
THE CLASSIC WORK
NEWLY UPDATED AND REVISED

The Art of Computer Programming

VOLUME 3

Sorting and Searching
Second Edition

DONALD E. KNUTH

This page intentionally left blank

THE ART OF COMPUTER PROGRAMMING

SECOND EDITION

DONALD E. KNUTH *Stanford University*



ADDISON-WESLEY

Volume 3 / **Sorting and Searching**

THE ART OF COMPUTER PROGRAMMING

SECOND EDITION

Boston · Columbus · New York · San Francisco · Amsterdam · Cape Town
Dubai · London · Madrid · Milan · Munich · Paris · Montréal · Toronto · Delhi · Mexico City
São Paulo · Sydney · Hong Kong · Seoul · Singapore · Taipei · Tokyo

T_EX is a trademark of the American Mathematical Society

METAFONT is a trademark of Addison–Wesley

The author and publisher have taken care in the preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

For government sales inquiries, please contact governmentsales@pearsoned.com

For questions about sales outside the U.S., please contact intlcs@pearson.com

Visit us on the Web: informit.com/aw

Library of Congress Cataloging-in-Publication Data

Knuth, Donald Ervin, 1938–

The art of computer programming / Donald Ervin Knuth.
xiv, 782 p. 24 cm.

Includes bibliographical references and index.

Contents: v. 1. Fundamental algorithms. -- v. 2. Seminumerical algorithms. -- v. 3. Sorting and searching. -- v. 4a. Combinatorial algorithms, part 1.

Contents: v. 3. Sorting and searching. -- 2nd ed.

ISBN 978-0-201-89683-1 (v. 1, 3rd ed.)

ISBN 978-0-201-89684-8 (v. 2, 3rd ed.)

ISBN 978-0-201-89685-5 (v. 3, 2nd ed.)

ISBN 978-0-201-03804-0 (v. 4a)

1. Electronic digital computers--Programming. 2. Computer algorithms. I. Title.

QA76.6.K64 1997

005.1--DC21

97-2147

Internet page <http://www-cs-faculty.stanford.edu/~knuth/taocp.html> contains current information about this book and related books.

Copyright © 1998 by Addison–Wesley

All rights reserved. Printed in the United States of America. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, request forms, and the appropriate contacts with the Pearson Education Global Rights & Permissions Department, please visit www.pearson.com/permissions/.

ISBN-13 978-0-201-89685-5

ISBN-10 0-201-89685-0

Fourth digital release, December 2023

PREFACE

*Cookery is become an art,
a noble science,
Cookees are Gentlemen.*

— TITUS LIVIUS, *Ab Urbe Condita* XXXIX.vi
(Robert Burton, *Anatomy of Melancholy* 1.2.2.2, 1624)

THIS BOOK forms a natural sequel to the material on information structures in Chapter 2 of Volume 1, because it adds the concept of linearly ordered data to the other basic structural ideas.

The title “Sorting and Searching” may sound as if this book is only for those systems programmers who are concerned with the preparation of general-purpose sorting routines or applications to information retrieval. But in fact the area of sorting and searching provides an ideal framework for discussing a wide variety of important general issues:

- How are good algorithms discovered?
- How can given algorithms and programs be improved?
- How can the efficiency of algorithms be analyzed mathematically?
- How can a person choose rationally between different algorithms for the same task?
- In what senses can algorithms be proved “best possible”?
- How does the theory of computing interact with practical considerations?
- How can external memories like tapes, drums, or disks be used efficiently with large databases?

Indeed, I believe that virtually *every* important aspect of programming arises somewhere in the context of sorting or searching!

This volume comprises Chapters 5 and 6 of the complete series. Chapter 5 is concerned with sorting into order; this is a large subject that has been divided chiefly into two parts, internal sorting and external sorting. There also are supplementary sections, which develop auxiliary theories about permutations (Section 5.1) and about optimum techniques for sorting (Section 5.3). Chapter 6 deals with the problem of searching for specified items in tables or files; this is subdivided into methods that search sequentially, or by comparison of keys, or by digital properties, or by hashing, and then the more difficult problem of secondary key retrieval is considered. There is a surprising amount of interplay

between both chapters, with strong analogies tying the topics together. Two important varieties of information structures are also discussed, in addition to those considered in Chapter 2, namely priority queues (Section 5.2.3) and linear lists represented as balanced trees (Section 6.2.3).

Like Volumes 1 and 2, this book includes a lot of material that does not appear in other publications. Many people have kindly written to me about their ideas, or spoken to me about them, and I hope that I have not distorted the material too badly when I have presented it in my own words.

I have not had time to search the patent literature systematically; indeed, I decry the current tendency to seek patents on algorithms (see Section 5.4.5). If somebody sends me a copy of a relevant patent not presently cited in this book, I will dutifully refer to it in future editions. However, I want to encourage people to continue the centuries-old mathematical tradition of putting newly discovered algorithms into the public domain. There are better ways to earn a living than to prevent other people from making use of one's contributions to computer science.

Before I retired from teaching, I used this book as a text for a student's second course in data structures, at the junior-to-graduate level, omitting most of the mathematical material. I also used the mathematical portions of this book as the basis for graduate-level courses in the analysis of algorithms, emphasizing especially Sections 5.1, 5.2.2, 6.3, and 6.4. A graduate-level course on concrete computational complexity could also be based on Sections 5.3, and 5.4.4, together with Sections 4.3.3, 4.6.3, and 4.6.4 of Volume 2.

For the most part this book is self-contained, except for occasional discussions relating to the MIX computer explained in Volume 1. Appendix B contains a summary of the mathematical notations used, some of which are a little different from those found in traditional mathematics books.

Preface to the Second Edition

This new edition matches the third editions of Volumes 1 and 2, in which I have been able to celebrate the completion of \TeX and \METAFONT by applying those systems to the publications they were designed for.

The conversion to electronic format has given me the opportunity to go over every word of the text and every punctuation mark. I've tried to retain the youthful exuberance of my original sentences while perhaps adding some more mature judgment. Dozens of new exercises have been added; dozens of old exercises have been given new and improved answers. Changes appear everywhere, but most significantly in Sections 5.1.4 (about permutations and tableaux), 5.3 (about optimum sorting), 5.4.9 (about disk sorting), 6.2.2 (about entropy), 6.4 (about universal hashing), and 6.5 (about multidimensional trees and tries).



The Art of Computer Programming is, however, still a work in progress. Research on sorting and searching continues to grow at a phenomenal rate. Therefore some parts of this book are headed by an “under construction” icon, to apologize for the fact that the material is not up-to-date. For example, if I were teaching an undergraduate class on data structures today, I would surely discuss randomized structures such as treaps at some length; but at present, I am only able to cite the principal papers on the subject, and to announce plans for a future Section 6.2.5 (see page 478). My files are bursting with important material that I plan to include in the final, glorious, third edition of Volume 3, perhaps 17 years from now. But I must finish Volumes 4 and 5 first, and I do not want to delay their publication any more than absolutely necessary.

I am enormously grateful to the many hundreds of people who have helped me to gather and refine this material during the past 35 years. Most of the hard work of preparing the new edition was accomplished by Phyllis Winkler (who put the text of the first edition into \TeX form), by Silvio Levy (who edited it extensively and helped to prepare several dozen illustrations), and by Jeffrey Oldham (who converted more than 250 of the original illustrations to $\text{\textsc{METAPOST}}$ format). The production staff at Addison–Wesley has also been extremely helpful, as usual.

I have corrected every error that alert readers detected in the first edition — as well as some mistakes that, alas, nobody else noticed — and I have tried to avoid introducing new errors in the new material. However, I suppose some defects still remain, and I want to fix them as soon as possible. Therefore I will cheerfully award \$2.56 to the first finder of each technical, typographical, or historical error. The webpage cited on page iv contains a current listing of all corrections that have been reported to me.

Stanford, California
February 1998

D. E. K.

*There are certain common Privileges of a Writer,
the Benefit whereof, I hope, there will be no Reason to doubt;
Particularly, that where I am not understood, it shall be concluded,
that something very useful and profound is coucht underneath.*

— JONATHAN SWIFT, *Tale of a Tub*, Preface (1704)

NOTES ON THE EXERCISES

THE EXERCISES in this set of books have been designed for self-study as well as for classroom study. It is difficult, if not impossible, for anyone to learn a subject purely by reading about it, without applying the information to specific problems and thereby being encouraged to think about what has been read. Furthermore, we all learn best the things that we have discovered for ourselves. Therefore the exercises form a major part of this work; a definite attempt has been made to keep them as informative as possible and to select problems that are enjoyable as well as instructive.

In many books, easy exercises are found mixed randomly among extremely difficult ones. A motley mixture is, however, often unfortunate because readers like to know in advance how long a problem ought to take—otherwise they may just skip over all the problems. A classic example of such a situation is the book *Dynamic Programming* by Richard Bellman; this is an important, pioneering work in which a group of problems is collected together at the end of some chapters under the heading “Exercises and Research Problems,” with extremely trivial questions appearing in the midst of deep, unsolved problems. It is rumored that someone once asked Dr. Bellman how to tell the exercises apart from the research problems, and he replied, “If you can solve it, it is an exercise; otherwise it’s a research problem.”

Good arguments can be made for including both research problems and very easy exercises in a book of this kind; therefore, to save the reader from the possible dilemma of determining which are which, *rating numbers* have been provided to indicate the level of difficulty. These numbers have the following general significance:

Rating Interpretation

- 00 An extremely easy exercise that can be answered immediately if the material of the text has been understood; such an exercise can almost always be worked “in your head.”
- 10 A simple problem that makes you think over the material just read, but is by no means difficult. You should be able to do this in one minute at most; pencil and paper may be useful in obtaining the solution.
- 20 An average problem that tests basic understanding of the text material, but you may need about fifteen or twenty minutes to answer it completely.

- 30 A problem of moderate difficulty and/or complexity; this one may involve more than two hours' work to solve satisfactorily, or even more if the TV is on.
- 40 Quite a difficult or lengthy problem that would be suitable for a term project in classroom situations. A student should be able to solve the problem in a reasonable amount of time, but the solution is not trivial.
- 50 A research problem that has not yet been solved satisfactorily, as far as the author knew at the time of writing, although many people have tried. If you have found an answer to such a problem, you ought to write it up for publication; furthermore, the author of this book would appreciate hearing about the solution as soon as possible (provided that it is correct).

By interpolation in this “logarithmic” scale, the significance of other rating numbers becomes clear. For example, a rating of 17 would indicate an exercise that is a bit simpler than average. Problems with a rating of 50 that are subsequently solved by some reader may appear with a 40 rating in later editions of the book, and in the errata posted on the Internet (see page iv).

The remainder of the rating number divided by 5 indicates the amount of detailed work required. Thus, an exercise rated 24 may take longer to solve than an exercise that is rated 25, but the latter will require more creativity. All exercises with ratings of 46 or more are open problems for future research, rated according to the number of different attacks that they've resisted so far.

The author has tried earnestly to assign accurate rating numbers, but it is difficult for the person who makes up a problem to know just how formidable it will be for someone else to find a solution; and everyone has more aptitude for certain types of problems than for others. It is hoped that the rating numbers represent a good guess at the level of difficulty, but they should be taken as general guidelines, not as absolute indicators.

This book has been written for readers with varying degrees of mathematical training and sophistication; as a result, some of the exercises are intended only for the use of more mathematically inclined readers. The rating is preceded by an *M* if the exercise involves mathematical concepts or motivation to a greater extent than necessary for someone who is primarily interested only in programming the algorithms themselves. An exercise is marked with the letters “*HM*” if its solution necessarily involves a knowledge of calculus or other higher mathematics not developed in this book. An “*HM*” designation does *not* necessarily imply difficulty.

Some exercises are preceded by an arrowhead, “►”; this designates problems that are especially instructive and especially recommended. Of course, no reader/student is expected to work *all* of the exercises, so those that seem to be the most valuable have been singled out. (This distinction is not meant to detract from the other exercises!) Each reader should at least make an attempt to solve all of the problems whose rating is 10 or less; and the arrows may help to indicate which of the problems with a higher rating should be given priority.

Solutions to most of the exercises appear in the answer section. Please use them wisely; do not turn to the answer until you have made a genuine effort to solve the problem by yourself, or unless you absolutely do not have time to work this particular problem. *After* getting your own solution or giving the problem a decent try, you may find the answer instructive and helpful. The solution given will often be quite short, and it will sketch the details under the assumption that you have earnestly tried to solve it by your own means first. Sometimes the solution gives less information than was asked; often it gives more. It is quite possible that you may have a better answer than the one published here, or you may have found an error in the published solution; in such a case, the author will be pleased to know the details. Later printings of this book will give the improved solutions together with the solver's name where appropriate.

When working an exercise you may generally use the answers to previous exercises, unless specifically forbidden from doing so. The rating numbers have been assigned with this in mind; thus it is possible for exercise $n + 1$ to have a lower rating than exercise n , even though it includes the result of exercise n as a special case.

Summary of codes:	00	Immediate
	10	Simple (one minute)
	20	Medium (quarter hour)
▶ Recommended	30	Moderately hard
<i>M</i> Mathematically oriented	40	Term project
<i>HM</i> Requiring "higher math"	50	Research problem

EXERCISES

- ▶ 1. [00] What does the rating " $M20$ " mean?
2. [10] Of what value can the exercises in a textbook be to the reader?
3. [*HM*45] Prove that when n is an integer, $n > 2$, the equation $x^n + y^n = z^n$ has no solution in positive integers x, y, z .

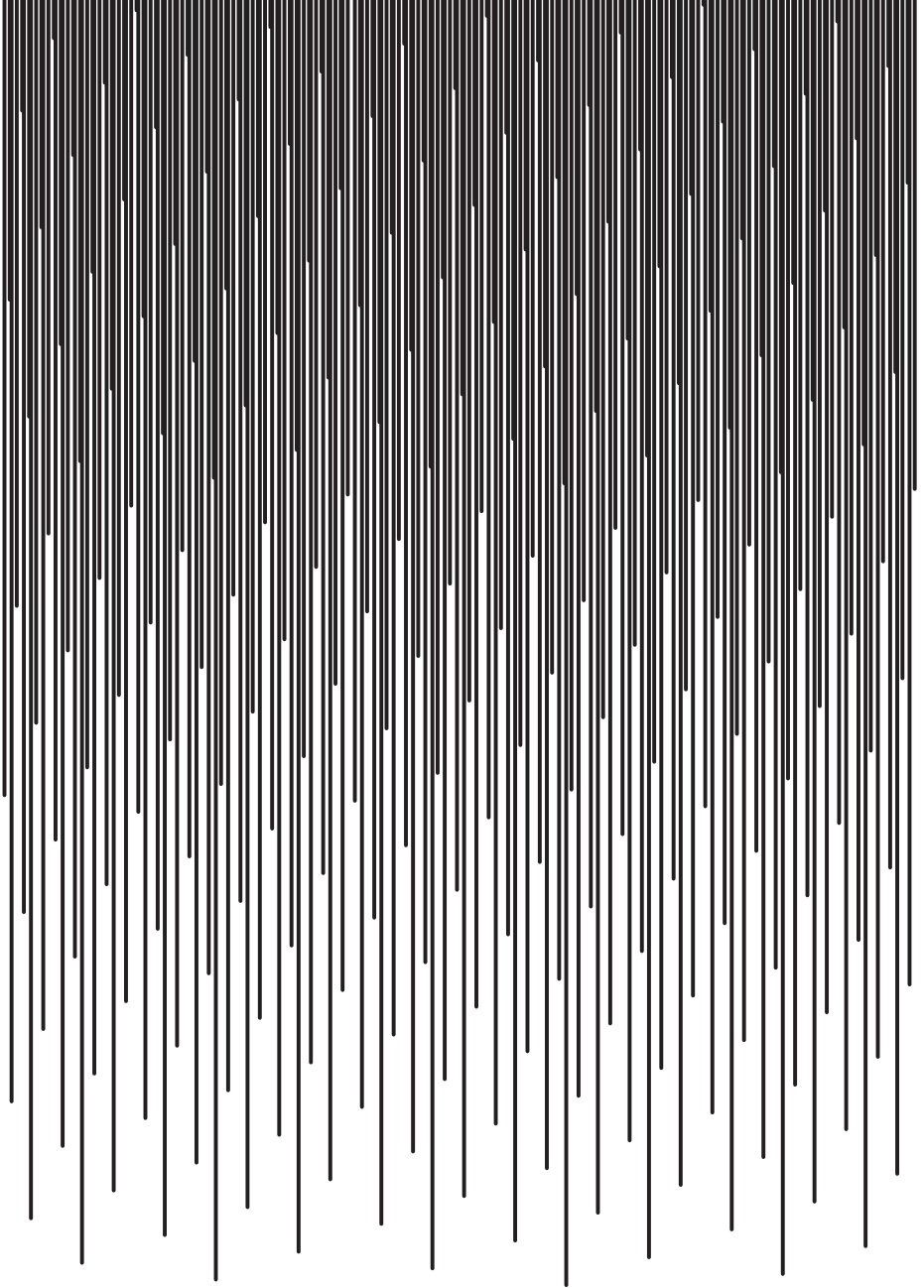
*Two hours' daily exercise . . . will be enough
to keep a hack fit for his work.*

— M. H. MAHON, *The Handy Horse Book* (1865)

CONTENTS

Preface	v
Notes on the Exercises	ix
Chapter 5—Sorting	1
*5.1. Combinatorial Properties of Permutations	11
*5.1.1. Inversions	11
*5.1.2. Permutations of a Multiset	22
*5.1.3. Runs	35
*5.1.4. Tableaux and Involutions	47
5.2. Internal Sorting	73
5.2.1. Sorting by Insertion	80
5.2.2. Sorting by Exchanging	105
5.2.3. Sorting by Selection	138
5.2.4. Sorting by Merging	158
5.2.5. Sorting by Distribution	168
5.3. Optimum Sorting	180
5.3.1. Minimum-Comparison Sorting	180
*5.3.2. Minimum-Comparison Merging	197
*5.3.3. Minimum-Comparison Selection	207
*5.3.4. Networks for Sorting	219
5.4. External Sorting	248
5.4.1. Multiway Merging and Replacement Selection	252
*5.4.2. The Polyphase Merge	267
*5.4.3. The Cascade Merge	288
*5.4.4. Reading Tape Backwards	299
*5.4.5. The Oscillating Sort	311
*5.4.6. Practical Considerations for Tape Merging	317
*5.4.7. External Radix Sorting	343
*5.4.8. Two-Tape Sorting	348
*5.4.9. Disks and Drums	356
5.5. Summary, History, and Bibliography	380
Chapter 6—Searching	392
6.1. Sequential Searching	396

6.2. Searching by Comparison of Keys	409
6.2.1. Searching an Ordered Table	409
6.2.2. Binary Tree Searching	426
6.2.3. Balanced Trees	458
6.2.4. Multiway Trees	481
6.3. Digital Searching	492
6.4. Hashing	513
6.5. Retrieval on Secondary Keys	559
Answers to Exercises	584
Appendix A — Tables of Numerical Quantities	748
1. Fundamental Constants (decimal)	748
2. Fundamental Constants (octal)	749
3. Harmonic Numbers, Bernoulli Numbers, Fibonacci Numbers	750
Appendix B — Index to Notations	752
Appendix C — Index to Algorithms and Theorems	757
Index and Glossary	759



CHAPTER FIVE

SORTING

*There is nothing more difficult to take in hand,
more perilous to conduct, or more uncertain in its success,
than to take the lead in the introduction of
a new order of things.*

— NICCOLÒ MACHIAVELLI, *The Prince* (1513)

*“But you can’t look up all those license
numbers in time,” Drake objected.
“We don’t have to, Paul. We merely arrange a list
and look for duplications.”*

— PERRY MASON, in *The Case of the Angry Mourner* (1951)

*“Treesort” Computer—With this new ‘computer-approach’
to nature study you can quickly identify over 260
different trees of U.S., Alaska, and Canada,
even palms, desert trees, and other exotics.
To sort, you simply insert the needle.*

— EDMUND SCIENTIFIC COMPANY, *Catalog* (1964)

IN THIS CHAPTER we shall study a topic that arises frequently in programming: the rearrangement of items into ascending or descending order. Imagine how hard it would be to use a dictionary if its words were not alphabetized! We will see that, in a similar way, the order in which items are stored in computer memory often has a profound influence on the speed and simplicity of algorithms that manipulate those items.

Although dictionaries of the English language define “sorting” as the process of separating or arranging things according to class or kind, computer programmers traditionally use the word in the much more special sense of marshaling things into ascending or descending order. The process should perhaps be called *ordering*, not sorting; but anyone who tries to call it “ordering” is soon led into confusion because of the many different meanings attached to that word. Consider the following sentence, for example: “Since only two of our tape drives were in working order, I was ordered to order more tape units in short order, in order to order the data several orders of magnitude faster.” Mathematical terminology abounds with still more senses of order (the order of a group, the order of a permutation, the order of a branch point, relations of order, etc., etc.). Thus we find that the word “order” can lead to chaos.

Some people have suggested that “sequencing” would be the best name for the process of sorting into order; but this word often seems to lack the right

connotation, especially when equal elements are present, and it occasionally conflicts with other terminology. It is quite true that “sorting” is itself an overused word (“I was sort of out of sorts after sorting that sort of data”), but it has become firmly established in computing parlance. Therefore we shall use the word “sorting” chiefly in the strict sense of sorting into order, without further apologies.

Some of the most important applications of sorting are:

a) *Solving the “togetherness” problem*, in which all items with the same identification are brought together. Suppose that we have 10000 items in arbitrary order, many of which have equal values; and suppose that we want to rearrange the data so that all items with equal values appear in consecutive positions. This is essentially the problem of sorting in the older sense of the word; and it can be solved easily by sorting the file in the new sense of the word, so that the values are in ascending order, $v_1 \leq v_2 \leq \dots \leq v_{10000}$. The efficiency achievable in this procedure explains why the original meaning of “sorting” has changed.

b) *Matching items in two or more files*. If several files have been sorted into the same order, it is possible to find all of the matching entries in one sequential pass through them, without backing up. This is the principle that Perry Mason used to help solve a murder case (see the quotation at the beginning of this chapter). We can usually process a list of information most quickly by traversing it in sequence from beginning to end, instead of skipping around at random in the list, unless the entire list is small enough to fit in a high-speed random-access memory. Sorting makes it possible to use sequential accessing on large files, as a feasible substitute for direct addressing.

c) *Searching for information by key values*. Sorting is also an aid to searching, as we shall see in Chapter 6, hence it helps us make computer output more suitable for human consumption. In fact, a listing that has been sorted into alphabetic order often looks quite authoritative even when the associated numerical information has been incorrectly computed.

Although sorting has traditionally been used mostly for business data processing, it is actually a basic tool that every programmer should keep in mind for use in a wide variety of situations. We have discussed its use for simplifying algebraic formulas, in exercise 2.3.2–17. The exercises below illustrate the diversity of typical applications.

One of the first large-scale software systems to demonstrate the versatility of sorting was the LARC Scientific Compiler developed by J. Erdwinn, D. E. Ferguson, and their associates at Computer Sciences Corporation in 1960. This optimizing compiler for an extended FORTRAN language made heavy use of sorting so that the various compilation algorithms were presented with relevant parts of the source program in a convenient sequence. The first pass was a lexical scan that divided the FORTRAN source code into individual tokens, each representing an identifier or a constant or an operator, etc. Each token was assigned several sequence numbers; when sorted on the name and an appropriate sequence number, all the uses of a given identifier were brought together. The

“defining entries” by which a user would specify whether an identifier stood for a function name, a parameter, or a dimensioned variable were given low sequence numbers, so that they would appear first among the tokens having a given identifier; this made it easy to check for conflicting usage and to allocate storage with respect to EQUIVALENCE declarations. The information thus gathered about each identifier was now attached to each token; in this way no “symbol table” of identifiers needed to be maintained in the high-speed memory. The updated tokens were then sorted on another sequence number, which essentially brought the source program back into its original order except that the numbering scheme was cleverly designed to put arithmetic expressions into a more convenient “Polish prefix” form. Sorting was also used in later phases of compilation, to facilitate loop optimization, to merge error messages into the listing, etc. In short, the compiler was designed so that virtually all the processing could be done sequentially from files that were stored in an auxiliary drum memory, since appropriate sequence numbers were attached to the data in such a way that it could be sorted into various convenient arrangements.

Computer manufacturers of the 1960s estimated that more than 25 percent of the running time on their computers was spent on sorting, when all their customers were taken into account. In fact, there were many installations in which the task of sorting was responsible for more than half of the computing time. From these statistics we may conclude that either (i) there are many important applications of sorting, or (ii) many people sort when they shouldn’t, or (iii) inefficient sorting algorithms have been in common use. The real truth probably involves all three of these possibilities, but in any event we can see that sorting is worthy of serious study, as a practical matter.

Even if sorting were almost useless, there would be plenty of rewarding reasons for studying it anyway! The ingenious algorithms that have been discovered show that sorting is an extremely interesting topic to explore in its own right. Many fascinating unsolved problems remain in this area, as well as quite a few solved ones.

From a broader perspective we will find also that sorting algorithms make a valuable *case study* of how to attack computer programming problems in general. Many important principles of data structure manipulation will be illustrated in this chapter. We will be examining the evolution of various sorting techniques in an attempt to indicate how the ideas were discovered in the first place. By extrapolating this case study we can learn a good deal about strategies that help us design good algorithms for other computer problems.

Sorting techniques also provide excellent illustrations of the general ideas involved in the *analysis of algorithms* — the ideas used to determine performance characteristics of algorithms so that an intelligent choice can be made between competing methods. Readers who are mathematically inclined will find quite a few instructive techniques in this chapter for estimating the speed of computer algorithms and for solving complicated recurrence relations. On the other hand, the material has been arranged so that readers without a mathematical bent can safely skip over these calculations.

Before going on, we ought to define our problem a little more clearly, and introduce some terminology. We are given N items

$$R_1, R_2, \dots, R_N$$

to be sorted; we shall call them *records*, and the entire collection of N records will be called a *file*. Each record R_j has a *key*, K_j , which governs the sorting process. Additional data, besides the key, is usually also present; this extra “satellite information” has no effect on sorting except that it must be carried along as part of each record.

An ordering relation “ $<$ ” is specified on the keys so that the following conditions are satisfied for any key values a, b, c :

- i) Exactly one of the possibilities $a < b$, $a = b$, $b < a$ is true. (This is called the law of trichotomy.)
- ii) If $a < b$ and $b < c$, then $a < c$. (This is the familiar law of transitivity.)

Properties (i) and (ii) characterize the mathematical concept of *linear ordering*, also called *total ordering*. Any relationship “ $<$ ” satisfying these two properties can be sorted by most of the methods to be mentioned in this chapter, although some sorting techniques are designed to work only with numerical or alphabetic keys that have the usual ordering.

