiley & Sons, Inc., third edition, 1968.

W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 2.* John Wiley & Sons, Inc., second edition, 1991.

H.P. Hsu. *Schaum's Outline of Theory and Problems of Probability, Random Variables, and Random Processes.* McGraw-Hill, first edition, 1997.

E. Kreyszig; H. Kreyszig; E.J. Norminton. *Advanced Engineering Mathematics.* John Wiley & Sons, Inc., tenth edition, 2011.

S. Lipschutz. *Schaum's Outline of Probability.* McGraw-Hill, first edition, 1968.

A.D. Polyanin; A.V. Manzhirov. *Handbook of Mathematics for Engineers and Scientists.* Chapman & Hall/CRC, first edition, 2007.

K.F. Riley; M.P. Hobson; S.J. Bence. *Mathematical Methods for Physics and Engineering.* Cambridge University Press, third edition, 2006.

D. Zwillinger, editor. *CRC Standard Mathematical Tables and Formulae.* CRC Press, 32nd edition, 2012.

Note: in addition to the above references, we also consulted during our work on the preparation of this book many other books, scientific papers and general articles about probability and related subjects.

Index

Author Notes

• All copyrights of this book are held by the author.
• This book, like any other academic document, is protected by the terms and conditions of the universally recognized intellectual property rights. Hence, any quotation or use of any part of the book should be acknowledged and cited according to the scholarly approved traditions.
• This book is totally made and prepared by the author including all the graphic illustrations, indexing, typesetting, book cover, and overall design.