The goal of sorting is to determine a permutation $p(1)p(2)\dots p(N)$ of the indices $\{1, 2, \dots, N\}$ that will put the keys into nondecreasing order:

$$K_{p(1)} \leq K_{p(2)} \leq \dots \leq K_{p(N)}. \quad (1)$$

The sorting is called *stable* if we make the further requirement that records with equal keys should retain their original relative order. In other words, stable sorting has the additional property that

$$p(i) < p(j) \quad \text{whenever} \quad K_{p(i)} = K_{p(j)} \quad \text{and} \quad i < j. \quad (2)$$

In some cases we will want the records to be physically rearranged in storage so that their keys are in order. But in other cases it will be sufficient merely to have an auxiliary table that specifies the permutation in some way, so that the records can be accessed in order of their keys.

A few of the sorting methods in this chapter assume the existence of either or both of the values “ ∞ ” and “ $-\infty$ ”, which are defined to be greater than or less than all keys, respectively:

$$-\infty < K_j < \infty, \quad \text{for } 1 \leq j \leq N. \quad (3)$$

Such extreme values are occasionally used as artificial keys or as sentinel indicators. The case of equality is excluded in (3); if equality can occur, the algorithms can be modified so that they will still work, but usually at the expense of some elegance and efficiency.

Sorting can be classified generally into *internal sorting*, in which the records are kept entirely in the computer’s high-speed random-access memory, and *external sorting*, when more records are present than can be held comfortably in

memory at once. Internal sorting allows more flexibility in the structuring and accessing of the data, while external sorting shows us how to live with rather stringent accessing constraints.

The time required to sort N records, using a decent general-purpose sorting algorithm, is roughly proportional to $N \log N$; we make about $\log N$ “passes” over the data. This is the minimum possible time, as we shall see in Section 5.3.1, if the records are in random order and if sorting is done by pairwise comparisons of keys. Thus if we double the number of records, it will take a little more than twice as long to sort them, all other things being equal. (Actually, as N approaches infinity, a better indication of the time needed to sort is $N(\log N)^2$, if the keys are distinct, since the size of the keys must grow at least as fast as $\log N$; but for practical purposes, N never really approaches infinity.)

On the other hand, if the keys are known to be randomly distributed with respect to some continuous numerical distribution, we will see that sorting can be accomplished in $O(N)$ steps on the average.

EXERCISES — First Set

1. [M20] Prove, from the laws of trichotomy and transitivity, that the permutation $p(1)p(2)\dots p(N)$ is *uniquely* determined when the sorting is assumed to be stable.

2. [21] Assume that each record R_j in a certain file contains *two* keys, a “major key” K_j and a “minor key” k_j , with a linear ordering $<$ defined on each of the sets of keys. Then we can define *lexicographic order* between pairs of keys (K, k) in the usual way:

$$(K_i, k_i) < (K_j, k_j) \quad \text{if } K_i < K_j \quad \text{or if } K_i = K_j \quad \text{and } k_i < k_j.$$

Alice took this file and sorted it first on the major keys, obtaining n groups of records with equal major keys in each group,

$$K_{p(1)} = \dots = K_{p(i_1)} < K_{p(i_1+1)} = \dots = K_{p(i_2)} < \dots < K_{p(i_{n-1}+1)} = \dots = K_{p(i_n)},$$

where $i_n = N$. Then she sorted each of the n groups $R_{p(i_{j-1}+1)}, \dots, R_{p(i_j)}$ on their minor keys.

Bill took the same original file and sorted it first on the minor keys; then he took the resulting file, and sorted it on the major keys.

Chris took the same original file and did a single sorting operation on it, using lexicographic order on the major and minor keys (K_j, k_j) .

Did everyone obtain the same result?

3. [M25] Let $<$ be a relation on K_1, \dots, K_N that satisfies the law of trichotomy but *not* the transitive law. Prove that even without the transitive law it is possible to sort the records in a stable manner, meeting conditions (1) and (2); in fact, there are at least three arrangements that satisfy the conditions!
- 4. [21] Lexicographers don’t actually use strict lexicographic order in dictionaries, because uppercase and lowercase letters must be interfiled. Thus they want an ordering such as this:

$$a < A < aa < AA < AAA < Aachen < aah < \dots < zzz < ZZZ.$$

Explain how to implement dictionary order.

- 5. [M28] Design a binary code for all nonnegative integers so that if n is encoded as the string $\rho(n)$ we have $m < n$ if and only if $\rho(m)$ is lexicographically less than $\rho(n)$. Moreover, $\rho(m)$ should not be a prefix of $\rho(n)$ for any $m \neq n$. If possible, the length of $\rho(n)$ should be at most $\lg n + O(\log \log n)$ for all large n . (Such a code is useful if we want to sort texts that mix words and numbers, or if we want to map arbitrarily large alphabets into binary strings.)

6. [15] Mr. B. C. Dull (a MIX programmer) wanted to know if the number stored in location A is greater than, less than, or equal to the number stored in location B. So he wrote 'LDA A; SUB B' and tested whether register A was positive, negative, or zero. What serious mistake did he make, and what should he have done instead?

7. [17] Write a MIX subroutine for multiprecision comparison of keys, having the following specifications:

Calling sequence: JMP COMPARE

Entry conditions: rI1 = n ; CONTENTS(A + k) = a_k and CONTENTS(B + k) = b_k , for $1 \leq k \leq n$; assume that $n \geq 1$.

Exit conditions: CI = GREATER, if $(a_n, \dots, a_1) > (b_n, \dots, b_1)$;
 CI = EQUAL, if $(a_n, \dots, a_1) = (b_n, \dots, b_1)$;
 CI = LESS, if $(a_n, \dots, a_1) < (b_n, \dots, b_1)$;
 rX and rI1 are possibly affected.

Here the relation $(a_n, \dots, a_1) < (b_n, \dots, b_1)$ denotes lexicographic ordering from left to right; that is, there is an index j such that $a_k = b_k$ for $n \geq k > j$, but $a_j < b_j$.

- 8. [30] Locations A and B contain two numbers a and b , respectively. Show that it is possible to write a MIX program that computes and stores $\min(a, b)$ in location C, *without using any jump operators*. (Caution: Since you will not be able to test whether or not arithmetic overflow has occurred, it is wise to guarantee that overflow is impossible regardless of the values of a and b .)

9. [M27] After N independent, uniformly distributed random variables between 0 and 1 have been sorted into nondecreasing order, what is the probability that the r th smallest of these numbers is $\leq x$?

EXERCISES — Second Set

Each of the following exercises states a problem that a computer programmer might have had to solve in the old days when computers didn't have much random-access memory. Suggest a "good" way to solve the problem, *assuming that only a few thousand words of internal memory are available*, supplemented by about half a dozen tape units (enough tape units for sorting). Algorithms that work well under such limitations also prove to be efficient on modern machines.

10. [15] You are given a tape containing one million words of data. How do you determine how many distinct words are present on the tape?

11. [18] You are the U. S. Internal Revenue Service; you receive millions of "information" forms from organizations telling how much income they have paid to people, and millions of "tax" forms from people telling how much income they have been paid. How do you catch people who don't report all of their income?

12. [M25] (*Transposing a matrix.*) You are given a magnetic tape containing one million words, representing the elements of a 1000×1000 matrix stored in order by rows: $a_{1,1} a_{1,2} \dots a_{1,1000} a_{2,1} \dots a_{2,1000} \dots a_{1000,1} \dots a_{1000,1000}$. How do you create a tape in which the

elements are stored by columns $a_{1,1} a_{2,1} \dots a_{1000,1} a_{1,2} \dots a_{1000,2} \dots a_{1000,1000}$ instead? (Try to make less than a dozen passes over the data.)

13. [M26] How could you “shuffle” a large file of N words into a random rearrangement?
14. [20] You are working with two computer systems that have different conventions for the “collating sequence” that defines the ordering of alphameric characters. How do you make one computer sort alphameric files in the order used by the other computer?
15. [18] You are given a list of the names of a fairly large number of people born in the U.S.A., together with the name of the state where they were born. How do you count the number of people born in each state? (Assume that nobody appears in the list more than once.)
16. [20] In order to make it easier to make changes to large FORTRAN programs, you want to design a “cross-reference” routine; such a routine takes FORTRAN programs as input and prints them together with an index that shows each use of each identifier (that is, each name) in the program. How should such a routine be designed?
- 17. [33] (*Library card sorting.*) Before the days of computerized databases, every library maintained a catalog of cards so that users could find the books they wanted. But the task of putting catalog cards into an order convenient for human use turned out to be quite complicated as library collections grew. The following “alphabetical” listing indicates many of the procedures recommended in the *American Library Association Rules for Filing Catalog Cards* (Chicago: 1942):

<i>Text of card</i>	<i>Remarks</i>
R. Accademia nazionale dei Lincei, Rome 1812; ein historischer Roman.	Ignore foreign royalty (except British) Achtzehnhundertzwölf
Bibliothèque d'histoire révolutionnaire.	Treat apostrophe as space in French
Bibliothèque des curiosités.	Ignore accents on letters
Brown, Mrs. J. Crosby	Ignore designation of rank
Brown, John	Names with dates follow those without
Brown, John, mathematician	... and the latter are subarranged
Brown, John, of Boston	by descriptive words
Brown, John, 1715–1766	Arrange identical names by birthdate
BROWN, JOHN, 1715–1766	Works “about” follow works “by”
Brown, John, d. 1811	Sometimes birthdate must be estimated
Brown, Dr. John, 1810–1882	Ignore designation of rank
Brown-Williams, Reginald Makepeace	Treat hyphen as space
Brown America.	Book titles follow compound names
Brown & Dallison's Nevada directory.	& in English becomes “and”
Brownjohn, Alan	
Den', Vladimir Éduardovich, 1867–	Ignore apostrophe in names
The den.	Ignore an initial article
Den lieben langen Tag.	... provided it's in nominative case
Dix, Morgan, 1827–1908	Names precede words
1812 ouverture.	Dix-huit cent douze
Le XIXe siècle français.	Dix-neuvième
The 1847 issue of U. S. stamps.	Eighteen forty-seven
1812 ouverture.	Eighteen twelve
I am a mathematician.	(a book by Norbert Wiener)

<i>Text of card</i>	<i>Remarks</i>
IBM journal of research and development.	Initials are like one-letter words
ha-I ha-ehad.	Ignore initial article
Ia; a love story.	Ignore punctuation in titles
International Business Machines Corporation	
al-Khwarizmi, Muḥammad ibn Mūsā,	
fl. 813–846	Ignore initial “al-” in Arabic names
Labour. A magazine for all workers.	Respell it “Labor”
Labor research association	
Labour, <i>see</i> Labor	Cross-reference card
McCall’s cook book	Ignore apostrophe in English
McCarthy, John, 1927–	Mc = Mac
Machine-independent computer	
programming.	Treat hyphen as space
MacMahon, Maj. Percy Alexander,	
1854–1929	Ignore designation of rank
Mrs. Dalloway.	“Mrs.” = “Mistress”
Mistress of mistresses.	
Royal society of London	Don’t ignore British royalty
St. Petersburger Zeitung.	“St.” = “Saint”, even in German
Saint-Saëns, Camille, 1835–1921	Treat hyphen as space
Ste-Marie, Gaston P	Sainte
Seminumerical algorithms.	(a book by Donald Ervin Knuth)
Uncle Tom’s cabin.	(a book by Harriet Beecher Stowe)
U. S. bureau of the census.	“U. S.” = “United States”
Vandermonde, Alexandre Théophile,	
1735–1796	
Van Valkenburg, Mac Elwyn, 1921–	Ignore space after prefix in surnames
Von Neumann, John, 1903–1957	
The whole art of legerdemain.	Ignore initial article
Who’s afraid of Virginia Woolf?	Ignore apostrophe in English
Wijngaarden, Adriaan van, 1916–	Surname begins with uppercase letter

(Most of these rules are subject to certain exceptions, and there are many other rules not illustrated here.)

If you were given the job of sorting large quantities of catalog cards by computer, and eventually maintaining a very large file of such cards, and if you had no chance to change these long-standing policies of card filing, how would you arrange the data in such a way that the sorting and merging operations are facilitated?

18. [M25] (E. T. Parker.) Leonhard Euler once conjectured [Nova Acta Acad. Sci. Petropolitanae **13** (1795), 45–63, §3; written in 1778] that there are no solutions to the equation

$$u^6 + v^6 + w^6 + x^6 + y^6 = z^6$$

in positive integers u, v, w, x, y, z . At the same time he conjectured that

$$x_1^n + \cdots + x_{n-1}^n = x_n^n$$

would have no positive integer solutions, for all $n \geq 3$, but this more general conjecture was disproved by the computer-discovered identity $27^5 + 84^5 + 110^5 + 133^5 = 144^5$; see L. J. Lander, T. R. Parkin, and J. L. Selfridge, *Math. Comp.* **21** (1967), 446–459.

Infinitely many counterexamples when $n = 4$ were subsequently found by Noam Elkies [*Math. Comp.* **51** (1988), 825–835]. Can you think of a way in which sorting would help in the search for counterexamples to Euler’s conjecture when $n = 6$?

- ▶ **19.** [24] Given a file containing a million or so distinct 30-bit binary words x_1, \dots, x_N , what is a good way to find all *complementary* pairs $\{x_i, x_j\}$ that are present? (Two words are complementary when one has 0 wherever the other has 1, and conversely; thus they are complementary if and only if their sum is $(11\dots 1)_2$, when they are treated as binary numbers.)
- ▶ **20.** [25] Given a file containing 1000 30-bit words x_1, \dots, x_{1000} , how would you prepare a list of all pairs (x_i, x_j) such that $x_i = x_j$ except in at most two bit positions?
- 21.** [22] How would you go about looking for five-letter anagrams such as CARET, CARTE, CATER, CRATE, REACT, RECTA, TRACE; CRUEL, LUCRE, ULCER; DOWRY, ROWDY, WORDY? [One might wish to know whether there are any sets of ten or more five-letter English anagrams besides the remarkable set

APERS, ASPER, PARES, PARSE, PEARS, PRAISE, PRESA, RAPES, REAPS, SPAER, SPARE, SPEAR,

to which we might add the French word APRÈS.]

- 22.** [M28] Given the specifications of a fairly large number of directed graphs, what approach will be useful for grouping the *isomorphic* ones together? (Directed graphs are isomorphic if there is a one-to-one correspondence between their vertices and a one-to-one correspondence between their arcs, where the correspondences preserve incidence between vertices and arcs.)
- 23.** [30] In a certain group of 4096 people, everyone has about 100 acquaintances. A file has been prepared listing all pairs of people who are acquaintances. (The relation is symmetric: If x is acquainted with y , then y is acquainted with x . Therefore the file contains roughly 200,000 entries.) How would you design an algorithm to list all the k -person *cliques* in this group of people, given k ? (A clique is an instance of mutual acquaintances: Everyone in the clique is acquainted with everyone else.) Assume that there are no cliques of size 25, so the total number of cliques cannot be enormous.
- ▶ **24.** [30] Three million men with distinct names were laid end-to-end, reaching from New York to California. Each participant was given a slip of paper on which he wrote down his own name and the name of the person immediately west of him in the line. The man at the extreme western end didn’t understand what to do, so he threw his paper away; the remaining 2,999,999 slips of paper were put into a huge basket and taken to the National Archives in Washington, D.C. Here the contents of the basket were shuffled completely and transferred to magnetic tapes.

At this point an information scientist observed that there was enough information on the tapes to reconstruct the list of people in their original order. And a computer scientist discovered a way to do the reconstruction with fewer than 1000 passes through the data tapes, using only sequential accessing of tape files and a small amount of random-access memory. How was that possible?

[In other words, given the pairs (x_i, x_{i+1}) , for $1 \leq i < N$, in random order, where the x_i are distinct, how can the sequence $x_1 x_2 \dots x_N$ be obtained, restricting all operations to serial techniques suitable for use with magnetic tapes? This is the problem of sorting into order when there is no easy way to tell which of two given keys precedes the other; we have already raised this question as part of exercise 2.2.3–25.]

25. [M21] (*Discrete logarithms.*) You know that p is a (rather large) prime number, and that a is a primitive root modulo p . Therefore, for all b in the range $1 \leq b < p$, there is a unique n such that $a^n \bmod p = b$, $1 \leq n < p$. (This n is called the index of b modulo p , with respect to a .) Explain how to find n , given b , without needing $\Omega(n)$ steps. [*Hint:* Let $m = \lceil \sqrt{p} \rceil$ and try to solve $a^{mn_1} \equiv ba^{-n_2}$ (modulo p) for $0 \leq n_1, n_2 < m$.]

*5.1. COMBINATORIAL PROPERTIES OF PERMUTATIONS

A PERMUTATION of a finite set is an arrangement of its elements into a row. Permutations are of special importance in the study of sorting algorithms, since they represent the unsorted input data. In order to study the efficiency of different sorting methods, we will want to be able to count the number of permutations that cause a certain step of a sorting procedure to be executed a certain number of times.

We have, of course, met permutations frequently in previous chapters. For example, in Section 1.2.5 we discussed two basic theoretical methods of constructing the $n!$ permutations of n objects; in Section 1.3.3 we analyzed some algorithms dealing with the cycle structure and multiplicative properties of permutations; in Section 3.3.2 we studied their “runs up” and “runs down.” The purpose of the present section is to study several other properties of permutations, and to consider the general case where equal elements are allowed to appear. In the course of this study we will learn a good deal about combinatorial mathematics.

The properties of permutations are sufficiently pleasing to be interesting in their own right, and it is convenient to develop them systematically in one place instead of scattering the material throughout this chapter. But readers who are not mathematically inclined and readers who are anxious to dive right into sorting techniques are advised to go on to Section 5.2 immediately, since the present section actually has little *direct* connection to sorting.

*5.1.1. Inversions

Let $a_1 a_2 \dots a_n$ be a permutation of the set $\{1, 2, \dots, n\}$. If $i < j$ and $a_i > a_j$, the pair (a_i, a_j) is called an *inversion* of the permutation; for example, the permutation 3 1 4 2 has three inversions: (3, 1), (3, 2), and (4, 2). Each inversion is a pair of elements that is out of sort, so the only permutation with no inversions is the sorted permutation 1 2 \dots n . This connection with sorting is the chief reason why we will be so interested in inversions, although we have already used the concept to analyze a dynamic storage allocation algorithm (see exercise 2.2.2–9).

The concept of inversions was introduced by G. Cramer in 1750 [*Intr. à l'Analyse des Lignes Courbes Algébriques* (Geneva: 1750), 657–659; see Thomas Muir, *Theory of Determinants* 1 (1906), 11–14], in connection with his famous rule for solving linear equations. In essence, Cramer defined the determinant of an $n \times n$ matrix in the following way:

$$\det \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nn} \end{pmatrix} = \sum (-1)^{\text{inv}(a_1 a_2 \dots a_n)} x_{1a_1} x_{2a_2} \dots x_{na_n},$$

summed over all permutations $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$, where $\text{inv}(a_1 a_2 \dots a_n)$ is the number of inversions of the permutation.

The *inversion table* $b_1 b_2 \dots b_n$ of the permutation $a_1 a_2 \dots a_n$ is obtained by letting b_j be the number of elements to the left of j that are greater than j .

In other words, b_j is the number of inversions whose second component is j . It follows, for example, that the permutation

$$5\ 9\ 1\ 8\ 2\ 6\ 4\ 7\ 3 \tag{1}$$

has the inversion table

$$2\ 3\ 6\ 4\ 0\ 2\ 2\ 1\ 0, \tag{2}$$

since 5 and 9 are to the left of 1; 5, 9, 8 are to the left of 2; etc. This permutation has 20 inversions in all. By definition the numbers b_j will always satisfy

$$0 \leq b_1 \leq n-1, \quad 0 \leq b_2 \leq n-2, \quad \dots, \quad 0 \leq b_{n-1} \leq 1, \quad b_n = 0. \tag{3}$$

Perhaps the most important fact about inversions is the simple observation that *an inversion table uniquely determines the corresponding permutation*. We can go back from any inversion table $b_1 b_2 \dots b_n$ satisfying (3) to the unique permutation that produces it, by successively determining the relative placement of the elements $n, n-1, \dots, 1$ (in this order). For example, we can construct the permutation corresponding to (2) as follows: Write down the number 9; then place 8 after 9, since $b_8 = 1$. Similarly, put 7 after both 8 and 9, since $b_7 = 2$. Then 6 must follow two of the numbers already written down, because $b_6 = 2$; the partial result so far is therefore

$$9\ 8\ 6\ 7.$$

Continue by placing 5 at the left, since $b_5 = 0$; put 4 after four of the numbers; and put 3 after six numbers (namely at the extreme right), giving

$$5\ 9\ 8\ 6\ 4\ 7\ 3.$$

The insertion of 2 and 1 in an analogous way yields (1).

This correspondence is important because we can often translate a problem stated in terms of permutations into an equivalent problem stated in terms of inversion tables, and the latter problem may be easier to solve. For example, consider the simplest question of all: How many permutations of $\{1, 2, \dots, n\}$ are possible? The answer must be the number of possible inversion tables, and they are easily enumerated since there are n choices for b_1 , independently $n-1$ choices for b_2 , \dots , 1 choice for b_n , making $n(n-1) \dots 1 = n!$ choices in all. Inversions are easy to count, because the b 's are completely independent of each other, while the a 's must be mutually distinct.

In Section 1.2.10 we analyzed the number of local maxima that occur when a permutation is read from right to left; in other words, we counted how many elements are larger than any of their successors. (The right-to-left maxima in (1), for example, are 3, 7, 8, and 9.) This is the number of j such that b_j has its maximum value, $n-j$. Since b_1 will equal $n-1$ with probability $1/n$, and (independently) b_2 will be equal to $n-2$ with probability $1/(n-1)$, etc., it is clear by consideration of the inversions that the average number of right-to-left

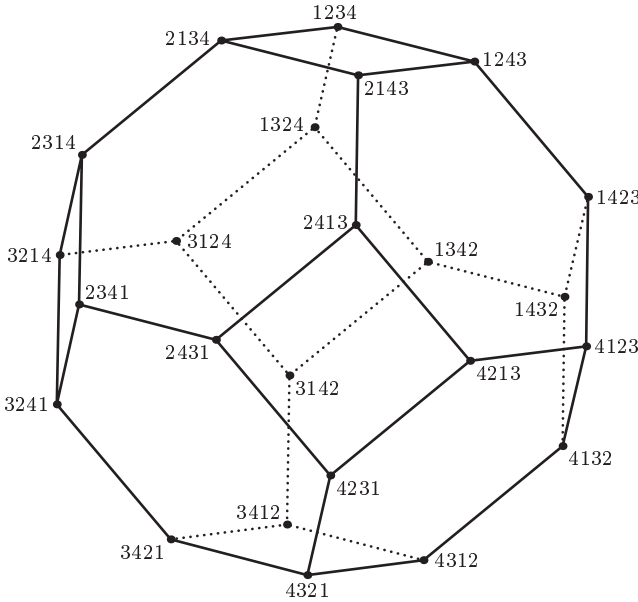


Fig. 1. The truncated octahedron, which shows the change in inversions when adjacent elements of a permutation are interchanged.

maxima is

$$\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1} = H_n.$$

The corresponding generating function is also easily derived in a similar way.

If we interchange two *adjacent* elements of a permutation, it is easy to see that the total number of inversions will increase or decrease by unity. Figure 1 shows the 24 permutations of $\{1, 2, 3, 4\}$, with lines joining permutations that differ by an interchange of adjacent elements; following any line downward inverts exactly one new pair. Hence the number of inversions of a permutation π is the length of a downward path from 1234 to π in Fig. 1; all such paths must have the same length.

Incidentally, the diagram in Fig. 1 may be viewed as a three-dimensional solid, the “truncated octahedron,” which has 8 hexagonal faces and 6 square faces. This is one of the classical uniform polyhedra attributed to Archimedes (see exercise 10).

The reader should not confuse inversions of a permutation with the *inverse* of a permutation. Recall that we can write a permutation in two-line form

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ a_1 & a_2 & a_3 & \dots & a_n \end{pmatrix}; \tag{4}$$

the inverse $a'_1 a'_2 a'_3 \dots a'_n$ of this permutation is the permutation obtained by interchanging the two rows and then sorting the columns into increasing order

of the new top row:

$$\begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_n \\ 1 & 2 & 3 & \dots & n \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ a'_1 & a'_2 & a'_3 & \dots & a'_n \end{pmatrix}. \quad (5)$$

For example, the inverse of 5 9 1 8 2 6 4 7 3 is 3 5 9 7 1 6 8 4 2, since

$$\begin{pmatrix} 5 & 9 & 1 & 8 & 2 & 6 & 4 & 7 & 3 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 3 & 5 & 9 & 7 & 1 & 6 & 8 & 4 & 2 \end{pmatrix}.$$

Another way to define the inverse is to say that $a'_j = k$ if and only if $a_k = j$.

The inverse of a permutation was first defined by H. A. Rothe [in *Sammlung combinatorisch-analytischer Abhandlungen*, edited by C. F. Hindenburg, 2 (Leipzig: 1800), 263–305], who noticed an interesting connection between inverses and inversions: *The inverse of a permutation has exactly as many inversions as the permutation itself.* Rothe's proof of this fact was not the simplest possible one, but it is instructive and quite pretty nevertheless. We construct an $n \times n$ chessboard having a dot in column j of row i whenever $a_i = j$. Then we put \times 's in all squares that have dots lying both below (in the same column) and to their right (in the same row). For example, the diagram for 5 9 1 8 2 6 4 7 3 is

\times	\times	\times	\times	\bullet					
\times	\times	\times	\times		\times	\times	\times	\bullet	
\bullet									
	\times	\times	\times		\times	\times	\bullet		
	\bullet								
		\times	\times		\bullet				
		\times	\bullet						
		\times				\bullet			
		\bullet							

The number of \times 's is the number of inversions, since it is easy to see that b_j is the number of \times 's in column j . Now if we transpose the diagram—interchanging rows and columns—we get the diagram corresponding to the inverse of the original permutation. Hence the number of \times 's (the number of inversions) is the same in both cases. Rothe used this fact to prove that the determinant of a matrix is unchanged when the matrix is transposed.

The analysis of several sorting algorithms involves the knowledge of how many permutations of n elements have exactly k inversions. Let us denote that number by $I_n(k)$; Table 1 lists the first few values of this function.

By considering the inversion table $b_1 b_2 \dots b_n$, it is obvious that $I_n(0) = 1$, $I_n(1) = n - 1$, and there is a symmetry property

$$I_n \left(\binom{n}{2} - k \right) = I_n(k). \quad (6)$$

Table 1
PERMUTATIONS WITH k INVERSIONS

n	$I_n(0)$	$I_n(1)$	$I_n(2)$	$I_n(3)$	$I_n(4)$	$I_n(5)$	$I_n(6)$	$I_n(7)$	$I_n(8)$	$I_n(9)$	$I_n(10)$	$I_n(11)$
1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0	0	0	0
3	1	2	2	1	0	0	0	0	0	0	0	0
4	1	3	5	6	5	3	1	0	0	0	0	0
5	1	4	9	15	20	22	20	15	9	4	1	0
6	1	5	14	29	49	71	90	101	101	90	71	49

Furthermore, since each of the b 's can be chosen independently of the others, it is not difficult to see that the generating function

$$G_n(z) = I_n(0) + I_n(1)z + I_n(2)z^2 + \dots \tag{7}$$

satisfies $G_n(z) = (1 + z + \dots + z^{n-1})G_{n-1}(z)$; hence it has the comparatively simple form noticed by O. Rodrigues [*J. de Math.* **4** (1839), 236–240]:

$$(1 + z + \dots + z^{n-1}) \dots (1 + z)(1) = (1 - z^n) \dots (1 - z^2)(1 - z)/(1 - z)^n. \tag{8}$$

From this generating function, we can easily extend Table 1, and we can verify that the numbers below the zigzag line in that table satisfy

$$I_n(k) = I_n(k - 1) + I_{n-1}(k), \quad \text{for } k < n. \tag{9}$$

(This relation does *not* hold *above* the zigzag line.) A more complicated argument (see exercise 14) shows that, in fact, we have the formulas

$$\begin{aligned} I_n(2) &= \binom{n}{2} - 1, & n \geq 2; \\ I_n(3) &= \binom{n+1}{3} - \binom{n}{1}, & n \geq 3; \\ I_n(4) &= \binom{n+2}{4} - \binom{n+1}{2}, & n \geq 4; \\ I_n(5) &= \binom{n+3}{5} - \binom{n+2}{3} + 1, & n \geq 5; \end{aligned}$$

in general, the formula for $I_n(k)$ contains about $1.6\sqrt{k}$ terms:

$$\begin{aligned} I_n(k) &= \binom{n+k-2}{k} - \binom{n+k-3}{k-2} + \binom{n+k-6}{k-5} + \binom{n+k-8}{k-7} - \dots \\ &+ (-1)^j \left(\binom{n+k-u_j-1}{k-u_j} + \binom{n+k-u_j-j-1}{k-u_j-j} \right) + \dots, \quad n \geq k, \tag{10} \end{aligned}$$

where $u_j = (3j^2 - j)/2$ is a so-called ‘‘pentagonal number.’’

If we divide $G_n(z)$ by $n!$ we get the generating function $g_n(z)$ for the probability distribution of the number of inversions in a random permutation

of n elements. This is the product

$$g_n(z) = h_1(z)h_2(z) \dots h_n(z), \quad (11)$$

where $h_k(z) = (1 + z + \dots + z^{k-1})/k$ is the generating function for the uniform distribution of a random nonnegative integer less than k . It follows that

$$\begin{aligned} \text{mean}(g_n) &= \text{mean}(h_1) + \text{mean}(h_2) + \dots + \text{mean}(h_n) \\ &= 0 + \frac{1}{2} + \dots + \frac{n-1}{2} = \frac{n(n-1)}{4}; \end{aligned} \quad (12)$$

$$\begin{aligned} \text{var}(g_n) &= \text{var}(h_1) + \text{var}(h_2) + \dots + \text{var}(h_n) \\ &= 0 + \frac{1}{4} + \dots + \frac{n^2-1}{12} = \frac{n(2n+5)(n-1)}{72}. \end{aligned} \quad (13)$$

So the average number of inversions is rather large, about $\frac{1}{4}n^2$; the standard deviation is also rather large, about $\frac{1}{6}n^{3/2}$.

A remarkable discovery about the distribution of inversions was made by P. A. MacMahon [*Amer. J. Math.* **35** (1913), 281–322]. Let us define the *index* of the permutation $a_1 a_2 \dots a_n$ as the sum of all subscripts j such that $a_j > a_{j+1}$, $1 \leq j < n$. For example, the index of 5 9 1 8 2 6 4 7 3 is $2 + 4 + 6 + 8 = 20$. By coincidence the index is the same as the number of inversions in this case. If we list the 24 permutations of $\{1, 2, 3, 4\}$, namely

Permutation	Index	Inversions	Permutation	Index	Inversions
1 2 3 4	0	0	3 1 2 4	1	2
1 2 4 3	3	1	3 1 4 2	4	3
1 3 2 4	2	1	3 2 1 4	3	3
1 3 4 2	3	2	3 2 4 1	4	4
1 4 2 3	2	2	3 4 1 2	2	4
1 4 3 2	5	3	3 4 2 1	5	5
2 1 3 4	1	1	4 1 2 3	1	3
2 1 4 3	4	2	4 1 3 2	4	4
2 3 1 4	2	2	4 2 1 3	3	4
2 3 4 1	3	3	4 2 3 1	4	5
2 4 1 3	2	3	4 3 1 2	3	5
2 4 3 1	5	4	4 3 2 1	6	6

we see that *the number of permutations having a given index, k , is the same as the number having k inversions.*

At first this fact might appear to be almost obvious, but further scrutiny makes it very mysterious. MacMahon gave an ingenious indirect proof, as follows: Let $\text{ind}(a_1 a_2 \dots a_n)$ be the index of the permutation $a_1 a_2 \dots a_n$, and let

$$H_n(z) = \sum z^{\text{ind}(a_1 a_2 \dots a_n)} \quad (14)$$

be the corresponding generating function; the sum in (14) is over all permutations of $\{1, 2, \dots, n\}$. We wish to show that $H_n(z) = G_n(z)$. For this purpose we will

define a one-to-one correspondence between arbitrary n -tuples (q_1, q_2, \dots, q_n) of nonnegative integers, on the one hand, and ordered pairs of n -tuples

$$((a_1, a_2, \dots, a_n), (p_1, p_2, \dots, p_n))$$

on the other hand, where $a_1 a_2 \dots a_n$ is a permutation of the indices $\{1, 2, \dots, n\}$ and $p_1 \geq p_2 \geq \dots \geq p_n \geq 0$. This correspondence will satisfy the condition

$$q_1 + q_2 + \dots + q_n = \text{ind}(a_1 a_2 \dots a_n) + (p_1 + p_2 + \dots + p_n). \quad (15)$$

The generating function $\sum z^{q_1+q_2+\dots+q_n}$, summed over all n -tuples of nonnegative integers (q_1, q_2, \dots, q_n) , is $Q_n(z) = 1/(1-z)^n$; and the generating function $\sum z^{p_1+p_2+\dots+p_n}$, summed over all n -tuples of integers (p_1, p_2, \dots, p_n) such that $p_1 \geq p_2 \geq \dots \geq p_n \geq 0$, is

$$P_n(z) = 1/(1-z)(1-z^2) \dots (1-z^n), \quad (16)$$

as shown in exercise 15. In view of (15), the one-to-one correspondence we are about to establish will prove that $Q_n(z) = H_n(z)P_n(z)$, that is,

$$H_n(z) = Q_n(z)/P_n(z). \quad (17)$$

But $Q_n(z)/P_n(z)$ is $G_n(z)$, by (8).

The desired correspondence is defined by a simple sorting procedure: Any n -tuple (q_1, q_2, \dots, q_n) can be rearranged into nonincreasing order $q_{a_1} \geq q_{a_2} \geq \dots \geq q_{a_n}$ in a stable manner, where $a_1 a_2 \dots a_n$ is a permutation such that $q_{a_j} = q_{a_{j+1}}$ implies $a_j < a_{j+1}$. We set $(p_1, p_2, \dots, p_n) = (q_{a_1}, q_{a_2}, \dots, q_{a_n})$ and then, for $1 \leq j < n$, subtract 1 from each of p_1, \dots, p_j for each j such that $a_j > a_{j+1}$. We still have $p_1 \geq p_2 \geq \dots \geq p_n$, because p_j was strictly greater than p_{j+1} whenever $a_j > a_{j+1}$. The resulting pair $((a_1, a_2, \dots, a_n), (p_1, p_2, \dots, p_n))$ satisfies (15), because the total reduction of the p 's is $\text{ind}(a_1 a_2 \dots a_n)$. For example, if $n = 9$ and $(q_1, \dots, q_9) = (3, 1, 4, 1, 5, 9, 2, 6, 5)$, we find $a_1 \dots a_9 = 685931724$ and $(p_1, \dots, p_9) = (5, 2, 2, 2, 2, 2, 1, 1, 1)$.

Conversely, we can easily go back to (q_1, q_2, \dots, q_n) when $a_1 a_2 \dots a_n$ and (p_1, p_2, \dots, p_n) are given. (See exercise 17.) So the desired correspondence has been established, and MacMahon's index theorem has been proved.

D. Foata and M. P. Schützenberger discovered a surprising extension of MacMahon's theorem, about 65 years after MacMahon's original publication: *The number of permutations of n elements that have k inversions and index l is the same as the number that have l inversions and index k .* In fact, Foata and Schützenberger found a simple one-to-one correspondence between permutations of the first kind and permutations of the second (see exercise 25).

EXERCISES

1. [10] What is the inversion table for the permutation 271845936? What permutation has the inversion table 50121200?

2. [M20] In the classical problem of Josephus (exercise 1.3.2–22), n men are initially arranged in a circle; the m th man is executed, the circle closes, and every m th man is repeatedly eliminated until all are dead. The resulting execution order is a permutation

of $\{1, 2, \dots, n\}$. For example, when $n = 8$ and $m = 4$ the order is 5 4 6 1 3 8 7 2 (man 1 is 5th out, etc.); the inversion table corresponding to this permutation is 3 6 3 1 0 0 1 0.

Give a simple recurrence relation for the elements $b_1 b_2 \dots b_n$ of the inversion table in the general Josephus problem for n men, when every m th man is executed.

3. [18] If the permutation $a_1 a_2 \dots a_n$ corresponds to the inversion table $b_1 b_2 \dots b_n$, what is the permutation $\bar{a}_1 \bar{a}_2 \dots \bar{a}_n$ that corresponds to the inversion table

$$(n - 1 - b_1)(n - 2 - b_2) \dots (0 - b_n)?$$

► 4. [20] Design an algorithm suitable for computer implementation that constructs the permutation $a_1 a_2 \dots a_n$ corresponding to a given inversion table $b_1 b_2 \dots b_n$ satisfying (3). [Hint: Consider a linked-memory technique.]

5. [35] The algorithm of exercise 4 requires an execution time roughly proportional to $n + b_1 + \dots + b_n$ on typical computers, and this is $\Theta(n^2)$ on the average. Is there an algorithm whose worst-case running time is substantially better than order n^2 ?

► 6. [26] Design an algorithm that computes the inversion table $b_1 b_2 \dots b_n$ corresponding to a given permutation $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$, where the running time is essentially proportional to $n \log n$ on typical computers.

7. [20] Several other kinds of inversion tables can be defined, corresponding to a given permutation $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$, besides the particular table $b_1 b_2 \dots b_n$ defined in the text; in this exercise we will consider three other types of inversion tables that arise in applications.

Let c_j be the number of inversions whose *first* component is j , that is, the number of elements to the *right* of j that are less than j . [Corresponding to (1) we have the table 0 0 0 1 4 2 1 5 7; clearly $0 \leq c_j < j$.] Let $B_j = b_{a_j}$ and $C_j = c_{a_j}$.

Show that $0 \leq B_j < j$ and $0 \leq C_j \leq n - j$, for $1 \leq j \leq n$; furthermore show that the permutation $a_1 a_2 \dots a_n$ can be determined uniquely when either $c_1 c_2 \dots c_n$ or $B_1 B_2 \dots B_n$ or $C_1 C_2 \dots C_n$ is given.

8. [M24] Continuing the notation of exercise 7, let $a'_1 a'_2 \dots a'_n$ be the inverse of the permutation $a_1 a_2 \dots a_n$, and let the corresponding inversion tables be $b'_1 b'_2 \dots b'_n$, $c'_1 c'_2 \dots c'_n$, $B'_1 B'_2 \dots B'_n$, and $C'_1 C'_2 \dots C'_n$. Find as many interesting relations as you can between the numbers $a_j, b_j, c_j, B_j, C_j, a'_j, b'_j, c'_j, B'_j, C'_j$.

► 9. [M21] Prove that, in the notation of exercise 7, the permutation $a_1 a_2 \dots a_n$ is an involution (that is, its own inverse) if and only if $b_j = C_j$ for $1 \leq j \leq n$.

10. [HM20] Consider Fig. 1 as a polyhedron in three dimensions. What is the diameter of the truncated octahedron (the distance between vertex 1234 and vertex 4321), if all of its edges have unit length?

11. [M25] If $\pi = a_1 a_2 \dots a_n$ is a permutation of $\{1, 2, \dots, n\}$, let

$$E(\pi) = \{(a_i, a_j) \mid i < j, a_i > a_j\}$$

be the set of its inversions, and let

$$\bar{E}(\pi) = \{(a_i, a_j) \mid i > j, a_i > a_j\}$$

be the non-inversions.

a) Prove that $E(\pi)$ and $\bar{E}(\pi)$ are transitive. (A set S of ordered pairs is called *transitive* if (a, c) is in S whenever both (a, b) and (b, c) are in S .)

b) Conversely, let E be any transitive subset of $T = \{(x, y) \mid 1 \leq y < x \leq n\}$ whose complement $\bar{E} = T \setminus E$ is also transitive. Prove that there exists a permutation π such that $E(\pi) = E$.

12. [M28] Continuing the notation of the previous exercise, prove that if π_1 and π_2 are permutations and if E is the smallest transitive set containing $E(\pi_1) \cup E(\pi_2)$, then \bar{E} is transitive. [Hence, if we say π_1 is “above” π_2 whenever $E(\pi_1) \subseteq E(\pi_2)$, a lattice of permutations is defined; there is a unique “lowest” permutation “above” two given permutations. Figure 1 is the lattice diagram when $n = 4$.]

13. [M23] It is well known that half of the terms in the expansion of a determinant have a plus sign, and half have a minus sign. In other words, there are just as many permutations with an *even* number of inversions as with an *odd* number, when $n \geq 2$. Show that, in general, the number of permutations having a number of inversions congruent to t modulo m is $n!/m$, regardless of the integer t , whenever $n \geq m$.

14. [M24] (F. Franklin.) A partition of n into k distinct parts is a representation $n = p_1 + p_2 + \cdots + p_k$, where $p_1 > p_2 > \cdots > p_k > 0$. For example, the partitions of 7 into distinct parts are 7, 6 + 1, 5 + 2, 4 + 3, 4 + 2 + 1. Let $f_k(n)$ be the number of partitions of n into k distinct parts; prove that $\sum_k (-1)^k f_k(n) = 0$, unless n has the form $(3j^2 \pm j)/2$, for some nonnegative integer j ; in the latter case the sum is $(-1)^j$. For example, when $n = 7$ the sum is $-1 + 3 - 1 = 1$, and $7 = (3 \cdot 2^2 + 2)/2$. [Hint: Represent a partition as an array of dots, putting p_i dots in the i th row, for $1 \leq i \leq k$. Find the smallest j such that $p_{j+1} < p_j - 1$, and encircle the rightmost dots in the first j rows. If $j < p_k$, these j dots can usually be removed, tilted 45° , and placed as a new $(k+1)$ st row. On the other hand if $j \geq p_k$, the k th row of dots can usually be removed, tilted 45° , and placed to the right of the circled dots. (See Fig. 2.) This process pairs off partitions having an odd number of rows with partitions having an even number of rows, in most cases, so only unpaired partitions must be considered in the sum.]

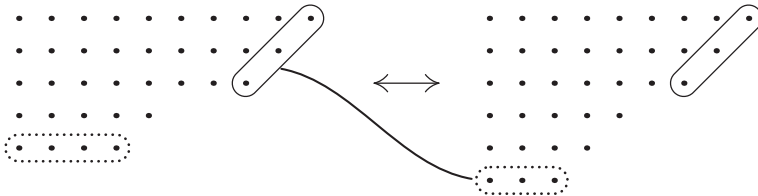


Fig. 2. Franklin's correspondence between partitions with distinct parts.

Note: As a consequence, we obtain Euler's formula

$$\begin{aligned} (1-z)(1-z^2)(1-z^3)\cdots &= 1 - z - z^2 + z^5 + z^7 - z^{12} - z^{15} + \cdots \\ &= \sum_{-\infty < j < \infty} (-1)^j z^{(3j^2+j)/2}. \end{aligned}$$

The generating function for ordinary partitions (whose parts are not necessarily distinct) is $\sum p(n)z^n = 1/(1-z)(1-z^2)(1-z^3)\cdots$; hence we obtain a nonobvious recurrence relation for the partition numbers,

$$p(n) = p(n-1) + p(n-2) - p(n-5) - p(n-7) + p(n-12) + p(n-15) - \cdots.$$

15. [M23] Prove that (16) is the generating function for partitions into at most n parts; that is, prove that the coefficient of z^m in $1/(1-z)(1-z^2)\dots(1-z^n)$ is the number of ways to write $m = p_1 + p_2 + \dots + p_n$ with $p_1 \geq p_2 \geq \dots \geq p_n \geq 0$. [Hint: Drawing dots as in exercise 14, show that there is a one-to-one correspondence between n -tuples (p_1, p_2, \dots, p_n) such that $p_1 \geq p_2 \geq \dots \geq p_n \geq 0$ and sequences (P_1, P_2, P_3, \dots) such that $n \geq P_1 \geq P_2 \geq P_3 \geq \dots \geq 0$, with the property that $p_1 + p_2 + \dots + p_n = P_1 + P_2 + P_3 + \dots$. In other words, partitions into at most n parts correspond to partitions into parts not exceeding n .]

16. [M25] (L. Euler.) Prove the following identities by interpreting both sides of the equations in terms of partitions:

$$\begin{aligned} \prod_{k \geq 0} \frac{1}{(1 - q^k z)} &= \frac{1}{(1 - z)(1 - qz)(1 - q^2 z) \dots} \\ &= 1 + \frac{z}{1 - q} + \frac{z^2}{(1 - q)(1 - q^2)} + \dots = \sum_{n \geq 0} z^n / \prod_{k=1}^n (1 - q^k). \end{aligned}$$

$$\begin{aligned} \prod_{k \geq 0} (1 + q^k z) &= (1 + z)(1 + qz)(1 + q^2 z) \dots \\ &= 1 + \frac{z}{1 - q} + \frac{z^2 q}{(1 - q)(1 - q^2)} + \dots = \sum_{n \geq 0} z^n q^{n(n-1)/2} / \prod_{k=1}^n (1 - q^k). \end{aligned}$$

17. [20] In MacMahon's correspondence defined at the end of this section, what are the 24 quadruples (q_1, q_2, q_3, q_4) for which $(p_1, p_2, p_3, p_4) = (0, 0, 0, 0)$?

18. [M30] (T. Hibbard, *CACM* **6** (1963), 210.) Let $n > 0$, and assume that a sequence of 2^n n -bit integers X_0, \dots, X_{2^n-1} has been generated at random, where each bit of each number is independently equal to 1 with probability p . Consider the sequence $X_0 \oplus 0, X_1 \oplus 1, \dots, X_{2^n-1} \oplus (2^n - 1)$, where \oplus denotes the "exclusive or" operation on the binary representations. Thus if $p = 0$, the sequence is $0, 1, \dots, 2^n - 1$, and if $p = 1$ it is $2^n - 1, \dots, 1, 0$; and when $p = \frac{1}{2}$, each element of the sequence is a random integer between 0 and $2^n - 1$. For general p this is a useful way to generate a sequence of random integers with a biased number of inversions, although the distribution of the elements of the sequence taken as a whole is uniform in the sense that each n -bit integer has the same distribution. What is the average number of inversions in such a sequence, as a function of the probability p ?

19. [M28] (C. Meyer.) When m is relatively prime to n , we know that the sequence $(m \bmod n)(2m \bmod n) \dots ((n-1)m \bmod n)$ is a permutation of $\{1, 2, \dots, n-1\}$. Show that the number of inversions of this permutation can be expressed in terms of Dedekind sums (see Section 3.3.3).

20. [M43] The following famous identity due to Jacobi [*Fundamenta Nova Theoriæ Functionum Ellipticarum* (1829), §64] is the basis of many remarkable relationships involving elliptic functions:

$$\begin{aligned} \prod_{k \geq 1} (1 - u^k v^{k-1})(1 - u^{k-1} v^k)(1 - u^k v^k) \\ &= (1 - u)(1 - v)(1 - uv)(1 - u^2 v)(1 - uv^2)(1 - u^2 v^2) \dots \\ &= 1 - (u + v) + (u^3 v + uv^3) - (u^6 v^3 + u^3 v^6) + \dots \\ &= \sum_{-\infty < j < +\infty} (-1)^j u^{\binom{j}{2}} v^{\binom{j+1}{2}}. \end{aligned}$$

For example, if we set $u = z$, $v = z^2$, we obtain Euler's formula of exercise 14. If we set $z = \sqrt{u/v}$, $q = \sqrt{uv}$, we obtain

$$\prod_{k \geq 1} (1 - q^{2k-1}z)(1 - q^{2k-1}z^{-1})(1 - q^{2k}) = \sum_{-\infty < n < \infty} (-1)^n z^n q^{n^2}.$$

Is there a combinatorial proof of Jacobi's identity, analogous to Franklin's proof of the special case in exercise 14? (Thus we want to consider "complex partitions"

$$m + ni = (p_1 + q_1i) + (p_2 + q_2i) + \cdots + (p_k + q_ki)$$

where the $p_j + q_ji$ are distinct nonzero complex numbers, p_j and q_j being nonnegative integers with $|p_j - q_j| \leq 1$. Jacobi's identity says that the number of such representations with k even is the same as the number with k odd, except when m and n are consecutive triangular numbers.) What other remarkable properties do complex partitions have?

► **21.** [M25] (G. D. Knott.) Show that the permutation $a_1 \dots a_n$ is obtainable with a stack, in the sense of exercise 2.2.1–5 or 2.3.1–6, if and only if $C_j \leq C_{j+1} + 1$ for $1 \leq j < n$ in the notation of exercise 7.

22. [M26] Given a permutation $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$, let h_j be the number of indices $i < j$ such that $a_i \in \{a_j+1, a_j+2, \dots, a_{j+1}\}$. (If $a_{j+1} < a_j$, the elements of this set "wrap around" from n to 1. When $j = n$ we use the set $\{a_n+1, a_n+2, \dots, n\}$.) For example, the permutation 5 9 1 8 2 6 4 7 3 leads to $h_1 \dots h_9 = 0 0 1 2 1 4 2 4 6$.

a) Prove that $a_1 a_2 \dots a_n$ can be reconstructed from the numbers $h_1 h_2 \dots h_n$.

b) Prove that $h_1 + h_2 + \cdots + h_n$ is the index of $a_1 a_2 \dots a_n$.

► **23.** [M27] (*Russian roulette.*) A group of n condemned men who prefer probability theory to number theory might choose to commit suicide by sitting in a circle and modifying Josephus's method (exercise 2) as follows: The first prisoner holds a gun and aims it at his head; with probability p he dies and leaves the circle. Then the second man takes the gun and proceeds in the same way. Play continues cyclically, with constant probability $p > 0$, until everyone is dead.

Let $a_j = k$ if man k is the j th to die. Prove that the death order $a_1 a_2 \dots a_n$ occurs with a probability that is a function only of n , p , and the index of the dual permutation $(n+1 - a_n) \dots (n+1 - a_2) (n+1 - a_1)$. What death order is least likely?

24. [M26] Given integers $t(1) t(2) \dots t(n)$ with $t(j) \geq j$, the *generalized index* of a permutation $a_1 a_2 \dots a_n$ is the sum of all subscripts j such that $a_j > t(a_{j+1})$, plus the total number of inversions such that $i < j$ and $t(a_j) \geq a_i > a_j$. Thus when $t(j) = j$ for all j , the generalized index is the same as the index; but when $t(j) \geq n$ for all j it is the number of inversions. Prove that the number of permutations whose generalized index equals k is the same as the number of permutations having k inversions. [Hint: Show that, if we take any permutation $a_1 \dots a_{n-1}$ of $\{1, \dots, n-1\}$ and insert the number n in all possible places, we increase the generalized index by the numbers $\{0, 1, \dots, n-1\}$ in some order.]

► **25.** [M30] (Foata and Schützenberger.) If $\alpha = a_1 \dots a_n$ is a permutation, let $\text{ind}(\alpha)$ be its index, and let $\text{inv}(\alpha)$ count its inversions.

a) Define a one-to-one correspondence that takes each permutation α of $\{1, \dots, n\}$ to a permutation $f(\alpha)$ that has the following two properties: (i) $\text{ind}(f(\alpha)) = \text{inv}(\alpha)$; (ii) for $1 \leq j < n$, the number j appears to the left of $j+1$ in $f(\alpha)$ if and only if it appears to the left of $j+1$ in α . What permutation does your

construction assign to $f(\alpha)$ when $\alpha = 198263745$? For what permutation α is $f(\alpha) = 198263745$? [*Hint*: If $n > 1$, write $\alpha = x_1\alpha_1x_2\alpha_2\dots x_k\alpha_k a_n$, where x_1, \dots, x_k are all the elements $< a_n$ if $a_1 < a_n$, otherwise x_1, \dots, x_k are all the elements $> a_n$; the other elements appear in (possibly empty) strings $\alpha_1, \dots, \alpha_k$. Compare the number of inversions of $h(\alpha) = \alpha_1x_1\alpha_2x_2\dots\alpha_kx_k$ to $\text{inv}(\alpha)$; in this construction the number a_n does not appear in $h(\alpha)$.]

- b) Use f to define another one-to-one correspondence g having the following two properties: (i) $\text{ind}(g(\alpha)) = \text{inv}(\alpha)$; (ii) $\text{inv}(g(\alpha)) = \text{ind}(\alpha)$. [*Hint*: Consider inverse permutations.]

26. [M25] What is the statistical correlation coefficient between the number of inversions and the index of a random permutation? (See Eq. 3.3.2–(24).)

27. [M37] Prove that, in addition to (15), there is a simple relationship between $\text{inv}(a_1 a_2 \dots a_n)$ and the n -tuple (q_1, q_2, \dots, q_n) . Use this fact to generalize the derivation of (17), obtaining an algebraic characterization of the bivariate generating function

$$H_n(w, z) = \sum w^{\text{inv}(a_1 a_2 \dots a_n)} z^{\text{ind}(a_1 a_2 \dots a_n)},$$

where the sum is over all $n!$ permutations $a_1 a_2 \dots a_n$.

- 28. [25] If $a_1 a_2 \dots a_n$ is a permutation of $\{1, 2, \dots, n\}$, its *total displacement* is defined to be $\sum_{j=1}^n |a_j - j|$. Find upper and lower bounds for total displacement in terms of the number of inversions.

29. [28] If $\pi = a_1 a_2 \dots a_n$ and $\pi' = a'_1 a'_2 \dots a'_n$ are permutations of $\{1, 2, \dots, n\}$, their product $\pi\pi'$ is $a'_{a_1} a'_{a_2} \dots a'_{a_n}$. Let $\text{inv}(\pi)$ denote the number of inversions, as in exercise 25. Show that $\text{inv}(\pi\pi') \leq \text{inv}(\pi) + \text{inv}(\pi')$, and that equality holds if and only if $\pi\pi'$ is “below” π' in the sense of exercise 12.

*5.1.2. Permutations of a Multiset

So far we have been discussing permutations of a *set* of elements; this is just a special case of the concept of permutations of a *multiset*. (A multiset is like a set except that it can have repetitions of identical elements. Some basic properties of multisets have been discussed in exercise 4.6.3–19.)

For example, consider the multiset

$$M = \{a, a, a, b, b, c, d, d, d, d\}, \tag{1}$$

which contains 3 a 's, 2 b 's, 1 c , and 4 d 's. We may also indicate the multiplicities of elements in another way, namely

$$M = \{3 \cdot a, 2 \cdot b, c, 4 \cdot d\}. \tag{2}$$

A permutation* of M is an arrangement of its elements into a row; for example,

$$c \ a \ b \ d \ d \ a \ b \ d \ a \ d.$$

From another point of view we would call this a string of letters, containing 3 a 's, 2 b 's, 1 c , and 4 d 's.

How many permutations of M are possible? If we regarded the elements of M as distinct, by subscripting them $a_1, a_2, a_3, b_1, b_2, c_1, d_1, d_2, d_3, d_4$,

* Sometimes called a “permatation.”

we would have $10! = 3,628,800$ permutations; but many of those permutations would actually be the same when we removed the subscripts. In fact, each permutation of M would occur exactly $3! 2! 1! 4! = 288$ times, since we can start with any permutation of M and put subscripts on the a 's in $3!$ ways, on the b 's (independently) in $2!$ ways, on the c in 1 way, and on the d 's in $4!$ ways. Therefore the true number of permutations of M is

$$\frac{10!}{3! 2! 1! 4!} = 12,600.$$

In general, we can see by this same argument that the number of permutations of any multiset is the multinomial coefficient

$$\binom{n}{n_1, n_2, \dots} = \frac{n!}{n_1! n_2! \dots}, \quad (3)$$

where n_1 is the number of elements of one kind, n_2 is the number of another kind, etc., and $n = n_1 + n_2 + \dots$ is the total number of elements.

The number of permutations of a set has been known for more than 1500 years. The Hebrew *Book of Creation* (c. A.D. 400), which was the earliest literary product of Jewish philosophical mysticism, gives the correct values of the first seven factorials, after which it says "Go on and compute what the mouth cannot express and the ear cannot hear." [*Sefer Yetzirah*, end of Chapter 4. See Solomon Gandz, *Studies in Hebrew Astronomy and Mathematics* (New York: Ktav, 1970), 494–496; Aryeh Kaplan, *Sefer Yetzirah* (York Beach, Maine: Samuel Weiser, 1993).] This is one of the first two known enumerations of permutations in history. The other occurs in the Indian classic *Anuyogadvārasūtra* (c. 500), rule 97, which gives the formula

$$6 \times 5 \times 4 \times 3 \times 2 \times 1 - 2$$

for the number of permutations of six elements that are neither in ascending nor descending order. [See G. Chakravarti, *Bull. Calcutta Math. Soc.* **24** (1932), 79–88. The *Anuyogadvārasūtra* is one of the books in the canon of Jainism, a religious sect that flourishes in India.]

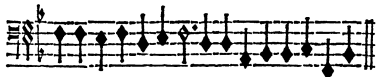
The corresponding formula for permutations of multisets seems to have appeared first in the *Līlāvātī* of Bhāskara (c. 1150), sections 270–271. Bhāskara stated the rule rather tersely, and illustrated it only with two simple examples $\{2, 2, 1, 1\}$ and $\{4, 8, 5, 5, 5\}$. Consequently the English translations of his work do not all state the rule correctly, although there is little doubt that Bhāskara knew what he was talking about. He went on to give the interesting formula

$$\frac{(4 + 8 + 5 + 5 + 5) \times 120 \times 11111}{5 \times 6}$$

for the sum of the 20 numbers $48555 + 45855 + \dots$.

The correct rule for counting permutations when elements are repeated was apparently unknown in Europe until Marin Mersenne stated it without proof as Proposition 10 in his elaborate treatise on melodic principles [*Harmonie Universelle* **2**, also entitled *Traitez de la Voix et des Chants* (1636), 129–130].

Mersenne was interested in the number of tunes that could be made from a given collection of notes; he observed, for example, that a theme by Boesset,



can be rearranged in exactly $15!/(4!3!3!2!) = 756,756,000$ ways.

The general rule (3) also appeared in Jean Prestet's *Éléments des Mathématiques* (Paris: 1675), 351–352, one of the very first expositions of combinatorial mathematics to be written in the Western world. Prestet stated the rule correctly for a general multiset, but illustrated it only in the simple case $\{a, a, b, b, c, c\}$. A few years later, John Wallis's *Discourse of Combinations* (Oxford: 1685), Chapter 2 (published with his *Treatise of Algebra*) gave a clearer and somewhat more detailed discussion of the rule.

In 1965, Dominique Foata introduced an ingenious idea called the “intercalation product,” which makes it possible to extend many of the known results about ordinary permutations to the general case of multiset permutations. [See *Publ. Inst. Statistique*, Univ. Paris, **14** (1965), 81–241; also *Lecture Notes in Math.* **85** (Springer, 1969).] Assuming that the elements of a multiset have been linearly ordered in some way, we may consider a *two-line notation* such as

$$\begin{pmatrix} a & a & a & b & b & c & d & d & d & d \\ c & a & b & d & d & a & b & d & a & d \end{pmatrix}, \quad (4)$$

where the top line contains the elements of M sorted into nondecreasing order and the bottom line is the permutation itself. The *intercalation product* $\alpha \uparrow \beta$ of two multiset permutations α and β is obtained by (a) expressing α and β in the two-line notation, (b) juxtaposing these two-line representations, and (c) sorting the columns into nondecreasing order of the top line. The sorting is supposed to be stable, in the sense that left-to-right order of elements in the bottom line is preserved when the corresponding top line elements are equal. For example, $c a d a b \uparrow b d d a d = c a b d d a b d a d$, since

$$\begin{pmatrix} a & a & b & c & d \\ c & a & d & a & b \end{pmatrix} \uparrow \begin{pmatrix} a & b & d & d & d \\ b & d & d & a & d \end{pmatrix} = \begin{pmatrix} a & a & a & b & b & c & d & d & d & d \\ c & a & b & d & d & a & b & d & a & d \end{pmatrix}. \quad (5)$$

It is easy to see that the intercalation product is associative:

$$(\alpha \uparrow \beta) \uparrow \gamma = \alpha \uparrow (\beta \uparrow \gamma); \quad (6)$$

it also satisfies two cancellation laws:

$$\begin{array}{ll} \pi \uparrow \alpha = \pi \uparrow \beta & \text{implies} \quad \alpha = \beta, \\ \alpha \uparrow \pi = \beta \uparrow \pi & \text{implies} \quad \alpha = \beta. \end{array} \quad (7)$$

There is an identity element,

$$\alpha \uparrow \epsilon = \epsilon \uparrow \alpha = \alpha, \quad (8)$$

where ϵ is the null permutation, the “arrangement” of the empty set. Although the commutative law is not valid in general (see exercise 2), we do have

$$\alpha \top \beta = \beta \top \alpha \quad \text{if } \alpha \text{ and } \beta \text{ have no elements in common.} \quad (9)$$

In an analogous fashion we can extend the concept of *cycles* in permutations to cases where elements are repeated; we let

$$(x_1 \ x_2 \ \dots \ x_n) \quad (10)$$

stand for the permutation obtained in two-line form by sorting the columns of

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ x_2 & x_3 & \dots & x_1 \end{pmatrix} \quad (11)$$

by their top elements in a stable manner. For example, we have

$$(d \ b \ d \ d \ a \ c \ a \ a \ b \ d) = \begin{pmatrix} d \ b \ d \ d \ a \ c \ a \ a \ b \ d \\ b \ d \ d \ a \ c \ a \ a \ b \ d \end{pmatrix} = \begin{pmatrix} a \ a \ a \ b \ b \ c \ d \ d \ d \ d \\ c \ a \ b \ d \ d \ a \ b \ d \ a \ d \end{pmatrix},$$

so the permutation (4) is actually a cycle. We might render this cycle in words by saying something like “ d goes to b goes to d goes to d goes ... goes to d goes back.” Note that these general cycles do not share all of the properties of ordinary cycles; $(x_1 \ x_2 \ \dots \ x_n)$ is not always the same as $(x_2 \ \dots \ x_n \ x_1)$.

We observed in Section 1.3.3 that every permutation of a set has a unique representation (up to order) as a product of disjoint cycles, where the “product” of permutations is defined by a law of composition. It is easy to see that *the product of disjoint cycles is exactly the same as their intercalation*; this suggests that we might be able to generalize the previous results, obtaining a unique representation (in some sense) for any permutation of a multiset, as the intercalation of cycles. In fact there are at least two natural ways to do this, each of which has important applications.

Equation (5) shows one way to factor $c \ a \ b \ d \ d \ a \ b \ d \ a \ d$ as the intercalation of shorter permutations; let us consider the general problem of finding all factorizations $\pi = \alpha \top \beta$ of a given permutation π . It will be helpful to consider a particular permutation, such as

$$\pi = \begin{pmatrix} a \ a \ b \ b \ b \ b \ b \ c \ c \ c \ d \ d \ d \ d \ d \\ d \ b \ c \ b \ c \ a \ c \ d \ a \ d \ d \ b \ b \ b \ d \end{pmatrix}, \quad (12)$$

as we investigate the factorization problem.

If we can write this permutation π in the form $\alpha \top \beta$, where α contains the letter a at least once, then the leftmost a in the top line of the two-line notation for α must appear over the letter d , so α must also contain at least one occurrence of the letter d . If we now look at the leftmost d in the top line of α , we see in the same way that it must appear over the letter d , so α must contain at least *two* d 's. Looking at the second d , we see that α also contains at least one b . We have deduced the partial result

$$\alpha = \begin{pmatrix} a & \dots & b & \dots & d & d & \dots \\ d & & & & d & b & \dots \end{pmatrix} \quad (13)$$

on the sole assumption that α is a left factor of π containing the letter a . Proceeding in the same manner, we find that the b in the top line of (13) must appear over the letter c , etc. Eventually this process will reach the letter a again, and we can identify this a with the first a if we choose to do so. The argument we have just made essentially proves that any left factor α of (12) that contains the letter a has the form $(d d b c d b b c a) \uparrow \alpha'$, for some permutation α' . (It is convenient to write the a last in the cycle, instead of first; this arrangement is permissible since there is only one a .) Similarly, if we had assumed that α contains the letter b , we would have deduced that $\alpha = (c d d b) \uparrow \alpha''$ for some α'' .

In general, this argument shows that, *if we have any factorization $\alpha \uparrow \beta = \pi$, where α contains a given letter y , exactly one cycle of the form*

$$(x_1 \dots x_n y), \quad n \geq 0, \quad x_1, \dots, x_n \neq y, \quad (14)$$

is a left factor of α . This cycle is easily determined when π and y are given; it is the shortest left factor of π that contains the letter y . One of the consequences of this observation is the following theorem:

Theorem A. *Let the elements of the multiset M be linearly ordered by the relation “ $<$ ”. Every permutation π of M has a unique representation as the intercalation*

$$\pi = (x_{11} \dots x_{1n_1} y_1) \uparrow (x_{21} \dots x_{2n_2} y_2) \uparrow \dots \uparrow (x_{t1} \dots x_{tn_t} y_t), \quad t \geq 0, \quad (15)$$

where the following two conditions are satisfied:

$$y_1 \leq y_2 \leq \dots \leq y_t \quad \text{and} \quad y_i < x_{ij} \quad \text{for } 1 \leq j \leq n_i, 1 \leq i \leq t. \quad (16)$$

(In other words, the last element in each cycle is smaller than every other element, and the sequence of last elements is in nondecreasing order.)

Proof. If $\pi = \epsilon$, we obtain such a factorization by letting $t = 0$. Otherwise we let y_1 be the smallest element permuted; and we determine $(x_{11} \dots x_{1n_1} y_1)$, the shortest left factor of π containing y_1 , as in the example above. Now $\pi = (x_{11} \dots x_{1n_1} y_1) \uparrow \rho$ for some permutation ρ ; by induction on the length, we can write

$$\rho = (x_{21} \dots x_{2n_2} y_2) \uparrow \dots \uparrow (x_{t1} \dots x_{tn_t} y_t), \quad t \geq 1,$$

where (16) is satisfied. This proves the existence of such a factorization.

Conversely, to prove that the representation (15) satisfying (16) is unique, clearly $t = 0$ if and only if π is the null permutation ϵ . When $t > 0$, (16) implies that y_1 is the smallest element permuted, and that $(x_{11} \dots x_{1n_1} y_1)$ is the shortest left factor containing y_1 . Therefore $(x_{11} \dots x_{1n_1} y_1)$ is uniquely determined; by the cancellation law (7) and induction, the representation is unique. ■

For example, the “canonical” factorization of (12), satisfying the given conditions, is

$$(d d b c d b b c a) \uparrow (b a) \uparrow (c d b) \uparrow (d), \quad (17)$$

if $a < b < c < d$.

It is important to note that we can actually drop the parentheses and the τ 's in this representation, without ambiguity! Each cycle ends just after the first appearance of the smallest remaining element. So this construction associates the permutation

$$\pi' = d d b c d b b c a b a c d b d$$

with the original permutation

$$\pi = d b c b c a c d a d d b b b d.$$

Whenever the two-line representation of π had a column of the form $\begin{smallmatrix} y \\ x \end{smallmatrix}$, where $x < y$, the associated permutation π' has a corresponding pair of adjacent elements $\dots y x \dots$. Thus our example permutation π has three columns of the form $\begin{smallmatrix} d \\ b \end{smallmatrix}$, and π' has three occurrences of the pair db . In general this construction establishes the following remarkable theorem:

Theorem B. *Let M be a multiset. There is a one-to-one correspondence between the permutations of M such that, if π corresponds to π' , the following conditions hold:*

- a) *The leftmost element of π' equals the leftmost element of π .*
- b) *For all pairs of permuted elements (x, y) with $x < y$, the number of occurrences of the column $\begin{smallmatrix} y \\ x \end{smallmatrix}$ in the two-line notation of π is equal to the number of times x is immediately preceded by y in π' . ■*

When M is a set, this is essentially the same as the “unusual correspondence” we discussed near the end of Section 1.3.3, with unimportant changes. The more general result in Theorem B is quite useful for enumerating special kinds of permutations, since we can often solve a problem based on a two-line constraint more easily than the equivalent problem based on an adjacent-pair constraint.

P. A. MacMahon considered problems of this type in his extraordinary book *Combinatory Analysis 1* (Cambridge Univ. Press, 1915), 168–186. He gave a constructive proof of Theorem B in the special case that M contains only two different kinds of elements, say a and b ; his construction for this case is essentially the same as that given here, although he expressed it quite differently. For the case of three different elements a, b, c , MacMahon gave a complicated nonconstructive proof of Theorem B; the general case was first proved constructively by Foata [*Comptes Rendus Acad. Sci.* **258** (Paris, 1964), 1672–1675].

As a nontrivial example of Theorem B, let us find the number of strings of letters a, b, c containing exactly

- A occurrences of the letter a ;
- B occurrences of the letter b ;
- C occurrences of the letter c ;
- k occurrences of the adjacent pair of letters ca ;
- l occurrences of the adjacent pair of letters cb ;
- m occurrences of the adjacent pair of letters ba . (18)

The theorem tells us that this is the same as the number of two-line arrays of the form

$$\begin{array}{ccc}
 A & B & C \\
 \left(\begin{array}{ccc} \overbrace{a \ \dots \ a} & \overbrace{b \ \dots \ b} & \overbrace{c \ \dots \ c} \\ \underbrace{\square \ \dots \ \square} & \underbrace{\square \ \dots \ \square} & \underbrace{\square \ \dots \ \square} \end{array} \right) \\
 A-k-m \ a's & m \ a's & k \ a's \\
 \underbrace{\hspace{10em}} & \underbrace{\hspace{4em}} & \\
 B-l \ b's & l \ b's & \\
 \underbrace{\hspace{14em}} & & \\
 C \ c's & &
 \end{array} \tag{19}$$

The a 's can be placed in the second line in

$$\binom{A}{A-k-m} \binom{B}{m} \binom{C}{k} \quad \text{ways;}$$

then the b 's can be placed in the remaining positions in

$$\binom{B+k}{B-l} \binom{C-k}{l} \quad \text{ways.}$$

The positions that are still vacant must be filled by c 's; hence the desired number is

$$\binom{A}{A-k-m} \binom{B}{m} \binom{C}{k} \binom{B+k}{B-l} \binom{C-k}{l}. \tag{20}$$

Let us return to the question of finding all factorizations of a given permutation. Is there such a thing as a “prime” permutation, one that has no intercalation factors except itself and ϵ ? The discussion preceding Theorem A leads us quickly to conclude that a permutation is prime if and only if it is a cycle with no repeated elements. For if it is such a cycle, our argument proves that there are no left factors except ϵ and the cycle itself. And if a permutation contains a repeated element y , it has a nontrivial cyclic left factor in which y appears only once.

A nonprime permutation can be factored into smaller and smaller pieces until it has been expressed as a product of primes. Furthermore we can show that the factorization is unique, if we neglect the order of factors that commute:

Theorem C. *Every permutation of a multiset can be written as a product*

$$\sigma_1 \tau \sigma_2 \tau \cdots \tau \sigma_t, \quad t \geq 0, \tag{21}$$

where each σ_j is a cycle having no repeated elements. This representation is unique, in the sense that any two such representations of the same permutation may be transformed into each other by successively interchanging pairs of adjacent disjoint cycles.

The term “disjoint cycles” means cycles having no elements in common. As an example of this theorem, we can verify that the permutation

$$\begin{pmatrix} a & a & b & b & c & c & d \\ b & a & a & c & d & b & c \end{pmatrix}$$

has exactly five factorizations into primes, namely

$$\begin{aligned} (a \ b) \top (a) \top (c \ d) \top (b \ c) &= (a \ b) \top (c \ d) \top (a) \top (b \ c) \\ &= (a \ b) \top (c \ d) \top (b \ c) \top (a) \\ &= (c \ d) \top (a \ b) \top (b \ c) \top (a) \\ &= (c \ d) \top (a \ b) \top (a) \top (b \ c). \end{aligned} \quad (22)$$

Proof. We must show that the stated uniqueness property holds. By induction on the length of the permutation, it suffices to prove that if ρ and σ are unequal cycles having no repeated elements, and if

$$\rho \top \alpha = \sigma \top \beta,$$

then ρ and σ are disjoint, and

$$\alpha = \sigma \top \theta, \quad \beta = \rho \top \theta,$$

for some permutation θ .

If y is any element of the cycle ρ , then any left factor of $\sigma \top \beta$ containing the element y must have ρ as a left factor. So if ρ and σ have an element in common, σ is a multiple of ρ ; hence $\sigma = \rho$ (since they are primes), contradicting our assumption. Therefore the cycle containing y , having no elements in common with σ , must be a left factor of β . The proof is completed by using the cancellation law (7). ■

As an example of Theorem C, let us consider permutations of the multiset $M = \{A \cdot a, B \cdot b, C \cdot c\}$ consisting of A a 's, B b 's, and C c 's. Let $N(A, B, C, m)$ be the number of permutations of M whose two-line representation contains *no* columns of the forms $\begin{smallmatrix} a \\ a \end{smallmatrix}$, $\begin{smallmatrix} b \\ b \end{smallmatrix}$, $\begin{smallmatrix} c \\ c \end{smallmatrix}$, and exactly m columns of the form $\begin{smallmatrix} a \\ b \end{smallmatrix}$. It follows that there are exactly $A - m$ columns of the form $\begin{smallmatrix} a \\ c \end{smallmatrix}$, $B - m$ of the form $\begin{smallmatrix} b \\ c \end{smallmatrix}$, $C - B + m$ of the form $\begin{smallmatrix} c \\ a \end{smallmatrix}$, $C - A + m$ of the form $\begin{smallmatrix} b \\ c \end{smallmatrix}$, and $A + B - C - m$ of the form $\begin{smallmatrix} b \\ a \end{smallmatrix}$. Hence

$$N(A, B, C, m) = \binom{A}{m} \binom{B}{C - A + m} \binom{C}{B - m}. \quad (23)$$

Theorem C tells us that we can count these permutations in another way: Since columns of the form $\begin{smallmatrix} a \\ a \end{smallmatrix}$, $\begin{smallmatrix} b \\ b \end{smallmatrix}$, $\begin{smallmatrix} c \\ c \end{smallmatrix}$ are excluded, the only possible prime factors of the permutation are

$$(a \ b), \quad (a \ c), \quad (b \ c), \quad (a \ b \ c), \quad (a \ c \ b). \quad (24)$$

Each pair of these cycles has at least one letter in common, so the factorization into primes is completely unique. If the cycle $(a \ b \ c)$ occurs k times in the factorization, our previous assumptions imply that $(a \ b)$ occurs $m - k$ times,

$(b\ c)$ occurs $C - A + m - k$ times, $(a\ c)$ occurs $C - B + m - k$ times, and $(a\ c\ b)$ occurs $A + B - C - 2m + k$ times. Hence $N(A, B, C, m)$ is the number of permutations of these cycles (a multinomial coefficient), summed over k :

$$\begin{aligned} N(A, B, C, m) &= \sum_k \frac{(C + m - k)!}{(m - k)! (C - A + m - k)! (C - B + m - k)! k! (A + B - C - 2m + k)!} \\ &= \sum_k \binom{m}{k} \binom{A}{m} \binom{A - m}{C - B + m - k} \binom{C + m - k}{A}. \end{aligned} \quad (25)$$

Comparing this with (23), we find that the following identity must be valid:

$$\sum_k \binom{m}{k} \binom{A - m}{C - B + m - k} \binom{C + m - k}{A} = \binom{B}{C - A + m} \binom{C}{B - m}. \quad (26)$$

This turns out to be the identity we met in exercise 1.2.6–31, namely

$$\sum_j \binom{M - R + S}{j} \binom{N + R - S}{N - j} \binom{R + j}{M + N} = \binom{R}{M} \binom{S}{N}, \quad (27)$$

with $M = A + B - C - m$, $N = C - B + m$, $R = B$, $S = C$, and $j = C - B + m - k$.

Similarly we can count the number of permutations of $\{A \cdot a, B \cdot b, C \cdot c, D \cdot d\}$ such that the number of columns of various types is specified as follows:

Column	a	a	b	b	c	c	d	d
type:	d	b	a	c	b	d	a	c
Frequency:	r	$A - r$	q	$B - q$	$B - A + r$	$D - r$	$A - q$	$D - A + q$

(28)

(Here $A + C = B + D$.) The possible cycles occurring in a prime factorization of such permutations are then

Cycle:	$(a\ b)$	$(b\ c)$	$(c\ d)$	$(d\ a)$	$(a\ b\ c\ d)$	$(d\ c\ b\ a)$
Frequency:	$A - r - s$	$B - q - s$	$D - r - s$	$A - q - s$	s	$q - A + r + s$

(29)

for some s (see exercise 12). In this case the cycles $(a\ b)$ and $(c\ d)$ commute with each other, and so do $(b\ c)$ and $(d\ a)$, so we must count the number of distinct prime factorizations. It turns out (see exercise 10) that there is always a unique factorization such that no $(c\ d)$ is immediately followed by $(a\ b)$, and no $(d\ a)$ is immediately followed by $(b\ c)$. Hence by the result of exercise 13, we have

$$\begin{aligned} &\sum_{s,t} \binom{B}{t} \binom{A - q - s}{A - r - s - t} \binom{B + D - r - s - t}{B - q - s} \\ &\quad \times \frac{D!}{(D - r - s)! (A - q - s)! s! (q - A + r + s)!} \\ &= \binom{A}{r} \binom{B + D - A}{D - r} \binom{B}{q} \binom{D}{A - q}. \end{aligned}$$

Taking out the factor $\binom{D}{A-q}$ from both sides and simplifying the factorials slightly leaves us with the complicated-looking five-parameter identity

$$\sum_{s,t} \binom{B}{t} \binom{A-r-t}{s} \binom{B+D-r-s-t}{D+q-r-t} \binom{D-A+q}{D-r-s} \binom{A-q}{r+t-q} \\ = \binom{A}{r} \binom{B+D-A}{D-r} \binom{B}{q}. \quad (30)$$

The sum on s can be performed using (27), and the resulting sum on t is easily evaluated; so, after all this work, we were not fortunate enough to discover any identities that we didn't already know how to derive. But at least we have learned how to count certain kinds of permutations, in two different ways, and these counting techniques are good training for the problems that lie ahead.

EXERCISES

1. [M05] *True or false:* Let M_1 and M_2 be multisets. If α is a permutation of M_1 and β is a permutation of M_2 , then $\alpha \uparrow \beta$ is a permutation of $M_1 \cup M_2$.
2. [10] The intercalation of $c a d a b$ and $b d d a d$ is computed in (5); find the intercalation $b d d a d \uparrow c a d a b$ that is obtained when the factors are interchanged.
3. [M13] Is the converse of (9) valid? In other words, if α and β commute under intercalation, must they have no letters in common?
4. [M11] The canonical factorization of (12), in the sense of Theorem A, is given in (17) when $a < b < c < d$. Find the corresponding canonical factorization when $d < c < b < a$.
5. [M23] Condition (b) of Theorem B requires $x < y$; what would happen if we weakened the relation to $x \leq y$?
6. [M15] How many strings are there that contain exactly m a 's, n b 's, and no other letters, with exactly k of the a 's preceded immediately by a b ?
7. [M21] How many strings on the letters a, b, c satisfying conditions (18) begin with the letter a ? with the letter b ? with c ?
- ▶ 8. [20] Find all factorizations of (12) into two factors $\alpha \uparrow \beta$.
9. [33] Write computer programs that perform the factorizations of a given multiset permutation into the forms mentioned in Theorems A and C.
- ▶ 10. [M30] *True or false:* Although the factorization into primes isn't quite unique, according to Theorem C, we can ensure uniqueness in the following way: "There is a linear ordering \prec of the set of primes such that every permutation of a multiset has a unique factorization $\sigma_1 \uparrow \sigma_2 \uparrow \cdots \uparrow \sigma_n$ into primes subject to the condition that $\sigma_i \preceq \sigma_{i+1}$ whenever σ_i commutes with σ_{i+1} , for $1 \leq i < n$."
- ▶ 11. [M26] Let $\sigma_1, \sigma_2, \dots, \sigma_t$ be cycles without repeated elements. Define a partial ordering \prec on the t objects $\{x_1, \dots, x_t\}$ by saying that $x_i \prec x_j$ if $i < j$ and σ_i has at least one letter in common with σ_j . Prove the following connection between Theorem C and the notion of "topological sorting" (Section 2.2.3): *The number of distinct prime factorizations of $\sigma_1 \uparrow \sigma_2 \uparrow \cdots \uparrow \sigma_t$ is the number of ways to sort the given partial ordering topologically.* (For example, corresponding to (22) we find that there are five ways to sort the ordering $x_1 \prec x_2, x_3 \prec x_4, x_1 \prec x_4$ topologically.) Conversely, given any partial ordering on t elements, there is a set of cycles $\{\sigma_1, \sigma_2, \dots, \sigma_t\}$ that defines it in the stated way.

12. [M16] Show that (29) is a consequence of the assumptions of (28).
 13. [M21] Prove that the number of permutations of the multiset

$$\{A \cdot a, B \cdot b, C \cdot c, D \cdot d, E \cdot e, F \cdot f\}$$

containing no occurrences of the adjacent pairs of letters ca and db is

$$\sum_t \binom{D}{A-t} \binom{A+B+E+F}{t} \binom{A+B+C+E+F-t}{B} \binom{C+D+E+F}{C, D, E, F}.$$

14. [M30] One way to define the inverse π^- of a general permutation π , suggested by other definitions in this section, is to interchange the lines of the two-line representation of π and then to do a stable sort of the columns in order to bring the top row into nondecreasing order. For example, if $a < b < c < d$, this definition implies that the inverse of $c a b d d a b d a d$ is $a c d a d a b b d d$.

Explore properties of this inversion operation; for example, does it have any simple relation with intercalation products? Can we count the number of permutations such that $\pi = \pi^-$?

- 15. [M25] Prove that the permutation $a_1 \dots a_n$ of the multiset

$$\{n_1 \cdot x_1, n_2 \cdot x_2, \dots, n_m \cdot x_m\},$$

where $x_1 < x_2 < \dots < x_m$ and $n_1 + n_2 + \dots + n_m = n$, is a cycle if and only if the directed graph with vertices $\{x_1, x_2, \dots, x_m\}$ and arcs from x_j to $a_{n_1+\dots+n_j}$ contains precisely one oriented cycle. In the latter case, the number of ways to represent the permutation in cycle form is the length of the oriented cycle. For example, the directed graph corresponding to

$$\left(\begin{array}{cccccccc} a & a & a & b & b & c & c & c & d & d \\ d & c & b & a & c & a & a & b & d & c \end{array} \right) \quad \text{is} \quad \begin{array}{c} a \bullet \longrightarrow b \\ \nearrow \\ d \bullet \longrightarrow c \\ \searrow \end{array}$$

and the two ways to represent the permutation as a cycle are $(b a d d c a c a b c)$ and $(c a d d c a c b a b)$.

16. [M35] We found the generating function for *inversions* of permutations in the previous section, Eq. 5.1.1-(8), in the special case that a set was being permuted. Show that, in general, if a *multiset* is permuted, the generating function for inversions of $\{n_1 \cdot x_1, n_2 \cdot x_2, \dots\}$ is the “ z -multinomial coefficient”

$$\binom{n}{n_1, n_2, \dots}_z = \frac{n!_z}{n_1!_z n_2!_z \dots}, \quad \text{where} \quad m!_z = \prod_{k=1}^m (1 + z + \dots + z^{k-1}).$$

[Compare with (3) and with the definition of z -nomial coefficients in Eq. 1.2.6-(40).]

17. [M24] Find the average and standard deviation of the number of inversions in a random permutation of a given multiset, using the generating function found in exercise 16.

18. [M30] (P. A. MacMahon.) The *index* of a permutation $a_1 a_2 \dots a_n$ was defined in the previous section; and we proved that the number of permutations of a given set that have a given index k is the same as the number of permutations that have k inversions. Does the same result hold for permutations of a given multiset?

19. [HM28] Define the *Möbius function* $\mu(\pi)$ of a permutation π to be 0 if π contains repeated elements, otherwise $(-1)^k$ if π is the product of k primes. (Compare with the definition of the ordinary Möbius function, exercise 4.5.2–10.)

a) Prove that if $\pi \neq \epsilon$, we have

$$\sum \mu(\lambda) = 0,$$

summed over all permutations λ that are left factors of π (namely all λ such that $\pi = \lambda \uparrow \rho$ for some ρ).

b) Given that $x_1 < x_2 < \dots < x_m$ and $\pi = x_{i_1} x_{i_2} \dots x_{i_n}$, where $1 \leq i_k \leq m$ for $1 \leq k \leq n$, prove that

$$\mu(\pi) = (-1)^n \epsilon(i_1 i_2 \dots i_n), \quad \text{where } \epsilon(i_1 i_2 \dots i_n) = \text{sign} \prod_{1 \leq j < k \leq n} (i_k - i_j).$$

► **20.** [HM33] (D. Foata.) Let (a_{ij}) be any matrix of real numbers. In the notation of exercise 19(b), define $\nu(\pi) = a_{i_1 j_1} \dots a_{i_n j_n}$, where the two-line notation for π is

$$\begin{pmatrix} x_{i_1} & x_{i_2} & \dots & x_{i_n} \\ x_{j_1} & x_{j_2} & \dots & x_{j_n} \end{pmatrix}.$$

This function is useful in the computation of generating functions for permutations of a multiset, because $\sum \nu(\pi)$, summed over all permutations π of the multiset

$$\{n_1 \cdot x_1, \dots, n_m \cdot x_m\},$$

will be the generating function for the number of permutations satisfying certain restrictions. For example, if we take $a_{ij} = z$ for $i = j$, and $a_{ij} = 1$ for $i \neq j$, then $\sum \nu(\pi)$ is the generating function for the number of “fixed points” (columns in which the top and bottom entries are equal). In order to study $\sum \nu(\pi)$ for all multisets simultaneously, we consider the function

$$G = \sum \pi \nu(\pi)$$

summed over all π in the set $\{x_1, \dots, x_m\}^*$ of all permutations of multisets involving the elements x_1, \dots, x_m , and we look at the coefficient of $x_1^{n_1} \dots x_m^{n_m}$ in G .

In this formula for G we are treating π as the product of the x 's. For example, when $m = 2$ we have

$$\begin{aligned} G &= 1 + x_1 \nu(x_1) + x_2 \nu(x_2) + x_1 x_1 \nu(x_1 x_1) + x_1 x_2 \nu(x_1 x_2) + x_2 x_1 \nu(x_2 x_1) + x_2 x_2 \nu(x_2 x_2) + \dots \\ &= 1 + x_1 a_{11} + x_2 a_{22} + x_1^2 a_{11}^2 + x_1 x_2 a_{11} a_{22} + x_1 x_2 a_{21} a_{12} + x_2^2 a_{22}^2 + \dots \end{aligned}$$

Thus the coefficient of $x_1^{n_1} \dots x_m^{n_m}$ in G is $\sum \nu(\pi)$ summed over all permutations π of $\{n_1 \cdot x_1, \dots, n_m \cdot x_m\}$. It is not hard to see that this coefficient is also the coefficient of $x_1^{n_1} \dots x_m^{n_m}$ in the expression

$$(a_{11} x_1 + \dots + a_{1m} x_m)^{n_1} (a_{21} x_1 + \dots + a_{2m} x_m)^{n_2} \dots (a_{m1} x_1 + \dots + a_{mm} x_m)^{n_m}.$$

The purpose of this exercise is to prove what P. A. MacMahon called a “Master Theorem” in his *Combinatory Analysis* **1** (1915), Section 3, namely the formula

$$G = 1/D, \quad \text{where } D = \det \begin{pmatrix} 1 - a_{11} x_1 & -a_{12} x_2 & \dots & -a_{1m} x_m \\ -a_{21} x_1 & 1 - a_{22} x_2 & & -a_{2m} x_m \\ \vdots & & & \vdots \\ -a_{m1} x_1 & -a_{m2} x_2 & \dots & 1 - a_{mm} x_m \end{pmatrix}.$$

For example, if $a_{ij} = 1$ for all i and j , this formula gives

$$G = 1/(1 - (x_1 + x_2 + \cdots + x_m)),$$

and the coefficient of $x_1^{n_1} \cdots x_m^{n_m}$ turns out to be $(n_1 + \cdots + n_m)!/n_1! \cdots n_m!$, as it should. To prove the Master Theorem, show that

- $\nu(\pi \tau \rho) = \nu(\pi)\nu(\rho)$;
- $D = \sum \pi \mu(\pi)\nu(\pi)$, in the notation of exercise 19, summed over all permutations π in $\{x_1, \dots, x_m\}^*$;
- therefore $D \cdot G = 1$.

21. [M21] Given n_1, \dots, n_m , and $d \geq 0$, how many permutations $a_1 a_2 \dots a_n$ of the multiset $\{n_1 \cdot 1, \dots, n_m \cdot m\}$ satisfy $a_{j+1} \geq a_j - d$ for $1 \leq j < n = n_1 + \cdots + n_m$?

22. [M30] Let $P(x_1^{n_1} \cdots x_m^{n_m})$ denote the set of all possible permutations of the multiset $\{n_1 \cdot x_1, \dots, n_m \cdot x_m\}$, and let $P_0(x_0^{n_0} x_1^{n_1} \cdots x_m^{n_m})$ be the subset of $P(x_0^{n_0} x_1^{n_1} \cdots x_m^{n_m})$ in which the first n_0 elements are $\neq x_0$.

- Given a number t with $1 \leq t < m$, find a one-to-one correspondence between $P(1^{n_1} \cdots m^{n_m})$ and the set of all ordered pairs of permutations that belong respectively to $P_0(0^p 1^{n_1} \cdots t^{n_t})$ and $P_0(0^p (t+1)^{n_{t+1}} \cdots m^{n_m})$, for some $p \geq 0$. [Hint: For each $\pi = a_1 \dots a_n \in P(1^{n_1} \cdots m^{n_m})$, let $l(\pi)$ be the permutation obtained by replacing $t+1, \dots, m$ by 0 and erasing all 0s in the last $n_{t+1} + \cdots + n_m$ positions; similarly, let $r(\pi)$ be the permutation obtained by replacing $1, \dots, t$ by 0 and erasing all 0s in the first $n_1 + \cdots + n_t$ positions.]
- Prove that the number of permutations of $P_0(0^{n_0} 1^{n_1} \cdots m^{n_m})$ whose two-line form has p_j columns $\begin{smallmatrix} 0 \\ j \end{smallmatrix}$ and q_j columns $\begin{smallmatrix} j \\ 0 \end{smallmatrix}$ is

$$\frac{|P(x_1^{p_1} \cdots x_m^{p_m} y_1^{n_1-p_1} \cdots y_m^{n_m-p_m})| |P(x_1^{q_1} \cdots x_m^{q_m} y_1^{n_1-q_1} \cdots y_m^{n_m-q_m})|}{|P_0(0^{n_0} 1^{n_1} \cdots m^{n_m})|}.$$

- Let $w_1, \dots, w_m, z_1, \dots, z_m$ be complex numbers on the unit circle. Define the weight $w(\pi)$ of a permutation $\pi \in P(1^{n_1} \cdots m^{n_m})$ as the product of the weights of its columns in two-line form, where the weight of $\begin{smallmatrix} j \\ k \end{smallmatrix}$ is w_j/w_k if j and k are both $\leq t$ or both $> t$, otherwise it is z_j/z_k . Prove that the sum of $w(\pi)$ over all $\pi \in P(1^{n_1} \cdots m^{n_m})$ is

$$\sum_{p \geq 0} \frac{p!^2 (n_{\leq t} - p)! (n_{> t} - p)!}{n_1! \cdots n_m!} \left| \sum \binom{n_1}{p_1} \cdots \binom{n_m}{p_m} \left(\frac{w_1}{z_1}\right)^{p_1} \cdots \left(\frac{w_m}{z_m}\right)^{p_m} \right|^2,$$

where $n_{\leq t}$ is $n_1 + \cdots + n_t$, $n_{> t}$ is $n_{t+1} + \cdots + n_m$, and the inner sum is over all (p_1, \dots, p_m) such that $p_{\leq t} = p_{> t} = p$.

23. [M23] A strand of DNA can be thought of as a word on a four-letter alphabet. Suppose we copy a strand of DNA and break it completely into one-letter bases, then recombine those bases at random. If the resulting strand is placed next to the original, prove that the number of places in which they differ is more likely to be even than odd. [Hint: Apply the previous exercise.]

24. [27] Consider any relation R that might hold between two unordered pairs of letters; if $\{w, x\}R\{y, z\}$ we say $\{w, x\}$ preserves $\{y, z\}$, otherwise $\{w, x\}$ moves $\{y, z\}$.

The operation of *transposing* $\begin{smallmatrix} w & x \\ y & z \end{smallmatrix}$ with respect to R replaces $\begin{smallmatrix} w & x \\ y & z \end{smallmatrix}$ by $\begin{smallmatrix} x & w \\ y & z \end{smallmatrix}$ or $\begin{smallmatrix} x & w \\ z & y \end{smallmatrix}$, according as the pair $\{w, x\}$ preserves or moves the pair $\{y, z\}$, assuming that $w \neq x$ and $y \neq z$; if $w = x$ or $y = z$ the transposition always produces $\begin{smallmatrix} x & w \\ z & y \end{smallmatrix}$.

The operation of *sorting* a two-line array $(\begin{smallmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{smallmatrix})$ with respect to R repeatedly finds the largest x_j such that $x_j > x_{j+1}$ and transposes columns j and $j + 1$, until eventually $x_1 \leq \dots \leq x_n$. (We do not require $y_1 \dots y_n$ to be a permutation of $x_1 \dots x_n$.)

- a) Given $(\begin{smallmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{smallmatrix})$, prove that for every $x \in \{x_1, \dots, x_n\}$ there is a unique $y \in \{y_1, \dots, y_n\}$ such that $\text{sort}(\begin{smallmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_n \end{smallmatrix}) = \text{sort}(\begin{smallmatrix} x'_1 & \dots & x'_n \\ y'_1 & \dots & y'_n \end{smallmatrix})$ for some $x'_1, y'_1, \dots, x'_n, y'_n$.
- b) Let $(\begin{smallmatrix} w_1 & \dots & w_k \\ y_1 & \dots & y_k \end{smallmatrix}) \textcircled{R} (\begin{smallmatrix} x_1 & \dots & x_l \\ z_1 & \dots & z_l \end{smallmatrix})$ denote the result of sorting $(\begin{smallmatrix} w_1 & \dots & w_k & x_1 & \dots & x_l \\ y_1 & \dots & y_k & z_1 & \dots & z_l \end{smallmatrix})$ with respect to R . For example, if R is always true, \textcircled{R} sorts $\{w_1, \dots, w_k, x_1, \dots, x_l\}$, but it simply juxtaposes $y_1 \dots y_k$ with $z_1 \dots z_l$; if R is always false, \textcircled{R} is the intercalation product Υ . Generalize Theorem A by proving that every permutation π of a multiset M has a unique representation of the form

$$\pi = (x_{11} \dots x_{1n_1} y_1) \textcircled{R} ((x_{21} \dots x_{2n_2} y_2) \textcircled{R} (\dots \textcircled{R} (x_{t1} \dots x_{tn_t} y_t) \dots))$$

satisfying (16), if we redefine cycle notation by letting the two-line array (11) correspond to the cycle $(x_2 \dots x_n x_1)$ instead of to $(x_1 x_2 \dots x_n)$. For example, suppose $\{w, x\}R\{y, z\}$ means that w, x, y , and z are distinct; then it turns out that the factorization of (12) analogous to (17) is

$$(d d b c a) \textcircled{R} ((c b b a) \textcircled{R} ((c d b) \textcircled{R} ((d b) \textcircled{R} (d))))).$$

(The operation \textcircled{R} does not always obey the associative law; parentheses in the generalized factorization should be nested from right to left.)

*5.1.3. Runs

In Chapter 3 we analyzed the lengths of upward runs in permutations, as a way to test the randomness of a sequence. If we place a vertical line at both ends of a permutation $a_1 a_2 \dots a_n$ and also between a_j and a_{j+1} whenever $a_j > a_{j+1}$, the *runs* are the segments between pairs of lines. For example, the permutation

$$| 3 \ 5 \ 7 \ | \ 1 \ 6 \ 8 \ 9 \ | \ 4 \ | \ 2 \ |$$

has four runs. The theory developed in Section 3.3.2G determines the average number of runs of length k in a random permutation of $\{1, 2, \dots, n\}$, as well as the covariance of the numbers of runs of lengths j and k . Runs are important in the study of sorting algorithms, because they represent sorted segments of the data, so we will now take up the subject of runs once again.

Let us use the notation

$$\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle \tag{1}$$

to stand for the number of permutations of $\{1, 2, \dots, n\}$ that have exactly k “descents” $a_j > a_{j+1}$, thus exactly $k + 1$ ascending runs. These numbers $\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle$ arise in several contexts, and they are usually called *Eulerian numbers* since Euler discussed them in his famous book *Institutiones Calculi Differentialis* (St. Petersburg: 1755), 485–487, after having introduced them several years earlier in a technical paper [*Comment. Acad. Sci. Imp. Petrop.* **8** (1736), 147–158, §13]. They should not be confused with the *Euler numbers* E_n discussed in exercise 5.1.4–23. The angle brackets in $\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle$ remind us of the “ $>$ ” sign in the definition of a descent. Of course $\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle$ is also the number of permutations that have k “ascents” $a_j < a_{j+1}$.

We can use any given permutation of $\{1, \dots, n-1\}$ to form n new permutations, by inserting the element n in all possible places. If the original permutation has k descents, exactly $k+1$ of these new permutations will have k descents; the remaining $n-1-k$ will have $k+1$, since we increase the number of descents unless we place the element n at the end of an existing run. For example, the six permutations formed from 3 1 2 4 5 are

$$\begin{array}{lll} 6\ 3\ 1\ 2\ 4\ 5, & 3\ 6\ 1\ 2\ 4\ 5, & 3\ 1\ 6\ 2\ 4\ 5, \\ 3\ 1\ 2\ 6\ 4\ 5, & 3\ 1\ 2\ 4\ 6\ 5, & 3\ 1\ 2\ 4\ 5\ 6; \end{array}$$

all but the second and last of these have two descents instead of one. Therefore we have the recurrence relation

$$\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle = (k+1) \left\langle \begin{matrix} n-1 \\ k \end{matrix} \right\rangle + (n-k) \left\langle \begin{matrix} n-1 \\ k-1 \end{matrix} \right\rangle, \quad \text{integer } n > 0, \text{ integer } k. \quad (2)$$

By convention we set

$$\left\langle \begin{matrix} 0 \\ k \end{matrix} \right\rangle = \delta_{k0}, \quad (3)$$

saying that the null permutation has no descents. The reader may find it interesting to compare (2) with the recurrence relations for Stirling numbers in Eqs. 1.2.6–(46). Table 1 lists the Eulerian numbers for small n .

Several patterns can be observed in Table 1. By definition, we have

$$\left\langle \begin{matrix} n \\ 0 \end{matrix} \right\rangle + \left\langle \begin{matrix} n \\ 1 \end{matrix} \right\rangle + \dots + \left\langle \begin{matrix} n \\ n \end{matrix} \right\rangle = n!; \quad (4)$$

$$\left\langle \begin{matrix} n \\ 0 \end{matrix} \right\rangle = 1; \quad (5)$$

$$\left\langle \begin{matrix} n \\ n-1 \end{matrix} \right\rangle = 1, \quad \left\langle \begin{matrix} n \\ n \end{matrix} \right\rangle = 0, \quad \text{for } n \geq 1. \quad (6)$$

Eq. (6) follows from (5) because of a general rule of symmetry,

$$\left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle = \left\langle \begin{matrix} n \\ n-1-k \end{matrix} \right\rangle, \quad \text{for } n \geq 1, \quad (7)$$

which comes from the fact that each nonnull permutation $a_1 a_2 \dots a_n$ having k descents has $n-1-k$ ascents.

Another important property of the Eulerian numbers is the formula

$$\sum_k \left\langle \begin{matrix} n \\ k \end{matrix} \right\rangle \binom{m+k}{n} = m^n, \quad n \geq 0, \quad (8)$$

which was discovered by the Chinese mathematician Li Shan-Lan and published in 1867. [See J.-C. Martzloff, *A History of Chinese Mathematics* (Berlin: Springer, 1997), 346–348; special cases for $n \leq 5$ had already been known to Yoshisuke Matsunaga in Japan, who died in 1744.] Li Shan-Lan's identity follows from the properties of sorting: Consider the m^n sequences $a_1 a_2 \dots a_n$ such that $1 \leq a_i \leq m$. We can sort any such sequence into nondecreasing order in a stable manner, obtaining

$$a_{i_1} \leq a_{i_2} \leq \dots \leq a_{i_n} \quad (9)$$

Table 1
EULERIAN NUMBERS

n	$\langle n \rangle_0$	$\langle n \rangle_1$	$\langle n \rangle_2$	$\langle n \rangle_3$	$\langle n \rangle_4$	$\langle n \rangle_5$	$\langle n \rangle_6$	$\langle n \rangle_7$	$\langle n \rangle_8$
0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0
3	1	4	1	0	0	0	0	0	0
4	1	11	11	1	0	0	0	0	0
5	1	26	66	26	1	0	0	0	0
6	1	57	302	302	57	1	0	0	0
7	1	120	1191	2416	1191	120	1	0	0
8	1	247	4293	15619	15619	4293	247	1	0
9	1	502	14608	88234	156190	88234	14608	502	1

where $i_1 i_2 \dots i_n$ is a uniquely determined permutation of $\{1, 2, \dots, n\}$ such that $a_{i_j} = a_{i_{j+1}}$ implies $i_j < i_{j+1}$; in other words, $i_j > i_{j+1}$ implies that $a_{i_j} < a_{i_{j+1}}$. If the permutation $i_1 i_2 \dots i_n$ has k runs, we will show that the number of corresponding sequences $a_1 a_2 \dots a_n$ is $\binom{m+n-k}{n}$. This will prove (8) if we replace k by $n - k$ and use (7), because $\langle n \rangle_k$ permutations have $n - k$ runs.

For example, if $n = 9$ and $i_1 i_2 \dots i_n = 3 5 7 1 6 8 9 4 2$, we want to count the number of sequences $a_1 a_2 \dots a_n$ such that

$$1 \leq a_3 \leq a_5 \leq a_7 < a_1 \leq a_6 \leq a_8 \leq a_9 < a_4 < a_2 \leq m; \tag{10}$$

this is the number of sequences $b_1 b_2 \dots b_9$ such that

$$1 \leq b_1 < b_2 < b_3 < b_4 < b_5 < b_6 < b_7 < b_8 < b_9 \leq m + 5,$$

since we can let $b_1 = a_3, b_2 = a_5 + 1, b_3 = a_7 + 2, b_4 = a_1 + 2, b_5 = a_6 + 3$, etc. The number of choices of the b 's is simply the number of ways of choosing 9 things out of $m + 5$, namely $\binom{m+5}{9}$; a similar proof works for general n and k , and for any permutation $i_1 i_2 \dots i_n$ with k runs.

Since both sides of (8) are polynomials in m , we may replace m by any real number x , and we obtain an interesting representation of powers in terms of consecutive binomial coefficients:

$$x^n = \langle n \rangle_0 \binom{x}{n} + \langle n \rangle_1 \binom{x+1}{n} + \dots + \langle n \rangle_{n-1} \binom{x+n-1}{n}, \quad n \geq 1. \tag{11}$$

For example,

$$x^3 = \binom{x}{3} + 4 \binom{x+1}{3} + \binom{x+2}{3}.$$

This is the key property of Eulerian numbers that makes them useful in the study of discrete mathematics.

Setting $x = 1$ in (11) proves again that $\langle n \rangle_{n-1} = 1$, since the binomial coefficients vanish in all but the last term. Setting $x = 2$ yields

$$\langle n \rangle_{n-2} = \langle n \rangle_1 = 2^n - n - 1, \quad n \geq 1. \tag{12}$$

Setting $x = 3, 4, \dots$ shows that relation (11) completely defines the numbers $\langle \binom{n}{k} \rangle$, and leads to a formula originally given by Euler:

$$\begin{aligned} \langle \binom{n}{k} \rangle &= (k+1)^n - k^n \binom{n+1}{1} + (k-1)^n \binom{n+1}{2} - \dots + (-1)^k 1^n \binom{n+1}{k} \\ &= \sum_{j=0}^k (-1)^j \binom{n+1}{j} (k+1-j)^n, \quad n \geq 0, k \geq 0. \end{aligned} \quad (13)$$

Now let us study the generating function for runs. If we set

$$g_n(z) = \sum_k \langle \binom{n}{k-1} \rangle \frac{z^k}{n!}, \quad (14)$$

the coefficient of z^k is the probability that a random permutation of $\{1, 2, \dots, n\}$ has exactly k runs. Since k runs are just as likely as $n+1-k$, the average number of runs must be $\frac{1}{2}(n+1)$, hence $g'_n(1) = \frac{1}{2}(n+1)$. Exercise 2(b) shows that there is a simple formula for *all* the derivatives of $g_n(z)$ at the point $z = 1$:

$$g_n^{(m)}(1) = \left\{ \binom{n+1}{n+1-m} \right\} / \binom{n}{m}, \quad n \geq m. \quad (15)$$

Thus in particular the variance $g''_n(1) + g'_n(1) - g'_n(1)^2$ comes to $(n+1)/12$, for $n \geq 2$, indicating a rather stable distribution about the mean. (We found this same quantity in Eq. 3.3.2-(18), where it was called $\text{covar}(R'_1, R'_1)$.) Since $g_n(z)$ is a polynomial, we can use formula (15) to deduce the Taylor series expansions

$$g_n(z) = \frac{1}{n!} \sum_{k=0}^n (z-1)^{n-k} k! \left\{ \binom{n+1}{k+1} \right\} = \frac{1}{n!} \sum_{k=0}^n z^{k+1} (1-z)^{n-k} k! \left\{ \binom{n+1}{k+1} \right\}. \quad (16)$$

The second of these equations follows from the first, since

$$g_n(z) = z^{n+1} g_n(1/z), \quad n \geq 1, \quad (17)$$

by the symmetry condition (7). The Stirling number recurrence

$$\left\{ \binom{n+1}{k+1} \right\} = (k+1) \left\{ \binom{n}{k+1} \right\} + \left\{ \binom{n}{k} \right\}$$

gives two slightly simpler representations,

$$g_n(z) = \frac{1}{n!} \sum_{k=0}^n z(z-1)^{n-k} k! \left\{ \binom{n}{k} \right\} = \frac{1}{n!} \sum_{k=0}^n z^k (1-z)^{n-k} k! \left\{ \binom{n}{k} \right\}, \quad (18)$$

when $n \geq 1$. The super generating function

$$g(z, x) = \sum_{n \geq 0} \frac{g_n(z) x^n}{z} = \sum_{k, n \geq 0} \langle \binom{n}{k} \rangle \frac{z^k x^n}{n!} \quad (19)$$

is therefore equal to

$$\sum_{k,n \geq 0} \frac{((z-1)x)^n}{(z-1)^k} \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \frac{k!}{n!} = \sum_{k \geq 0} \left(\frac{e^{(z-1)x} - 1}{z-1} \right)^k = \frac{1-z}{e^{(z-1)x} - z}; \quad (20)$$

this is another relation discussed by Euler.

Further properties of the Eulerian numbers may be found in a survey paper by L. Carlitz [*Math. Magazine* **32** (1959), 247–260]. See also J. Riordan, *Introduction to Combinatorial Analysis* (New York: Wiley, 1958), 38–39, 214–219, 234–237; D. Foata and M. P. Schützenberger, *Lecture Notes in Math.* **138** (Berlin: Springer, 1970).

Let us now consider the length of runs; how long will a run be, on the average? We have already studied the expected number of runs having a given length, in Section 3.3.2; the average run length is approximately 2, in agreement with the fact that about $\frac{1}{2}(n+1)$ runs appear in a random permutation of length n . For applications to sorting algorithms, a slightly different viewpoint is useful; we will consider the length of the k th run of the permutation from left to right, for $k = 1, 2, \dots$

For example, how long is the first (leftmost) run of a random permutation $a_1 a_2 \dots a_n$? Its length is always ≥ 1 , and its length is ≥ 2 exactly one-half the time (namely when $a_1 < a_2$). Its length is ≥ 3 exactly one-sixth of the time (when $a_1 < a_2 < a_3$), and, in general, its length is $\geq m$ with probability $q_m = 1/m!$, for $1 \leq m \leq n$. The probability that its length is exactly equal to m is therefore

$$\begin{aligned} p_m &= q_m - q_{m+1} = 1/m! - 1/(m+1)!, & \text{for } 1 \leq m < n; \\ p_n &= 1/n!. \end{aligned} \quad (21)$$

The average length of the first run therefore equals

$$\begin{aligned} p_1 + 2p_2 + \dots + np_n &= (q_1 - q_2) + 2(q_2 - q_3) + \dots + (n-1)(q_{n-1} - q_n) + nq_n \\ &= q_1 + q_2 + \dots + q_n = \frac{1}{1!} + \frac{1}{2!} + \dots + \frac{1}{n!}. \end{aligned} \quad (22)$$

If we let $n \rightarrow \infty$, the limit is $e - 1 = 1.71828\dots$, and for finite n the value is $e - 1 - \delta_n$ where δ_n is quite small;

$$\delta_n = \frac{1}{(n+1)!} \left(1 + \frac{1}{n+2} + \frac{1}{(n+2)(n+3)} + \dots \right) \leq \frac{e-1}{(n+1)!}.$$

For practical purposes it is therefore convenient to study runs in a random *infinite* sequence of distinct numbers

$$a_1, a_2, a_3, \dots;$$

by “random” we mean in this case that each of the $n!$ possible relative orderings of the first n elements in the sequence is equally likely. The average length of the first run in a random infinite sequence is exactly $e - 1$.

By slightly sharpening our analysis of the first run, we can ascertain the average length of the k th run in a random sequence. Let q_{km} be the probability

that the first k runs have total length $\geq m$; then q_{km} is $1/m!$ times the number of permutations of $\{1, 2, \dots, m\}$ that have $\leq k$ runs,

$$q_{km} = \left(\binom{m}{0} + \dots + \binom{m}{k-1} \right) / m!. \quad (23)$$

The probability that the first k runs have total length m is $q_{km} - q_{k(m+1)}$. Therefore if L_k denotes the average length of the k th run, we find that

$$\begin{aligned} L_1 + \dots + L_k &= \text{average total length of first } k \text{ runs} \\ &= (q_{k1} - q_{k2}) + 2(q_{k2} - q_{k3}) + 3(q_{k3} - q_{k4}) + \dots \\ &= q_{k1} + q_{k2} + q_{k3} + \dots. \end{aligned}$$

Subtracting $L_1 + \dots + L_{k-1}$ and using the value of q_{km} in (23) yields the desired formula

$$L_k = \frac{1}{1!} \binom{1}{k-1} + \frac{1}{2!} \binom{2}{k-1} + \frac{1}{3!} \binom{3}{k-1} + \dots = \sum_{m \geq 1} \binom{m}{k-1} \frac{1}{m!}. \quad (24)$$

Since $\binom{0}{k-1} = 0$ except when $k = 1$, L_k turns out to be the coefficient of z^{k-1} in the generating function $g(z, 1) - 1$ (see Eq. (19)), so we have

$$L(z) = \sum_{k \geq 0} L_k z^k = \frac{z(1-z)}{e^{z-1} - z} - z. \quad (25)$$

From Euler's formula (13) we obtain a representation of L_k as a polynomial in e :

$$\begin{aligned} L_k &= \sum_{m \geq 0} \sum_{j=0}^k (-1)^{k-j} \binom{m+1}{k-j} \frac{j^m}{m!} \\ &= \sum_{j=0}^k (-1)^{k-j} \sum_{m \geq 0} \binom{m}{k-j} \frac{j^m}{m!} + \sum_{j=0}^{k-1} (-1)^{k-j} \sum_{m \geq 0} \binom{m}{k-j-1} \frac{j^m}{m!} \\ &= \sum_{j=0}^k \frac{(-1)^{k-j} j^{k-j}}{(k-j)!} \sum_{n \geq 0} \frac{j^n}{n!} + \sum_{j=0}^{k-1} \frac{(-1)^{k-j} j^{k-j-1}}{(k-j-1)!} \sum_{n \geq 0} \frac{j^n}{n!} \\ &= k \sum_{j=0}^k \frac{(-1)^{k-j} j^{k-j-1}}{(k-j)!} e^j. \end{aligned} \quad (26)$$

This formula for L_k was first obtained by B. J. Gassner [see *CACM* **10** (1967), 89–93]. In particular, we have

$$\begin{aligned} L_1 &= e - 1 &&= 1.71828\dots; \\ L_2 &= e^2 - 2e &&= 1.95249\dots; \\ L_3 &= e^3 - 3e^2 + \frac{3}{2}e &&= 1.99579\dots \end{aligned}$$

The second run is expected to be longer than the first, and the third run will be longer yet, on the average. This may seem surprising at first glance, but a moment's reflection shows that the first element of the second run tends to be

Table 2

AVERAGE LENGTH OF THE k TH RUN			
k	L_k	k	L_k
1	1.71828 18284 59045+	10	2.00000 00012 05997+
2	1.95249 24420 12560-	11	2.00000 00001 93672+
3	1.99579 13690 84285-	12	1.99999 99999 99909+
4	2.00003 88504 76806-	13	1.99999 99999 97022-
5	2.00005 75785 89716+	14	1.99999 99999 99719+
6	2.00000 50727 55710-	15	2.00000 00000 00019+
7	1.99999 96401 44022+	16	2.00000 00000 00006+
8	1.99999 98889 04744+	17	2.00000 00000 00000+
9	1.99999 99948 43434-	18	2.00000 00000 00000-

small (it caused the first run to terminate); hence there is a better chance for the second run to go on longer. The first element of the third run will tend to be even smaller than that of the second.

The numbers L_k are important in the theory of replacement-selection sorting (Section 5.4.1), so it is interesting to study their values in detail. Table 2 shows the first 18 values of L_k to 15 decimal places. Our discussion in the preceding paragraph might lead us to suspect at first that $L_{k+1} > L_k$, but in fact the values oscillate back and forth. Notice that L_k rapidly approaches the limiting value 2; it is quite remarkable to see these monic polynomials in the transcendental number e converging to the rational number 2 so quickly! The polynomials (26) are also somewhat interesting from the standpoint of numerical analysis, since they provide an excellent example of the loss of significant figures when nearly equal numbers are subtracted; using 19-digit floating point arithmetic, Gassner concluded incorrectly that $L_{12} > 2$, and John W. Wrench, Jr., has remarked that 42-digit floating point arithmetic gives L_{28} correct to only 29 significant digits.

The asymptotic behavior of L_k can be determined by using simple principles of complex variable theory. The denominator of (25) is zero only when $e^{z-1} = z$, namely when

$$e^{x-1} \cos y = x \quad \text{and} \quad e^{x-1} \sin y = y, \quad (27)$$

if we write $z = x + iy$. Figure 3 shows the superimposed graphs of these two equations, and we note that they intersect at the points $z = z_0, z_1, \bar{z}_1, z_2, \bar{z}_2, \dots$, where $z_0 = 1$,

$$z_1 = (3.08884 30156 13044-) + (7.46148 92856 54255-)i, \quad (28)$$

and the imaginary part $\Im(z_{k+1})$ is roughly equal to $\Im(z_k) + 2\pi$ for large k . Since

$$\lim_{z \rightarrow z_k} \left(\frac{1-z}{e^{z-1}-z} \right) (z-z_k) = -1, \quad \text{for } k > 0,$$

and since the limit is -2 for $k = 0$, the function

$$R_m(z) = L(z) + \frac{2z}{z-z_0} + \frac{z}{z-z_1} + \frac{z}{z-\bar{z}_1} + \frac{z}{z-z_2} + \frac{z}{z-\bar{z}_2} + \cdots + \frac{z}{z-z_m} + \frac{z}{z-\bar{z}_m}$$

has no singularities in the complex plane for $|z| < |z_{m+1}|$. Hence $R_m(z)$ has a power series expansion $\sum_k \rho_k z^k$ that converges absolutely when $|z| < |z_{m+1}|$; it follows that $\rho_k M^k \rightarrow 0$ as $k \rightarrow \infty$, where $M = |z_{m+1}| - \epsilon$. The coefficients of $L(z)$ are the coefficients of

$$\frac{2z}{1-z} + \frac{z/z_1}{1-z/z_1} + \frac{z/\bar{z}_1}{1-z/\bar{z}_1} + \cdots + \frac{z/z_m}{1-z/z_m} + \frac{z/\bar{z}_m}{1-z/\bar{z}_m} + R_m(z),$$

namely,

$$L_n = 2 + 2r_1^{-n} \cos n\theta_1 + 2r_2^{-n} \cos n\theta_2 + \cdots + 2r_m^{-n} \cos n\theta_m + O(r_{m+1}^{-n}), \quad (29)$$

if we let

$$z_k = r_k e^{i\theta_k}. \quad (30)$$

This shows the asymptotic behavior of L_n . We have

$$\begin{aligned} r_1 &= 8.07556\ 64528\ 89526-, & \theta_1 &= 1.17830\ 39784\ 74668+; \\ r_2 &= 14.35456\ 68997\ 62106-, & \theta_2 &= 1.31268\ 53883\ 87636+; \\ r_3 &= 20.62073\ 15381\ 80628-, & \theta_3 &= 1.37427\ 90757\ 91688-; \\ r_4 &= 26.88795\ 29424\ 54546-, & \theta_4 &= 1.41049\ 72786\ 51865-; \end{aligned} \quad (31)$$

so the main contribution to $L_n - 2$ is due to r_1 and θ_1 , and convergence of (29) is quite rapid. Further analysis [W. W. Hooker, *CACM* **12** (1969), 411–413] shows that $R_m(z) \rightarrow cz$ for some constant c as $m \rightarrow \infty$; hence the series $2 \sum_{k \geq 0} r_k^{-n} \cos n\theta_k$ actually converges to L_n when $n > 1$. (See also exercise 28.)

A more careful examination of probabilities can be carried out to determine the complete probability distribution for the length of the k th run and for the total length of the first k runs (see exercises 9, 10, 11). The sum $L_1 + \cdots + L_k$ turns out to be asymptotically $2k - \frac{1}{3} + O(8^{-k})$.

Let us conclude this section by considering the properties of runs when equal elements are allowed to appear in the permutations. The famous nineteenth-century American astronomer Simon Newcomb amused himself by playing a game of solitaire related to this question. He would deal a deck of cards into a pile, so long as the face values were in nondecreasing order; but whenever the next card to be dealt had a face value lower than its predecessor, he would start a new pile. He wanted to know the probability that a given number of piles would be formed after the entire deck had been dealt out in this manner.

Thus Simon Newcomb's problem was to find the probability distribution of runs in a random permutation of a certain multiset. The general answer is rather complicated (see exercise 12), although we have already seen how to solve the special case when all cards have a distinct face value. We will content ourselves here with a derivation of the *average* number of piles that appear in his game.

Suppose first that there are m different types of cards, each occurring exactly p times. An ordinary bridge deck, for example, has $m = 13$ and $p = 4$ if suits are disregarded. A remarkable symmetry applying to this case was discovered

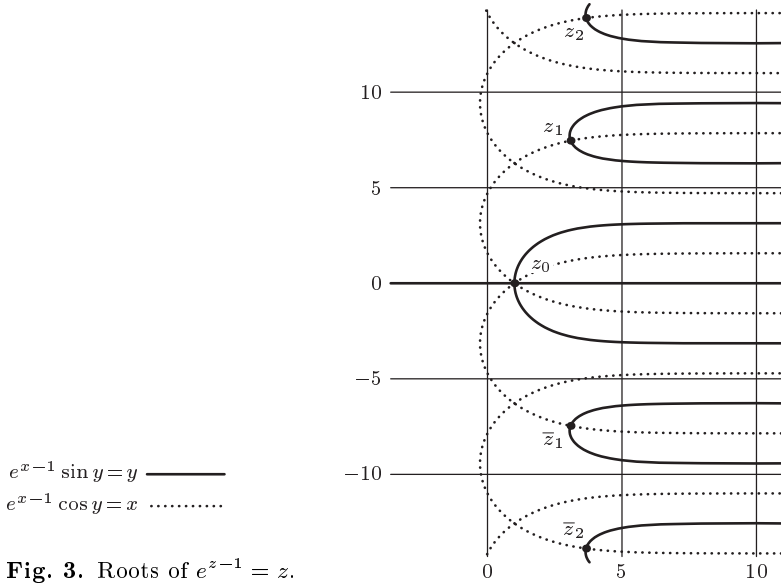


Fig. 3. Roots of $e^{z-1} = z$.

by P. A. MacMahon [*Combinatory Analysis 1* (Cambridge, 1915), 212–213]: The number of permutations with $k + 1$ runs is the same as the number with $mp - p - k + 1$ runs. When $p = 1$, this relation is Eq. (7), but for $p > 1$ it is quite surprising.

We can prove the symmetry by setting up a one-to-one correspondence between the permutations in such a way that each permutation with $k + 1$ runs corresponds to another having $mp - p - k + 1$ runs. The reader is urged to try discovering such a correspondence before reading further.

No very simple correspondence is evident; MacMahon’s proof was based on generating functions instead of a combinatorial construction. But Foata’s correspondence (Theorem 5.1.2B) provides a useful simplification, because it tells us that there is a one-to-one correspondence between multiset permutations with $k + 1$ runs and permutations whose two-line notation contains exactly k columns $\frac{y}{x}$ with $x < y$.

Suppose the given multiset is $\{p \cdot 1, p \cdot 2, \dots, p \cdot m\}$, and consider the permutation whose two-line notation is

$$\left(\begin{array}{cccccccc} 1 & \dots & 1 & 2 & \dots & 2 & \dots & m & \dots & m \\ x_{11} & \dots & x_{1p} & x_{21} & \dots & x_{2p} & \dots & x_{m1} & \dots & x_{mp} \end{array} \right). \quad (32)$$

We can associate this permutation with another one,

$$\left(\begin{array}{cccccccc} 1 & \dots & 1 & 2 & \dots & 2 & \dots & m & \dots & m \\ x'_{11} & \dots & x'_{1p} & x'_{m1} & \dots & x'_{mp} & \dots & x'_{21} & \dots & x'_{2p} \end{array} \right), \quad (33)$$

where $x' = m + 1 - x$. If (32) contains k columns of the form $\frac{y}{x}$ with $x < y$, then (33) contains $(m - 1)p - k$ such columns; for we need only consider the case $y > 1$, and $x < y$ is equivalent to $x' \geq m + 2 - y$. Now (32) corresponds to a permutation

with $k + 1$ runs, and (33) corresponds to a permutation with $mp - p - k + 1$ runs, and the transformation that takes (32) into (33) is reversible—it takes (33) back into (32). Therefore MacMahon's symmetry condition has been established. See exercise 14 for an example of this construction.

Because of the symmetry property, the average number of runs in a random permutation must be $\frac{1}{2}((k + 1) + (mp - p - k + 1)) = 1 + \frac{1}{2}p(m - 1)$. For example, the average number of piles resulting from Simon Newcomb's solitaire game using a standard deck will be 25 (so it doesn't appear to be a very exciting way to play solitaire).

We can actually determine the average number of runs in general, using a fairly simple argument, given *any* multiset $\{n_1 \cdot x_1, n_2 \cdot x_2, \dots, n_m \cdot x_m\}$ where the x 's are distinct. Let $n = n_1 + n_2 + \dots + n_m$, and imagine that all of the permutations $a_1 a_2 \dots a_n$ of this multiset have been written down; we will count how often a_i is greater than a_{i+1} , for each fixed value of i , $1 \leq i < n$. The number of times $a_i > a_{i+1}$ is just half of the number of times $a_i \neq a_{i+1}$; and it is not difficult to see that $a_i = a_{i+1} = x_j$ exactly $Nn_j(n_j - 1)/n(n - 1)$ times, where N is the total number of permutations. Hence $a_i = a_{i+1}$ exactly

$$\frac{N}{n(n - 1)}(n_1(n_1 - 1) + \dots + n_m(n_m - 1)) = \frac{N}{n(n - 1)}(n_1^2 + \dots + n_m^2 - n)$$

times, and $a_i > a_{i+1}$ exactly

$$\frac{N}{2n(n - 1)}(n^2 - (n_1^2 + \dots + n_m^2))$$

times. Summing over i and adding N , since a run ends at a_n in each permutation, we obtain the total number of runs among all N permutations:

$$N\left(\frac{n}{2} - \frac{1}{2n}(n_1^2 + \dots + n_m^2) + 1\right). \quad (34)$$

Dividing by N gives the desired average number of runs.

Since runs are important in the study of "order statistics," there is a fairly large literature dealing with them, including several other types of runs not considered here. For additional information, see the book *Combinatorial Chance* by F. N. David and D. E. Barton (London: Griffin, 1962), Chapter 10; and the survey paper by D. E. Barton and C. L. Mallows, *Annals of Math. Statistics* **36** (1965), 236–260.

EXERCISES

1. [M26] Derive Euler's formula (13).
- ▶ 2. [M22] (a) Extend the idea used in the text to prove (8), considering those sequences $a_1 a_2 \dots a_n$ that contain exactly q distinct elements, in order to prove the formula

$$\sum_k \langle n \rangle_k \binom{k}{n - q} = \left\{ \begin{matrix} n \\ q \end{matrix} \right\} q!, \quad \text{integer } q \geq 0.$$

(b) Use this identity to prove that

$$\sum_k \binom{n}{k} \binom{k+1}{m} = \binom{n+1}{n+1-m} (n-m)!, \quad \text{for } n \geq m.$$

3. [HM25] Evaluate the sum $\sum_k \binom{n}{k} (-1)^k$.
4. [M21] What is the value of $\sum_k (-1)^k \binom{n}{k} k!$ ($\binom{n-k}{m}$)?
5. [M20] Deduce the value of $\binom{p}{k} \pmod p$ when p is prime.
- ▶ 6. [M21] Mr. B. C. Dull noticed that, by Eqs. (4) and (13),

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

Carrying out the sum on k first, he found that $\sum_{k \geq 0} (-1)^{k-j} \binom{n+1}{k-j} = 0$ for all $j \geq 0$; hence $n! = 0$ for all $n \geq 0$. Did he make a mistake?

7. [HM40] Is the probability distribution of runs, given by (14), asymptotically normal? (See exercise 1.2.10–13.)
8. [M24] (P. A. MacMahon.) Show that the probability that the first run of a sufficiently long permutation has length l_1 , the second has length l_2 , ..., and the k th has length $\geq l_k$, is

$$\det \begin{pmatrix} 1/l_1! & 1/(l_1+l_2)! & 1/(l_1+l_2+l_3)! & \dots & 1/(l_1+l_2+l_3+\dots+l_k)! \\ 1 & 1/l_2! & 1/(l_2+l_3)! & \dots & 1/(l_2+l_3+\dots+l_k)! \\ 0 & 1 & 1/l_3! & \dots & 1/(l_3+\dots+l_k)! \\ \vdots & & & & \vdots \\ 0 & 0 & \dots & 1 & 1/l_k! \end{pmatrix}.$$

9. [M30] Let $h_k(z) = \sum p_{km} z^m$, where p_{km} is the probability that m is the total length of the first k runs in a random (infinite) sequence. Find “simple” expressions for $h_1(z)$, $h_2(z)$, and the super generating function $h(z, x) = \sum_k h_k(z) x^k$.
10. [HM30] Find the asymptotic behavior of the mean and variance of the distributions $h_k(z)$ in the preceding exercise, for large k .
11. [M40] Let $H_k(z) = \sum P_{km} z^m$, where P_{km} is the probability that m is the length of the k th run in a random (infinite) sequence. Express $H_1(z)$, $H_2(z)$, and the super generating function $H(z, x) = \sum_k H_k(z) x^k$ in terms of familiar functions.
12. [M33] (P. A. MacMahon.) Generalize Eq. (13) to permutations of a multiset, by proving that the number of permutations of $\{n_1 \cdot 1, n_2 \cdot 2, \dots, n_m \cdot m\}$ having exactly k runs is

$$\sum_{j=0}^k (-1)^j \binom{n+1}{j} \binom{n_1-1+k-j}{n_1} \binom{n_2-1+k-j}{n_2} \dots \binom{n_m-1+k-j}{n_m},$$

where $n = n_1 + n_2 + \dots + n_m$.

13. [05] If Simon Newcomb’s solitaire game is played with a standard bridge deck, ignoring face value but treating clubs < diamonds < hearts < spades, what is the average number of piles?
14. [M18] The permutation 3 1 1 1 2 3 1 4 2 3 3 4 2 2 4 4 has 5 runs; find the corresponding permutation with 9 runs, according to the text’s construction for MacMahon’s symmetry condition.

- **15.** [M21] (*Alternating runs.*) The classical nineteenth-century literature of combinatorial analysis did not treat the topic of runs in permutations, as we have considered them, but several authors studied “runs” that are alternately ascending and descending. Thus 53247618 was considered to have 4 runs: 532, 247, 761, and 18. (The first run would be ascending or descending, according as $a_1 < a_2$ or $a_1 > a_2$; thus $a_1 a_2 \dots a_n$ and $a_n \dots a_2 a_1$ and $(n+1-a_1)(n+1-a_2) \dots (n+1-a_n)$ all have the same number of alternating runs.) When n elements are being permuted, the maximum number of runs of this kind is $n-1$.

Find the average number of alternating runs in a random permutation of the set $\{1, 2, \dots, n\}$. [Hint: Consider the proof of (34).]

- 16.** [M30] Continuing the previous exercise, let $\chi_{k,n}$ be the number of permutations of $\{1, 2, \dots, n\}$ that have exactly k alternating runs. Find a recurrence relation, by means of which a table of $\chi_{k,n}$ can be computed; and find the corresponding recurrence relation for the generating function $G_n(z) = \sum_k \chi_{k,n} z^k / n!$. Use the latter recurrence to discover a simple formula for the *variance* of the number of alternating runs in a random permutation of $\{1, 2, \dots, n\}$.

- 17.** [M25] Among all 2^n sequences $a_1 a_2 \dots a_n$, where each a_j is either 0 or 1, how many have exactly k runs (that is, $k-1$ occurrences of $a_j > a_{j+1}$)?

- 18.** [M28] Among all $n!$ sequences $b_1 b_2 \dots b_n$ such that each b_j is an integer in the range $0 \leq b_j \leq n-j$, how many have (a) exactly k descents (that is, k occurrences of $b_j > b_{j+1}$)? (b) exactly k distinct elements?

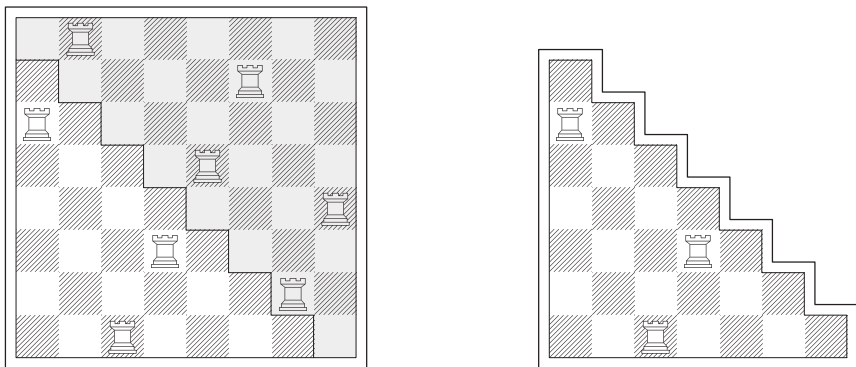


Fig. 4. Nonattacking rooks on a chessboard, with $k = 3$ rooks below the main diagonal.

- **19.** [M26] (I. Kaplansky and J. Riordan, 1946.) (a) In how many ways can n nonattacking rooks — no two in the same row or column — be placed on an $n \times n$ chessboard, so that exactly k lie below the main diagonal? (b) In how many ways can k nonattacking rooks be placed below the main diagonal of an $n \times n$ chessboard?

For example, Fig. 4 shows one of the 15619 ways to put eight nonattacking rooks on a standard chessboard with exactly three rooks in the unshaded portion below the main diagonal, together with one of the 1050 ways to put three nonattacking rooks on a triangular board.

- **20.** [M21] A permutation is said to require k *readings* if we must scan it k times from left to right in order to read off its elements in nondecreasing order. For example, the

permutation 491825367 requires four readings: On the first we obtain 1, 2, 3; on the second we get 4, 5, 6, 7; then 8; then 9. Find a connection between runs and readings.

21. [M22] If the permutation $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$ has k runs and requires j readings, in the sense of exercise 20, what can be said about $a_n \dots a_2 a_1$?

22. [M26] (L. Carlitz, D. P. Roselle, and R. A. Scoville.) Show that there is no permutation of $\{1, 2, \dots, n\}$ with $n + 1 - r$ runs, and requiring s readings, if $rs < n$; but such permutations do exist if $n \geq n + 1 - r \geq s \geq 1$ and $rs \geq n$.

23. [HM42] (Walter Weissblum.) The “long runs” of a permutation $a_1 a_2 \dots a_n$ are obtained by placing vertical lines just before a segment fails to be monotonic; long runs are either increasing or decreasing, depending on the order of their first two elements, so the length of each long run (except possibly the last) is ≥ 2 . For example, $7\ 5\ |\ 6\ 2\ |\ 3\ 8\ 9\ |\ 1\ 4$ has four long runs. Find the average length of the first two long runs of an infinite permutation, and prove that the limiting long-run length is

$$(1 + \cot \frac{1}{2}) / (3 - \cot \frac{1}{2}) \approx 2.4202.$$

24. [M30] What is the average number of runs in sequences generated as in exercise 5.1.1–18, as a function of p ?

25. [M25] Let U_1, \dots, U_n be independent uniform random numbers in $[0..1)$. What is the probability that $\lfloor U_1 + \dots + U_n \rfloor = k$?

26. [M20] Let ϑ be the operation $z \frac{d}{dz}$, which multiplies the coefficient of z^n in a generating function by n . Show that the result of applying ϑ to $1/(1 - z)$ repeatedly, m times, can be expressed in terms of Eulerian numbers.

▶ 27. [M21] An *increasing forest* is an oriented forest in which the nodes are labeled $\{1, 2, \dots, n\}$ in such a way that parents have smaller numbers than their children. Show that $\langle n \rangle_k$ is the number of n -node increasing forests with $k + 1$ leaves.

28. [HM35] Find the asymptotic value of the numbers z_m in Fig. 3 as $m \rightarrow \infty$, and prove that $\sum_{m=1}^{\infty} (z_m^{-1} + \bar{z}_m^{-1}) = e - 5/2$.

▶ 29. [M30] The permutation $a_1 \dots a_n$ has a “peak” at a_j if $1 < j < n$ and $a_{j-1} < a_j > a_{j+1}$. Let s_{nk} be the number of permutations with exactly k peaks, and let t_{nk} be the number with k peaks and k descents. Prove that (a) $s_{nk} = \frac{1}{2} \langle n \rangle_{2k} + \langle n \rangle_{2k+1} + \frac{1}{2} \langle n \rangle_{2k+2}$ (see exercise 16); (b) $s_{nk} = 2^{n-1-2k} t_{nk}$; (c) $\sum_k \langle n \rangle_k x^k = \sum_k t_{nk} x^k (1 + x)^{n-1-2k}$.

***5.1.4. Tableaux and Involution**

To complete our survey of the combinatorial properties of permutations, we will discuss some remarkable relations that connect permutations with arrays of integers called tableaux. A *Young tableau of shape* (n_1, n_2, \dots, n_m) , where $n_1 \geq n_2 \geq \dots \geq n_m > 0$, is an arrangement of $n_1 + n_2 + \dots + n_m$ distinct integers in an array of left-justified rows, with n_i elements in row i , such that the entries of each row are in increasing order from left to right, and the entries of each column are increasing from top to bottom. For example,

1	2	5	9	10	15
3	6	7	13		
4	8	12	14		
11					

(1)

is a Young tableau of shape $(6, 4, 4, 1)$. Such arrangements were introduced by Alfred Young as an aid to the study of matrix representations of permutations [see *Proc. London Math. Soc.* (2) **28** (1928), 255–292; Bruce E. Sagan, *The Symmetric Group* (Pacific Grove, Calif.: Wadsworth & Brooks/Cole, 1991)]. For simplicity, we will simply say “tableau” instead of “Young tableau.”

An *involution* is a permutation that is its own inverse. For example, there are ten involutions of $\{1, 2, 3, 4\}$:

$$\begin{array}{ccccc} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 3 & 4 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 2 & 3 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 4 & 3 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 1 & 2 \end{pmatrix} & \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \end{array} \quad (2)$$

The term “involution” originated in classical geometry problems; involutions in the general sense considered here were first studied by H. A. Rothe when he introduced the concept of inverses (see Section 5.1.1).

It may appear strange that we should be discussing both tableaux and involutions at the same time, but there is an extraordinary connection between these two apparently unrelated concepts: *The number of involutions of $\{1, 2, \dots, n\}$ is the same as the number of tableaux that can be formed from the elements $\{1, 2, \dots, n\}$.* For example, exactly ten tableaux can be formed from $\{1, 2, 3, 4\}$, namely,

$$\begin{array}{ccccc} \begin{array}{|c|c|c|c|} \hline 1 & 2 & 3 & 4 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline 1 & 3 & 4 \\ \hline 2 \\ \hline \end{array} & \begin{array}{|c|c|} \hline 1 & 4 \\ \hline 2 \\ \hline 3 \\ \hline \end{array} & \begin{array}{|c|c|} \hline 1 & 3 \\ \hline 2 \\ \hline 4 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline 1 & 2 & 4 \\ \hline 3 \\ \hline \end{array} \\ \begin{array}{|c|c|} \hline 1 & 2 \\ \hline 3 \\ \hline 4 \\ \hline \end{array} & \begin{array}{|c|c|c|} \hline 1 & 2 & 3 \\ \hline 4 \\ \hline \end{array} & \begin{array}{|c|c|} \hline 1 & 3 \\ \hline 2 & 4 \\ \hline \end{array} & \begin{array}{|c|c|} \hline 1 & 2 \\ \hline 3 & 4 \\ \hline \end{array} & \begin{array}{|c|} \hline 1 \\ \hline 2 \\ \hline 3 \\ \hline 4 \\ \hline \end{array} \end{array} \quad (3)$$

corresponding respectively to the ten involutions (2).

This connection between involutions and tableaux is by no means obvious, and there is probably no very simple way to prove it. The proof we will discuss involves an interesting tableau-construction algorithm that has several other surprising properties. It is based on a special procedure that inserts new elements into a tableau.

For example, suppose that we want to insert the element 8 into the tableau

$$\begin{array}{|c|c|c|c|c|c|} \hline 1 & 3 & 5 & 9 & 12 & 16 \\ \hline 2 & 6 & 10 & 15 & & \\ \hline 4 & 13 & 14 & & & \\ \hline 11 & & & & & \\ \hline 17 & & & & & \\ \hline \end{array} \quad (4)$$

The method we will use starts by placing the 8 into row 1, in the spot previously occupied by 9, since 9 is the least element greater than 8 in that row. Element 9 is “bumped down” into row 2, where it displaces the 10. The 10 then “bumps” the 13 from row 3 to row 4; and since row 4 contains no element greater than 13, the process terminates by inserting 13 at the right end of row 4. Thus, tableau (4) has been transformed into

1	3	5	8	12	16
2	6	9	15		
4	10	14			
11	13				
17					

(5)

A precise description of this process, together with a proof that it always preserves the tableau properties, appears in Algorithm I.

Algorithm I (*Insertion into a tableau*). Let $P = (P_{ij})$ be a tableau of positive integers, and let x be a positive integer not in P . This algorithm transforms P into another tableau that contains x in addition to its original elements. The new tableau has the same shape as the old, except for the addition of a new position in row s , column t , where s and t are quantities determined by the algorithm.

(Parenthesized remarks in this algorithm serve to prove its validity, since it is easy to verify inductively that the remarks are valid and that the array P remains a tableau throughout the process. For convenience we will assume that the tableau has been bordered by zeros at the top and left and with ∞ 's to the right and below, so that P_{ij} is defined for all $i, j \geq 0$. If we define the relation

$$a \lesssim b \quad \text{if and only if} \quad a < b \quad \text{or} \quad a = b = 0 \quad \text{or} \quad a = b = \infty, \quad (6)$$

the tableau inequalities can be expressed in the convenient form

$$\begin{aligned} P_{ij} = 0 & \quad \text{if and only if} \quad i = 0 \quad \text{or} \quad j = 0; \\ P_{ij} \lesssim P_{i(j+1)} & \quad \text{and} \quad P_{ij} \lesssim P_{(i+1)j}, \quad \text{for all } i, j \geq 0. \end{aligned} \quad (7)$$

The statement “ $x \notin P$ ” means that either $x = \infty$ or $x \neq P_{ij}$ for all $i, j \geq 0$.)

11. [Input x .] Set $i \leftarrow 1$, set $x_1 \leftarrow x$, and set j to the smallest value such that $P_{1j} = \infty$.
12. [Find x_{i+1} .] (At this point $P_{(i-1)j} < x_i < P_{ij}$ and $x_i \notin P$.) If $x_i < P_{i(j-1)}$, decrease j by 1 and repeat this step. Otherwise set $x_{i+1} \leftarrow P_{ij}$ and set $r_i \leftarrow j$.
13. [Replace by x_i .] (Now $P_{i(j-1)} < x_i < x_{i+1} = P_{ij} \lesssim P_{i(j+1)}$, $P_{(i-1)j} < x_i < x_{i+1} = P_{ij} \lesssim P_{(i+1)j}$, and $r_i = j$.) Set $P_{ij} \leftarrow x_i$.
14. [Is $x_{i+1} = \infty$?] (Now $P_{i(j-1)} < P_{ij} = x_i < x_{i+1} \lesssim P_{i(j+1)}$, $P_{(i-1)j} < P_{ij} = x_i < x_{i+1} \lesssim P_{(i+1)j}$, $r_i = j$, and $x_{i+1} \notin P$.) If $x_{i+1} \neq \infty$, increase i by 1 and return to step 12.

I5. [Determine s, t .] Set $s \leftarrow i, t \leftarrow j$, and terminate the algorithm. (At this point the conditions

$$P_{st} \neq \infty \quad \text{and} \quad P_{(s+1)t} = P_{s(t+1)} = \infty \quad (8)$$

are satisfied.) **■**

Algorithm I defines a “bumping sequence”

$$x = x_1 < x_2 < \cdots < x_s < x_{s+1} = \infty, \quad (9)$$

as well as an auxiliary sequence of column indices

$$r_1 \geq r_2 \geq \cdots \geq r_s = t; \quad (10)$$

element P_{ir_i} has been changed from x_{i+1} to x_i , for $1 \leq i \leq s$. For example, when we inserted 8 into (4), the bumping sequence was 8, 9, 10, 13, ∞ , and the auxiliary sequence was 4, 3, 2, 2. We could have reformulated the algorithm so that it used much less temporary storage; only the current values of i, j, x_i , and x_{i+1} need to be remembered. But sequences (9) and (10) have been introduced so that we can prove interesting things about the algorithm.

The key fact we will use about Algorithm I is that it can be run backwards: Given the values of s and t determined in step I5, we can transform P back into its original form again, determining and removing the element x that was inserted. For example, consider (5) and suppose we are told that element 13 is in the position that used to be blank. Then 13 must have been bumped down from row 3 by the 10, since 10 is the greatest element less than 13 in that row; similarly the 10 must have been bumped from row 2 by the 9, and the 9 must have been bumped from row 1 by the 8. Thus we can go from (5) back to (4). The following algorithm specifies this process in detail:

Algorithm D (*Deletion from a tableau*). Given a tableau P and positive integers s, t satisfying (8), this algorithm transforms P into another tableau, having almost the same shape, but with ∞ in column t of row s . An element x , determined by the algorithm, is deleted from P .

(As in Algorithm I, parenthesized assertions are included here to facilitate a proof that P remains a tableau throughout the process.)

D1. [Input s, t .] Set $j \leftarrow t, i \leftarrow s, x_{s+1} \leftarrow \infty$.

D2. [Find x_i .] (At this point $P_{ij} < x_{i+1} \lesssim P_{(i+1)j}$ and $x_{i+1} \notin P$.) If $P_{i(j+1)} < x_{i+1}$, increase j by 1 and repeat this step. Otherwise set $x_i \leftarrow P_{ij}$ and $r_i \leftarrow j$.

D3. [Replace by x_{i+1} .] (Now $P_{i(j-1)} < P_{ij} = x_i < x_{i+1} \lesssim P_{i(j+1)}, P_{(i-1)j} < P_{ij} = x_i < x_{i+1} \lesssim P_{(i+1)j}$, and $r_i = j$.) Set $P_{ij} \leftarrow x_{i+1}$.

D4. [Is $i = 1$?] (Now $P_{i(j-1)} < x_i < x_{i+1} = P_{ij} \lesssim P_{i(j+1)}, P_{(i-1)j} < x_i < x_{i+1} = P_{ij} \lesssim P_{(i+1)j}$, and $r_i = j$.) If $i > 1$, decrease i by 1 and return to step D2.

D5. [Determine x .] Set $x \leftarrow x_1$; the algorithm terminates. (Now $0 < x < \infty$.) **■**

The parenthesized assertions appearing in Algorithms I and D are not only a useful way to prove that the algorithms preserve the tableau structure; they also serve to verify that *Algorithms I and D are perfect inverses of each other*. If we perform Algorithm I first, given some tableau P and some positive integer $x \notin P$, it will insert x and determine positive integers s, t satisfying (8); Algorithm D applied to the result will recompute x and will restore P . Conversely, if we perform Algorithm D first, given some tableau P and some positive integers s, t satisfying (8), it will modify P , deleting some positive integer x ; Algorithm I applied to the result will recompute s, t and will restore P . The reason is that the parenthesized assertions of steps I3 and D4 are identical, as are the assertions of steps I4 and D3, and these assertions characterize the value of j uniquely. Hence the auxiliary sequences (9), (10) are the same in each case.

Now we are ready to prove a basic property of tableaux:

Theorem A. *There is a one-to-one correspondence between the set of all permutations of $\{1, 2, \dots, n\}$ and the set of ordered pairs (P, Q) of tableaux formed from $\{1, 2, \dots, n\}$, where P and Q have the same shape.*

(An example of this theorem appears within the proof that follows.)

Proof. It is convenient to prove a slightly more general result. Given any two-line array

$$\begin{pmatrix} q_1 & q_2 & \dots & q_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad \begin{array}{l} q_1 < q_2 < \dots < q_n, \\ p_1, p_2, \dots, p_n \text{ distinct,} \end{array} \quad (11)$$

we will construct two corresponding tableaux P and Q , where the elements of P are $\{p_1, \dots, p_n\}$ and the elements of Q are $\{q_1, \dots, q_n\}$ and the shape of P is the shape of Q .

Let P and Q be empty initially. Then, for $i = 1, 2, \dots, n$ (in this order), do the following operation: Insert p_i into tableau P using Algorithm I; then set $Q_{st} \leftarrow q_i$, where s and t specify the newly filled position of P .

For example, if the given permutation is $\begin{pmatrix} 1 & 3 & 5 & 6 & 8 \\ 7 & 2 & 9 & 5 & 3 \end{pmatrix}$, we obtain

	P	Q	
Insert 7:	7	1	
Insert 2:	2 7	1 3	
Insert 9:	2 9 7	1 5 3	(12)
Insert 5:	2 5 7 9	1 5 3 6	
Insert 3:	2 3 5 9 7	1 5 3 6 8	

so the tableaux (P, Q) corresponding to $\left(\begin{smallmatrix} 1 & 3 & 5 & 6 & 8 \\ 7 & 2 & 9 & 5 & 3 \end{smallmatrix}\right)$ are

$$P = \begin{array}{|c|c|} \hline 2 & 3 \\ \hline 5 & 9 \\ \hline 7 & \\ \hline \end{array}, \quad Q = \begin{array}{|c|c|} \hline 1 & 5 \\ \hline 3 & 6 \\ \hline 8 & \\ \hline \end{array}. \quad (13)$$

It is clear from this construction that P and Q always have the same shape; furthermore, since we always add elements on the periphery of Q , in increasing order, Q is a tableau.

Conversely, given two equal-shape tableaux P and Q , we can find the corresponding two-line array (11) as follows. Let the elements of Q be

$$q_1 < q_2 < \cdots < q_n.$$

For $i = n, \dots, 2, 1$ (in this order), let p_i be the element x that is removed when Algorithm D is applied to P , using the values s and t such that $Q_{st} = q_i$.

For example, this construction will start with (13) and will successively undo the calculation (12) until P is empty, and $\left(\begin{smallmatrix} 1 & 3 & 5 & 6 & 8 \\ 7 & 2 & 9 & 5 & 3 \end{smallmatrix}\right)$ is obtained.

Since Algorithms I and D are inverses of each other, the two constructions we have described are inverses of each other, and the one-to-one correspondence has been established. ■

The correspondence defined in the proof of Theorem A has many startling properties, and we will now proceed to derive some of them. The reader is urged to work out the example in exercise 1, in order to become familiar with the construction, before proceeding further.

Once an element has been bumped from row 1 to row 2, it doesn't affect row 1 any longer; furthermore rows 2, 3, \dots are built up from the sequence of bumped elements in exactly the same way as rows 1, 2, \dots are built up from the original permutation. These facts suggest that we can look at the construction of Theorem A in another way, concentrating only on the first rows of P and Q . For example, the permutation $\left(\begin{smallmatrix} 1 & 3 & 5 & 6 & 8 \\ 7 & 2 & 9 & 5 & 3 \end{smallmatrix}\right)$ causes the following action in row 1, according to (12):

- 1: Insert 7, set $Q_{11} \leftarrow 1$.
 - 3: Insert 2, bump 7.
 - 5: Insert 9, set $Q_{12} \leftarrow 5$.
 - 6: Insert 5, bump 9.
 - 8: Insert 3, bump 5.
- (14)

Thus the first row of P is 2 3, and the first row of Q is 1 5. Furthermore, the remaining rows of P and Q are the tableaux corresponding to the "bumped" two-line array

$$\left(\begin{array}{ccc} 3 & 6 & 8 \\ 7 & 9 & 5 \end{array}\right). \quad (15)$$

In order to study the behavior of the construction on row 1, we can consider the elements that go into a given column of this row. Let us say that (q_i, p_i) is

in class t with respect to the two-line array

$$\begin{pmatrix} q_1 & q_2 & \cdots & q_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad \begin{array}{l} q_1 < q_2 < \cdots < q_n, \\ p_1, p_2, \dots, p_n \text{ distinct,} \end{array} \quad (16)$$

if $p_i = P_{1t}$ after Algorithm I has been applied successively to p_1, p_2, \dots, p_i , starting with an empty tableau P . (Remember that Algorithm I always inserts the given element into row 1.)

It is easy to see that (q_i, p_i) is in class 1 if and only if p_i has $i - 1$ inversions, that is, if and only if $p_i = \min\{p_1, p_2, \dots, p_i\}$ is a “left-to-right minimum.” If we cross out the columns of class 1 in (16), we obtain another two-line array

$$\begin{pmatrix} q'_1 & q'_2 & \cdots & q'_m \\ p'_1 & p'_2 & \cdots & p'_m \end{pmatrix} \quad (17)$$

such that (q, p) is in class t with respect to (17) if and only if it is in class $t + 1$ with respect to (16). The operation of going from (16) to (17) represents removing the leftmost position of row 1. This gives us a systematic way to determine the classes. For example in $\begin{pmatrix} 1 & 3 & 5 & 6 & 8 \\ 7 & 2 & 9 & 5 & 3 \end{pmatrix}$ the elements that are left-to-right minima are 7 and 2, so class 1 is $\{(1, 7), (3, 2)\}$; in the remaining array $\begin{pmatrix} 5 & 6 & 8 \\ 9 & 5 & 3 \end{pmatrix}$ all elements are minima, so class 2 is $\{(5, 9), (6, 5), (8, 3)\}$. In the “bumped” array (15), class 1 is $\{(3, 7), (8, 5)\}$ and class 2 is $\{(6, 9)\}$.

For any fixed value of t , the elements of class t can be labeled

$$(q_{i_1}, p_{i_1}), \dots, (q_{i_k}, p_{i_k})$$

in such a way that

$$\begin{array}{l} q_{i_1} < q_{i_2} < \cdots < q_{i_k}, \\ p_{i_1} > p_{i_2} > \cdots > p_{i_k}, \end{array} \quad (18)$$

since the tableau position P_{1t} takes on the decreasing sequence of values p_{i_1}, \dots, p_{i_k} as the insertion algorithm proceeds. At the end of the construction we have

$$P_{1t} = p_{i_k}, \quad Q_{1t} = q_{i_1}; \quad (19)$$

and the “bumped” two-line array that defines rows 2, 3, \dots of P and Q contains the columns

$$\begin{pmatrix} q_{i_2} & q_{i_3} & \cdots & q_{i_k} \\ p_{i_1} & p_{i_2} & \cdots & p_{i_{k-1}} \end{pmatrix} \quad (20)$$

plus other columns formed in a similar way from the other classes.

These observations lead to a simple method for calculating P and Q by hand (see exercise 3), and they also provide us with the means to prove a rather unexpected result:

Theorem B. *If the permutation*

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}$$

corresponds to tableaux (P, Q) in the construction of Theorem A, then the inverse permutation corresponds to (Q, P) .

This fact is quite startling, since P and Q are formed by such completely different methods in Theorem A, and since the inverse of a permutation is obtained by juggling the columns of the two-line array rather capriciously.

Proof. Suppose that we have a two-line array (16); its columns are essentially independent and can be rearranged. Interchanging the lines and sorting the columns so that the new top line is in increasing order gives the “inverse” array

$$\begin{aligned} \begin{pmatrix} q_1 & q_2 & \cdots & q_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}^{-1} &= \begin{pmatrix} p_1 & p_2 & \cdots & p_n \\ q_1 & q_2 & \cdots & q_n \end{pmatrix} \\ &= \begin{pmatrix} p'_1 & p'_2 & \cdots & p'_n \\ q'_1 & q'_2 & \cdots & q'_n \end{pmatrix}, \quad p'_1 < p'_2 < \cdots < p'_n; \\ & \quad q'_1, q'_2, \dots, q'_n \text{ distinct.} \end{aligned} \quad (21)$$

We will show that this operation corresponds to interchanging P and Q in the construction of Theorem A.

Exercise 2 reformulates our remarks about class determination so that the class of (q_i, p_i) doesn't depend on the fact that q_1, q_2, \dots, q_n are in ascending order. Since the resulting condition is symmetrical in the q 's and the p 's, the operation (21) does not destroy the class structure; if (q, p) is in class t with respect to (16), then (p, q) is in class t with respect to (21). If we therefore arrange the elements of the latter class t as

$$\begin{aligned} p_{i_k} < \cdots < p_{i_2} < p_{i_1}, \\ q_{i_k} > \cdots > q_{i_2} > q_{i_1}, \end{aligned} \quad (22)$$

by analogy with (18), we have

$$P_{1t} = q_{i_1}, \quad Q_{1t} = p_{i_k} \quad (23)$$

as in (19), and the columns

$$\begin{pmatrix} p_{i_{k-1}} & \cdots & p_{i_2} & p_{i_1} \\ q_{i_k} & \cdots & q_{i_3} & q_{i_2} \end{pmatrix} \quad (24)$$

go into the “bumped” array as in (20). Hence the first rows of P and Q are interchanged. Furthermore the “bumped” two-line array for (21) is the inverse of the “bumped” two-line array for (16), so the proof is completed by induction on the number of rows in the tableaux. ■

Corollary B. *The number of tableaux that can be formed from $\{1, 2, \dots, n\}$ is the number of involutions on $\{1, 2, \dots, n\}$.*

Proof. If π is an involution corresponding to (P, Q) , then $\pi = \pi^{-1}$ corresponds to (Q, P) ; hence $P = Q$. Conversely, if π is any permutation corresponding to (P, P) , then π^{-1} also corresponds to (P, P) ; hence $\pi = \pi^{-1}$. So there is a one-to-one correspondence between involutions π and tableaux P . ■

It is clear that the upper-left corner element of a tableau is always the smallest. This suggests a possible way to sort a set of numbers: First we can put the numbers into a tableau, by using Algorithm I repeatedly; this brings the smallest element to the corner. Then we delete the smallest element, rearranging

the remaining elements so that they form another tableau; then we delete the new smallest element; and so on.

Let us therefore consider what happens when we delete the corner element from the tableau

1	3	5	7	11	15	(25)
2	6	8	14			
4	9	13				
10	12					
16						

If the 1 is removed, the 2 must come to take its place. Then we can move the 4 up to where the 2 was, but we can't move the 10 to the position of the 4; the 9 can be moved instead, then the 12 in place of the 9. In general, we are led to the following procedure.

Algorithm S (*Delete corner element*). Given a tableau P , this algorithm deletes the upper left corner element of P and moves other elements so that the tableau properties are preserved. The notational conventions of Algorithms I and D are used.

- S1. [Initialize.] Set $r \leftarrow 1$, $s \leftarrow 1$.
- S2. [Done?] If $P_{rs} = \infty$, the process is complete.
- S3. [Compare.] If $P_{(r+1)s} \lesssim P_{r(s+1)}$, go to step S5. (We examine the elements just below and to the right of the vacant cell, and we will move the smaller of the two.)
- S4. [Shift left.] Set $P_{rs} \leftarrow P_{r(s+1)}$, $s \leftarrow s + 1$, and return to S3.
- S5. [Shift up.] Set $P_{rs} \leftarrow P_{(r+1)s}$, $r \leftarrow r + 1$, and return to S2. ■

It is easy to prove that P is still a tableau after Algorithm S has deleted its corner element (see exercise 10). So if we repeat Algorithm S until P is empty, we can read out its elements in increasing order. Unfortunately this doesn't turn out to be as efficient a sorting algorithm as other methods we will see; its minimum running time is proportional to $n^{1.5}$, but similar algorithms that use trees instead of tableau structures have an execution time on the order of $n \log n$.

In spite of the fact that Algorithm S doesn't lead to a superbly efficient sorting algorithm, it has some very interesting properties.

Theorem C (M. P. Schützenberger). *If P is the tableau formed by the construction of Theorem A from the permutation $a_1 a_2 \dots a_n$, and if*

$$a_i = \min\{a_1, a_2, \dots, a_n\},$$

then Algorithm S changes P to the tableau corresponding to $a_1 \dots a_{i-1} a_{i+1} \dots a_n$.

Proof. See exercise 13. ■

After we apply Algorithm S to a tableau, let us put the deleted element into the newly vacated place P_{rs} , but in *italic type* to indicate that it isn't really part of the tableau. For example, after applying this procedure to the tableau (25) we would have

2	3	5	7	11	15
4	6	8	14		
9	12	13			
10	<i>1</i>				
16					

and two more applications yield

4	5	7	11	15	<i>2</i>
6	8	13	14		
9	12	<i>3</i>			
10	<i>1</i>				
16					

Continuing until all elements are removed gives

16	14	13	12	10	<i>2</i>
15	9	6	4		
11	5	<i>3</i>			
8	<i>1</i>				
7					

(26)

which has the same shape as the original tableau (25). This configuration may be called a *dual tableau*, since it is like a tableau except that the “dual order” has been used (reversing the roles of $<$ and $>$). Let us denote the dual tableau formed from P in this way by the symbol P^S .

From P^S we can determine P uniquely; in fact, we can obtain the original tableau P from P^S , by applying exactly the same algorithm — but reversing the order and the roles of italic and regular type, since P^S is a dual tableau. For example, two steps of the algorithm applied to (26) give

14	13	12	10	<i>2</i>	15
11	9	6	4		
8	5	<i>3</i>			
7	<i>1</i>				
16					

and eventually (25) will be reproduced again! This remarkable fact is one of the consequences of our next theorem.

Theorem D (C. Schensted, M. P. Schützenberger). *Let*

$$\begin{pmatrix} q_1 & q_2 & \cdots & q_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix} \quad (27)$$

be the two-line array corresponding to the tableaux (P, Q) .

a) *Using dual (reverse) order on the q 's, but not on the p 's, the two-line array*

$$\begin{pmatrix} q_n & \cdots & q_2 & q_1 \\ p_n & \cdots & p_2 & p_1 \end{pmatrix} \quad (28)$$

corresponds to $(P^T, (Q^S)^T)$.

As usual, “ T ” denotes the operation of transposing rows and columns; P^T is a tableau, while $(Q^S)^T$ is a dual tableau, since the order of the q 's is reversed.

b) *Using dual order on the p 's, but not on the q 's, the two-line array (27) corresponds to $((P^S)^T, Q^T)$.*

c) *Using dual order on both the p 's and the q 's, the two-line array (28) corresponds to (P^S, Q^S) .*

Proof. No simple proof of this theorem is known. The fact that case (a) corresponds to (P^T, X) for some dual tableau X is proved in exercise 5; hence by Theorem B, case (b) corresponds to (Y, Q^T) for some dual tableau Y , and Y must have the shape of P^T .

Let $p_i = \min\{p_1, \dots, p_n\}$; since p_i is the “largest” element in the dual order, it appears on the periphery of Y , and it doesn't bump any elements in the construction of Theorem A. Thus, if we successively insert $p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n$ using the dual order, we get $Y - \{p_i\}$, that is, Y with p_i removed. By Theorem C if we successively insert $p_1, \dots, p_{i-1}, p_{i+1}, \dots, p_n$ using the normal order, we get the tableau $d(P)$ obtained by applying Algorithm S to P . By induction on n , $Y - \{p_i\} = (d(P)^S)^T$. But since

$$(P^S)^T - \{p_i\} = (d(P)^S)^T, \quad (29)$$

by definition of the operation S , and since Y has the same shape as $(P^S)^T$, we must have $Y = (P^S)^T$.

This proves part (b), and part (a) follows by an application of Theorem B. Applying parts (a) and (b) successively then shows that case (c) corresponds to $((P^T)^S)^T, ((Q^S)^T)^T$; and this is (P^S, Q^S) since $(P^S)^T = (P^T)^S$ by the row-column symmetry of operation S . ■

In particular, this theorem establishes two surprising facts about the tableau insertion algorithm: If successive insertion of distinct elements p_1, \dots, p_n into an empty tableau yields tableau P , insertion in the opposite order p_n, \dots, p_1 yields the *transposed* tableau P^T . And if we not only insert the p 's in this order p_n, \dots, p_1 but also interchange the roles of $<$ and $>$, as well as 0 and ∞ , in the insertion process, we obtain the dual tableau P^S . The reader is urged to try out these processes on some simple examples. The unusual nature of these coincidences might lead us to suspect that some sort of witchcraft is operating

behind the scenes! No simple explanation for these phenomena is yet known; there seems to be no obvious way to prove even that case (c) corresponds to tableaux having the same *shape* as P and Q , although the characterization of classes in exercise 2 does provide a significant clue.

The correspondence of Theorem A was given by G. de B. Robinson [*American J. Math.* **60** (1938), 745–760, §5], in a somewhat vague and different form, as part of his solution to a rather difficult problem in group theory. Robinson stated Theorem B without proof. Many years later, C. Schensted independently rediscovered the correspondence, which he described in terms of “bumping” as we have done in Algorithm I; Schensted also proved the “ P ” part of Theorem D(a) [see *Canadian J. Math.* **13** (1961), 179–191]. M. P. Schützenberger [*Math. Scand.* **12** (1963), 117–128] proved Theorem C and the “ Q ” part of Theorem D(a), from which (b) and (c) follow. It is possible to extend the correspondence to permutations of *multisets*; the case that p_1, \dots, p_n need not be distinct was considered by Schensted, and the “ultimate” generalization to the case that both the p 's and the q 's may contain repeated elements was investigated by Knuth [*Pacific J. Math.* **34** (1970), 709–727].

Let us now turn to a related question: *How many tableaux formed from $\{1, 2, \dots, n\}$ have a given shape (n_1, n_2, \dots, n_m) , where $n_1 + n_2 + \dots + n_m = n$?* If we denote this number by $f(n_1, n_2, \dots, n_m)$, and if we allow the parameters n_j to be arbitrary integers, the function f must satisfy the relations

$$f(n_1, n_2, \dots, n_m) = 0 \quad \text{unless} \quad n_1 \geq n_2 \geq \dots \geq n_m \geq 0; \quad (30)$$

$$f(n_1, n_2, \dots, n_m, 0) = f(n_1, n_2, \dots, n_m); \quad (31)$$

$$f(n_1, n_2, \dots, n_m) = f(n_1 - 1, n_2, \dots, n_m) + f(n_1, n_2 - 1, \dots, n_m) \\ + \dots + f(n_1, n_2, \dots, n_m - 1), \\ \text{if } n_1 \geq n_2 \geq \dots \geq n_m \geq 1. \quad (32)$$

Recurrence (32) comes from the fact that a tableau with its largest element removed is always another tableau; for example, the number of tableaux of shape $(6, 4, 4, 1)$ is $f(5, 4, 4, 1) + f(6, 3, 4, 1) + f(6, 4, 3, 1) + f(6, 4, 4, 0) = f(5, 4, 4, 1) + f(6, 4, 3, 1) + f(6, 4, 4)$, since every tableau of shape $(6, 4, 4, 1)$ on $\{1, 2, \dots, 15\}$ is formed by inserting the element 15 into the appropriate place in a tableau of shape $(5, 4, 4, 1)$, $(6, 4, 3, 1)$, or $(6, 4, 4)$. Schematically:

$$\begin{array}{|c|c|c|c|c|c|} \hline \square & \square & \square & \square & \square & \square \\ \hline \square & \square & \square & \square & & \\ \hline \square & \square & \square & \square & & \\ \hline \square & & & & & \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|c|} \hline \square & \square & \square & \square & \square & \square \\ \hline \square & \square & \square & \square & & \\ \hline \square & \square & \square & \square & & \\ \hline \square & & & & & \\ \hline \end{array}^{15} + \begin{array}{|c|c|c|c|c|c|} \hline \square & \square & \square & \square & \square & \square \\ \hline \square & \square & \square & \square & & \\ \hline \square & \square & \square & \square & & \\ \hline \square & & & & & \\ \hline \end{array}^{15} + \begin{array}{|c|c|c|c|c|c|} \hline \square & \square & \square & \square & \square & \square \\ \hline \square & \square & \square & \square & & \\ \hline \square & \square & \square & \square & & \\ \hline \square & & & & & \\ \hline \end{array}^{15} \quad (33)$$

The function $f(n_1, n_2, \dots, n_m)$ that satisfies these relations has a fairly simple form,

$$f(n_1, n_2, \dots, n_m) = \frac{\Delta(n_1 + m - 1, n_2 + m - 2, \dots, n_m) n!}{(n_1 + m - 1)! (n_2 + m - 2)! \dots n_m!}, \quad (34)$$

provided that the relatively mild conditions

$$n_1 + m - 1 \geq n_2 + m - 2 \geq \dots \geq n_m$$

are satisfied; here Δ denotes the “square root of the discriminant” function

$$\Delta(x_1, x_2, \dots, x_m) = \det \begin{pmatrix} x_1^{m-1} & x_2^{m-1} & \dots & x_m^{m-1} \\ \vdots & \vdots & & \vdots \\ x_1^2 & x_2^2 & & x_m^2 \\ x_1 & x_2 & & x_m \\ 1 & 1 & \dots & 1 \end{pmatrix} = \prod_{1 \leq i < j \leq m} (x_i - x_j). \quad (35)$$

Formula (34) was derived by G. Frobenius [*Sitzungsberichte preuß. Akad. der Wissenschaften* (1900), 516–534, §3], in connection with an equivalent problem in group theory, using a rather deep group-theoretical argument; a combinatorial proof was given independently by MacMahon [*Philosophical Trans.* **A209** (1909), 153–175]. The formula can be established by induction, since relations (30) and (31) are readily proved and (32) follows by setting $y = -1$ in the identity of exercise 17.

Theorem A gives a remarkable identity in connection with this formula for the number of tableaux. If we sum over all shapes, we have

$$\begin{aligned} n! &= \sum_{\substack{k_1 \geq k_2 \geq \dots \geq k_n \geq 0 \\ k_1 + k_2 + \dots + k_n = n}} f(k_1, k_2, \dots, k_n)^2 \\ &= n!^2 \sum_{\substack{k_1 \geq k_2 \geq \dots \geq k_n \geq 0 \\ k_1 + k_2 + \dots + k_n = n}} \frac{\Delta(k_1 + n - 1, k_2 + n - 2, \dots, k_n)^2}{(k_1 + n - 1)!^2 (k_2 + n - 2)!^2 \dots k_n!^2} \\ &= n!^2 \sum_{\substack{q_1 > q_2 > \dots > q_n \geq 0 \\ q_1 + q_2 + \dots + q_n = (n+1)n/2}} \frac{\Delta(q_1, q_2, \dots, q_n)^2}{q_1!^2 q_2!^2 \dots q_n!^2}; \end{aligned}$$

hence

$$\sum_{\substack{q_1 + q_2 + \dots + q_n = (n+1)n/2 \\ q_1, q_2, \dots, q_n \geq 0}} \frac{\Delta(q_1, q_2, \dots, q_n)^2}{q_1!^2 q_2!^2 \dots q_n!^2} = 1. \quad (36)$$

The inequalities $q_1 > q_2 > \dots > q_n$ have been removed in the latter sum, since the summand is a symmetric function of the q ’s that vanishes when $q_i = q_j$. A similar identity appears in exercise 24.

The formula for the number of tableaux can also be expressed in a much more interesting way, based on the idea of “hooks.” The *hook* corresponding to a cell in a tableau is defined to be the cell itself plus the cells lying below and to its right. For example, the shaded area in Fig. 5 is the hook corresponding to cell (2, 3) in row 2, column 3; it contains six cells. Each cell of Fig. 5 has been filled in with the length of its hook.

12	11	8	7	5	4	1	
10	9	6	5	3	2		•
9	8	5	4	2	1		•
6	5	2	1				•
3	2						
2	1						

Fig. 5. Hooks and hook lengths.

If the shape of the tableau is (n_1, n_2, \dots, n_m) , the longest hook has length $n_1 + m - 1$. Further examination of the hook lengths shows that row 1 contains all the lengths $n_1 + m - 1, n_1 + m - 2, \dots, 1$ *except* for $(n_1 + m - 1) - (n_m)$, $(n_1 + m - 1) - (n_{m-1} + 1)$, \dots , $(n_1 + m - 1) - (n_2 + m - 2)$. In Fig. 5, for example, the hook lengths in row 1 are 12, 11, 10, \dots , 1 except for 10, 9, 6, 3, 2; the exceptions correspond to five nonexistent hooks, from nonexistent cells (6, 3), (5, 3), (4, 5), (3, 7), (2, 7) leading up to cell (1, 7). Similarly, row j contains all lengths $n_j + m - j, \dots, 1$, except for $(n_j + m - j) - (n_m), \dots, (n_j + m - j) - (n_{j+1} + m - j - 1)$. It follows that the product of all the hook lengths is equal to

$$\frac{(n_1 + m - 1)! (n_2 + m - 2)! \dots n_m!}{\Delta(n_1 + m - 1, n_2 + m - 2, \dots, n_m)}.$$

This is just what happens in Eq. (34), so we have derived the following celebrated result due to J. S. Frame, G. de B. Robinson, and R. M. Thrall [*Canadian J. Math.* **6** (1954), 316–318]:

Theorem H. *The number of tableaux on $\{1, 2, \dots, n\}$ having a specified shape is $n!$ divided by the product of the hook lengths.* ■

Since this is such a simple rule, it deserves a simple proof; a heuristic argument runs as follows: Each element of the tableau is the smallest in its hook. If we fill the tableau shape at random, the probability that cell (i, j) will contain the minimum element of the corresponding hook is the reciprocal of the hook length; multiplying these probabilities over all i and j gives Theorem H. But unfortunately this argument is fallacious, since the probabilities are far from independent! No direct proof of Theorem H, based on combinatorial properties of hooks used correctly, was known until 1992 (see exercise 39), although researchers did discover several instructive indirect proofs (exercises 35, 36, and 38).

Theorem H has an interesting connection with the enumeration of trees, which we considered in Chapter 2. We observed that binary trees with n nodes correspond to permutations that can be obtained with a stack, and that such permutations correspond to sequences $a_1 a_2 \dots a_{2n}$ of n S's and n X's, where the number of S's is never less than the number of X's as we read from left to right. (See exercises 2.2.1–3 and 2.3.1–6.) The latter sequences correspond in a natural way to tableaux of shape (n, n) ; we place in row 1 the indices i such that $a_i = S$, and in row 2 we put those indices with $a_i = X$. For example, the sequence

S S S X X S S X X S X X

corresponds to the tableau

$$\begin{array}{|c|c|c|c|c|c|} \hline 1 & 2 & 3 & 6 & 7 & 10 \\ \hline 4 & 5 & 8 & 9 & 11 & 12 \\ \hline \end{array} . \tag{37}$$

The column constraint is satisfied in this tableau if and only if the number of X's never exceeds the number of S's from left to right. By Theorem H, the number of tableaux of shape (n, n) is

$$\frac{(2n)!}{(n+1)!n!};$$

so this is the number of binary trees, in agreement with Eq. 2.3.4.4-(14). Furthermore, this argument solves the more general “ballot problem” considered in the answer to exercise 2.2.1-4, if we use tableaux of shape (n, m) for $n \geq m$. So Theorem H includes some rather complex enumeration problems as simple special cases.

Any tableau A of shape (n, n) on the elements $\{1, 2, \dots, 2n\}$ corresponds to two tableaux (P, Q) of the same shape, in the following way suggested by MacMahon [*Combinatory Analysis 1* (1915), 130–131]: Let P consist of the elements $\{1, \dots, n\}$ as they appear in A ; then Q is formed by taking the remaining elements, rotating the configuration by 180° , and replacing $n+1, n+2, \dots, 2n$ by $n, n-1, \dots, 1$, respectively. For example, (37) splits into

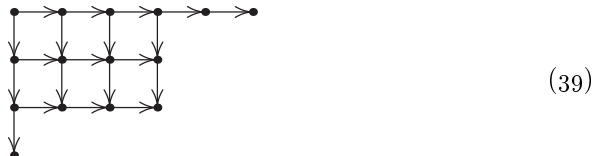
$$\begin{array}{|c|c|c|c|} \hline 1 & 2 & 3 & 6 \\ \hline 4 & 5 & & \\ \hline \end{array} \quad \text{and} \quad \begin{array}{|c|c|c|c|} \hline & & 7 & 10 \\ \hline 8 & 9 & 11 & 12 \\ \hline \end{array} ;$$

rotation and renaming of the latter yields

$$P = \begin{array}{|c|c|c|c|} \hline 1 & 2 & 3 & 6 \\ \hline 4 & 5 & & \\ \hline \end{array} , \quad Q = \begin{array}{|c|c|c|c|} \hline 1 & 2 & 4 & 5 \\ \hline 3 & 6 & & \\ \hline \end{array} . \tag{38}$$

Conversely, any pair of equal-shape tableaux of at most two rows, each containing n cells, corresponds in this way to a tableau of shape (n, n) . Hence by exercise 7 *the number of permutations $a_1 a_2 \dots a_n$ of $\{1, 2, \dots, n\}$ containing no decreasing subsequence $a_i > a_j > a_k$ for $i < j < k$ is the number of binary trees with n nodes*. An interesting one-to-one correspondence between such permutations and binary trees, more direct than the roundabout method via Algorithm I that we have used here, has been found by D. Rotem [*Inf. Proc. Letters 4* (1975), 58–61]; similarly there is a rather direct correspondence between binary trees and permutations having no instances of $a_i > a_k > a_j$ for $i < j < k$ (see exercise 2.2.1-5).

The number of ways to fill a tableau of shape $(6, 4, 4, 1)$ is obviously the number of ways to put the labels $\{1, 2, \dots, 15\}$ onto the vertices of the directed graph



(39)

in such a way that the label of vertex u is less than the label of vertex v whenever $u \rightarrow v$. In other words, it is the number of ways to sort the partial ordering (39) topologically, in the sense of Section 2.2.3.

In general, we can ask the same question for any directed graph that contains no oriented cycles. It would be nice if there were some simple formula generalizing Theorem H to the case of an arbitrary directed graph; but not all graphs have such pleasant properties as the graphs corresponding to tableaux. Some other classes of directed graphs for which the labeling problem has a simple solution are discussed in the exercises at the close of this section. Other exercises show that some directed graphs have *no* simple formula corresponding to Theorem H. For example, the number of ways to do the labeling is not always a divisor of $n!$.

To complete our investigations, let us count the total number of tableaux that can be formed from n distinct elements; we will denote this number by t_n . By Corollary B, t_n is the number of involutions of $\{1, 2, \dots, n\}$. A permutation is its own inverse if and only if its cycle form consists solely of one-cycles (fixed points) and two-cycles (transpositions). Since t_{n-1} of the t_n involutions have (n) as a one-cycle, and since t_{n-2} of them have $(j \ n)$ as a two-cycle, for fixed $j < n$, we obtain the formula

$$t_n = t_{n-1} + (n-1)t_{n-2}, \quad (40)$$

which Rothe devised in 1800 to tabulate t_n for small n . The values for $n \geq 0$ are 1, 1, 2, 4, 10, 26, 76, 232, 764, 2620, 9496, \dots

Counting another way, let us suppose that there are k two-cycles and $(n-2k)$ one-cycles. There are $\binom{n}{2k}$ ways to choose the fixed points, and the multinomial coefficient $(2k)!/(2!)^k$ is the number of ways to arrange the other elements into k distinguishable transpositions; dividing by $k!$ to make the transpositions indistinguishable we therefore obtain

$$t_n = \sum_{k=0}^{\lfloor n/2 \rfloor} t_n(k), \quad t_n(k) = \frac{n!}{(n-2k)! 2^k k!}. \quad (41)$$

Unfortunately, this sum has no simple closed form (unless we choose to regard the Hermite polynomial $i^n 2^{-n/2} H_n(-i/\sqrt{2})$ as simple). So we resort to two indirect approaches in order to understand t_n better:

a) We can find the generating function

$$\sum_n t_n z^n / n! = e^{z+z^2/2}; \quad (42)$$

see exercise 25.

b) We can determine the asymptotic behavior of t_n . This is an instructive problem, because it involves some general techniques that will be useful to us in other connections, so we will conclude this section with an analysis of the asymptotic behavior of t_n .

The first step in analyzing the asymptotic behavior of (41) is to locate the main contribution to the sum. Since

$$\frac{t_n(k+1)}{t_n(k)} = \frac{(n-2k)(n-2k-1)}{2(k+1)}, \quad (43)$$

we can see that the terms gradually increase from $k=0$ until $t_n(k+1) \approx t_n(k)$ when k is approximately $\frac{1}{2}(n-\sqrt{n})$; then they decrease to zero when k exceeds $\frac{1}{2}n$. The main contribution clearly comes from the vicinity of $k = \frac{1}{2}(n-\sqrt{n})$. It is usually preferable to have the main contribution at the value 0, so we write

$$k = \frac{1}{2}(n-\sqrt{n}) + x, \quad (44)$$

and we will investigate the size of $t_n(k)$ as a function of x .

One useful way to get rid of the factorials in $t_n(k)$ is to use Stirling's approximation, Eq. 1.2.11.2-(18). For this purpose it is convenient (as we shall see in a moment) to restrict x to the range

$$-n^{\epsilon+1/4} \leq x \leq n^{\epsilon+1/4}, \quad (45)$$

where $\epsilon = 0.001$, say, so that an error term can be included. A somewhat laborious calculation, which the author did by hand in the 60s but which is now easily done with the help of computer algebra, yields the formula

$$t_n(k) = \exp\left(\frac{1}{2}n \ln n - \frac{1}{2}n + \sqrt{n} - \frac{1}{4} \ln n - 2x^2/\sqrt{n} - \frac{1}{4} - \frac{1}{2} \ln \pi - \frac{4}{3}x^3/n + 2x/\sqrt{n} + \frac{1}{3}/\sqrt{n} - \frac{4}{3}x^4/n\sqrt{n} + O(n^{5\epsilon-3/4})\right). \quad (46)$$

The restriction on x in (45) can be justified by the fact that we may set $x = \pm n^{\epsilon+1/4}$ to get an upper bound for all of the discarded terms, namely

$$e^{-2n^{2\epsilon}} \exp\left(\frac{1}{2}n \ln n - \frac{1}{2}n + \sqrt{n} - \frac{1}{4} \ln n - \frac{1}{4} - \frac{1}{2} \ln \pi + O(n^{3\epsilon-1/4})\right), \quad (47)$$

and if we multiply this by n we get an upper bound for the sum of the excluded terms. The upper bound is of lesser order than the terms we will compute for x in the restricted range (45), because of the factor $\exp(-2n^{2\epsilon})$, which is much smaller than any polynomial in n .

We can evidently remove the factor

$$\exp\left(\frac{1}{2}n \ln n - \frac{1}{2}n + \sqrt{n} - \frac{1}{4} \ln n - \frac{1}{4} - \frac{1}{2} \ln \pi + \frac{1}{3}/\sqrt{n}\right) \quad (48)$$

from the sum, and this leaves us with the task of summing

$$\begin{aligned} & \exp\left(-2x^2/\sqrt{n} - \frac{4}{3}x^3/n + 2x/\sqrt{n} - \frac{4}{3}x^4/n\sqrt{n} + O(n^{5\epsilon-3/4})\right) \\ &= \exp\left(\frac{-2x^2}{\sqrt{n}}\right) \left(1 - \frac{4}{3}\frac{x^3}{n} + \frac{8}{9}\frac{x^6}{n^2}\right) \left(1 + 2\frac{x}{\sqrt{n}} + 2\frac{x^2}{n}\right) \\ & \quad \times \left(1 - \frac{4}{3}\frac{x^4}{n\sqrt{n}}\right) (1 + O(n^{9\epsilon-3/4})) \quad (49) \end{aligned}$$

over the range $x = \alpha, \alpha+1, \dots, \beta-2, \beta-1$, where $-\alpha$ and β are approximately equal to $n^{\epsilon+1/4}$ (and not necessarily integers). Euler's summation formula, Eq. 1.2.11.2-(10), can be written

$$\sum_{\alpha \leq x < \beta} f(x) = \int_{\alpha}^{\beta} f(x) dx - \frac{1}{2} f(x) \Big|_{\alpha}^{\beta} + \frac{1}{2} B_2 \frac{f'(x)}{1!} \Big|_{\alpha}^{\beta} + \cdots + \frac{1}{m+1} B_{m+1} \frac{f^{(m)}(x)}{m!} \Big|_{\alpha}^{\beta} + R_{m+1}, \quad (50)$$

by translation of the summation interval. Here $|R_m| \leq (4/(2\pi)^m) \int_{\alpha}^{\beta} |f^{(m)}(x)| dx$. If we let $f(x) = x^t \exp(-2x^2/\sqrt{n})$, where t is a fixed nonnegative integer, Euler's summation formula will give an asymptotic series for $\sum f(x)$ as $n \rightarrow \infty$, since

$$f^{(m)}(x) = n^{(t-m)/4} g^{(m)}(n^{-1/4}x), \quad g(y) = y^t e^{-2y^2}, \quad (51)$$

and $g(y)$ is a well-behaved function independent of n . The derivative $g^{(m)}(y)$ is e^{-2y^2} times a polynomial in y , hence $R_m = O(n^{(t+1-m)/4}) \int_{-\infty}^{+\infty} |g^{(m)}(y)| dy = O(n^{(t+1-m)/4})$. Furthermore if we replace α and β by $-\infty$ and $+\infty$ in the right-hand side of (50), we make an error of at most $O(\exp(-2n^{2\epsilon}))$ in each term. Thus

$$\sum_{\alpha \leq x < \beta} f(x) = \int_{-\infty}^{\infty} f(x) dx + O(n^{-m}), \quad \text{for all } m \geq 0; \quad (52)$$

only the integral is really significant, given this particular choice of $f(x)$! The integral is not difficult to evaluate (see exercise 26), so we can multiply out and sum formula (49), giving $\sqrt{\pi/2}(n^{1/4} - \frac{1}{24}n^{-1/4} + O(n^{-1/2}))$. Thus

$$t_n = \frac{1}{\sqrt{2}} n^{n/2} e^{-n/2 + \sqrt{n}-1/4} \left(1 + \frac{7}{24}n^{-1/2} + O(n^{-3/4})\right). \quad (53)$$

Actually the O -terms here should have an extra 9ϵ in the exponent, but our manipulations make it clear that this 9ϵ would disappear if we had carried further accuracy in the intermediate calculations. In principle, the method we have used could be extended to obtain $O(n^{-k})$ for any k , instead of $O(n^{-3/4})$. This asymptotic series for t_n was first determined (using a different method) by Moser and Wyman, *Canadian J. Math.* **7** (1955), 159–168.

The method we have used to derive (53) is an extremely useful technique for asymptotic analysis that was introduced by P. S. Laplace [*Mémoires Acad. Sci.* (Paris, 1782), 1–88]; it is discussed under the name “trading tails” in *CMath*, §9.4. For further examples and extensions of tail-trading, see the conclusion of Section 5.2.2.

EXERCISES

- [16] What tableaux (P, Q) correspond to the two-line array

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 6 & 4 & 9 & 5 & 7 & 1 & 2 & 8 & 3 \end{pmatrix},$$

in the construction of Theorem A? What two-line array corresponds to the tableaux

$$P = \begin{array}{|c|c|c|} \hline 1 & 4 & 7 \\ \hline 2 & 8 & \\ \hline 5 & 9 & \\ \hline \end{array}, \quad Q = \begin{array}{|c|c|c|} \hline 1 & 3 & 7 \\ \hline 4 & 5 & \\ \hline 8 & 9 & \\ \hline \end{array} ?$$

2. [M21] Prove that (q, p) belongs to class t with respect to (16) if and only if t is the largest number of indices i_1, \dots, i_t such that

$$p_{i_1} < p_{i_2} < \dots < p_{i_t} = p, \quad q_{i_1} < q_{i_2} < \dots < q_{i_t} = q.$$

► 3. [M24] Show that the correspondence defined in the proof of Theorem A can also be carried out by constructing a table such as this:

Line 0	1	3	5	6	8
Line 1	7	2	9	5	3
Line 2	∞	7	∞	9	5
Line 3		∞		∞	7
Line 4					∞

Here lines 0 and 1 constitute the given two-line array. For $k \geq 1$, line $k + 1$ is formed from line k by the following procedure:

- a) Set $p \leftarrow \infty$.
- b) Let column j be the leftmost column in which line k contains an integer $< p$, but line $k + 1$ is blank. If no such columns exist, and if $p = \infty$, line $k + 1$ is complete; if no such columns exist and $p < \infty$, return to (a).
- c) Insert p into column j in line $k + 1$, then set p equal to the entry in column j of line k and return to (b).

Once the table has been constructed in this way, row k of P consists of those integers in line k that are not in line $(k + 1)$; row k of Q consists of those integers in line 0 that appear in a column containing ∞ in line $k + 1$.

► 4. [M30] Let $a_1 \dots a_{j-1} a_j \dots a_n$ be a permutation of distinct elements, and assume that $1 < j \leq n$. The permutation $a_1 \dots a_{j-2} a_j a_{j-1} a_{j+1} \dots a_n$, obtained by interchanging a_{j-1} with a_j , is called “admissible” if either

- i) $j \geq 3$ and a_{j-2} lies between a_{j-1} and a_j ; or
- ii) $j < n$ and a_{j+1} lies between a_{j-1} and a_j .

For example, exactly three admissible interchanges can be performed on the permutation 1 5 4 6 8 3 7; we can interchange the 1 and the 5 since $1 < 4 < 5$; we can interchange the 8 and the 3 since $3 < 6 < 8$ (or since $3 < 7 < 8$); but we cannot interchange the 5 and the 4, or the 3 and the 7.

- a) Prove that an admissible interchange does not change the tableau P formed from the permutation by successive insertion of the elements a_1, a_2, \dots, a_n into an initially empty tableau.
- b) Conversely, prove that any two permutations that have the same P tableau can be transformed into each other by a sequence of one or more admissible interchanges. [Hint: Given that the shape of P is (n_1, n_2, \dots, n_m) , show that any permutation that corresponds to P can be transformed into the “canonical permutation” $P_{m1} \dots P_{mn_m} \dots P_{21} \dots P_{2n_2} P_{11} \dots P_{1n_1}$ by a sequence of admissible interchanges.]

► 5. [M22] Let P be the tableau corresponding to the permutation $a_1 a_2 \dots a_n$; use exercise 4 to prove that P^T is the tableau corresponding to $a_n \dots a_2 a_1$.