APPLIED COMBINATORICS

THIRD EDITION

ALAN TUCKER

Department of Applied Mathematics and Statistics
State University of New York at Stony Brook

JOHN WILEY & SONS, INC.
NEW YORK CHICHESTER BRISBANE TORONTO SINGAPORE

CHAPTER 4

4. COMBINATORICS AND PROBABILITY

"CLASSICAL PROBABILITY"

$\zeta \Rightarrow$ sample outcomes are equally likely in Ω

i.e $\Omega \equiv \{ \omega_1, \omega_2, \dots, \omega_{n_\Omega} \}$

$\#$ outcomes in $\Omega = |\Omega| = n_\Omega$

$$P(\{\omega_i\}) = \frac{1}{n_\Omega} \quad \forall i.$$

If $E \subseteq \Omega$, and

$|E| = n_E$

Then

$$P(E) = \frac{n_E}{n_\Omega}$$

Hence need to evaluate / enumerate

n_E, n_Ω

Such problems arise in simple "mechanical" experiments

- coins / dice / cards / lottery

where knowledge of physical system is sufficient to use "symmetry" assumptions etc. to justify the "equally likely" assumption.

It is not always straightforward to enumerate n_E, n_Ω.

4.1 COUNTING OPERATIONS

– BASIC TECHNIQUES AND TERMINOLOGY

THE MULTIPLICATION PRINCIPLE

A sequence of k operations, in which operation i can result in n_i outcomes, can result in a total of

$$n_1 \times n_2 \times \ldots \times n_k$$

sequences of outcomes.

Recall, for $n \in \mathbb{Z}^+$, the factorial

$$n! = n(n-1)(n-2)\ldots 3.2.1$$

$(0! \equiv 1)$ —

Many combinatorics problems arise in the context of

"SAMPLING FROM A FINITE POPULATION

i.e. N items from which we wish to select a sample of size n, say, under certain conditions.

e.g. context could dictate that the successive sampling operations are carried out

"WITH REPLACEMENT" of items

or

"WITHOUT REPLACEMENT" of items

e.g. could be that, in the resulting sample, the items are regarded as

"ORDERED" – order of sampling is important

or

"UNORDERED" – order not important.

EXAMPLE

For the population $\{1, 2, \ldots, N\}$, to obtain a sample of size $n = 4$.

Number of equally likely samples :

WITH REP. $n_1 \times n_2 \times n_3 \times n_4$

$$= N \times N \times N \times N = N^4$$

WITHOUT REP $n_1 \times n_2 \times n_3 \times n_4$

$$= N(N-1)(N-2)(N-3)$$

ORDERED Results $(3,6,2,7)$ and $(2,3,6,7)$ are distinct

UNORDERED $(3,6,2,7) \equiv (2,3,6,7)$

Finally, could have that items in population are

DISTINGUISHABLE

(i.e. individually labelled)

or

INDISTINGUISHABLE

(i.e. labelled according to a <u>class</u> <u>type</u>, but not labelled individually).

<u>EXAMPLE</u> Students in IC Maths

DISTINGUISHABLE	00MATH001
	00MATH002
	⋮
	99MATH001
	⋮
INDISTINGUISHABLE	YEAR 1
	YEAR 2
	YEAR 3
	YEAR 4

<u>DEFINITION</u>

An ordered arrangement of n items is a <u>PERMUTATION</u>. The number of ways of ordering a subset of $r \leq n$ items is

$$^nP_r = \frac{n!}{(n-r)!} = (n)_r$$

$$= n(n-1)\dots(n-r+1)$$

- the number of ways of producing an <u>ordered</u> sample of r items from a population of size n.

<u>DEFINITION</u>

An unordered arrangement of r items from n is a <u>COMBINATION</u>. The number of combinations is

$$^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- must have

$$^nP_r = r!\,^nC_r$$

- the number of <u>unordered</u> samples of size r from n.
- number of ways of choosing r from n.

E.G. <u>National Lottery</u>

Sample of size $r = 6$ from $n = 49$

$\{1, 2, \ldots, 49\}$ DISTINGUISHABLE

ORDERED : $^{49}P_6 = \dfrac{49!}{43!}$

UNORDERED : $^{49}C_6 = \binom{49}{6}$

- here order is <u>not</u> important in resulting sample.

BINARY SEQUENCE REPRESENTATIONS

For sampling without replacement
- unordered sample

write

e.g. 0100100100

to represent that

ITEMS 2, 5, 8 IN SAMPLE

ITEMS 1, 3, 4, 6, 7, 9, 10 NOT IN SAMPLE

- often a useful representation for combinatorial problems.

$\binom{n}{r}$ is the number of subsets of size r from n items

Recall the BINOMIAL EXPANSION

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$$

or

$$(1+x)^n = \sum_{r=0}^{n} \binom{n}{r} x^r$$

Thus the function $(1+x)^n$ is a "GENERATING FUNCTION" for the coefficients $a_r = \binom{n}{r}$

In summary: in sampling r from n items

	WITHOUT REPLACEMENT	WITH REPLACEMENT
ORDERED	$(n)_r$	n^r
UNORDERED	$\binom{n}{r}$	$\binom{n+r-1}{r}$

- table gives total number of possible samples.

4.2 COMBINATORIAL IDENTITIES

e.g. $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$

– consider binary sequence rep.

- SEE HANDOUT FOR OTHER IDENTITIES.

BINOMIAL IDENTITIES

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{n-k}$$

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n$$

$$\binom{n}{0} + \binom{n+1}{1} + \binom{n+2}{2} + \cdots + \binom{n+r}{r} = \binom{n+r+1}{r}$$

$$\binom{r}{r} + \binom{r+1}{r} + \binom{r+2}{r} + \cdots + \binom{n}{r} = \binom{n+1}{r+1}$$

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}$$

$$\sum_{k=0}^{r} \binom{m}{k}\binom{n}{r-k} = \binom{m+n}{r}$$

$$\sum_{k=0}^{m} \binom{m}{k}\binom{n}{r+k} = \binom{m+n}{m+r}$$

$$\sum_{k=0}^{m} \binom{m-k}{r}\binom{n+k}{s} = \binom{m+n+1}{r+s+1}$$

4.3 THE HYPERGEOMETRIC FORMULA

Consider

Finite pop^n : N items

of which R TYPE I

$N-R$ TYPE II

– a sample of size n is obtained without replacement

What is the composition of the sample?

What is the probability that the sample contains r TYPE I items?

$$n_E = \binom{R}{r}\binom{N-R}{n-r} \qquad (1)$$

$$n_\Omega = \binom{N}{n}$$

$\therefore$ Prob is $\dfrac{n_E}{n_\Omega}$.

- consider selecting items for the sample

For example

$$\overbrace{\circ\,\circ\,\circ\,\circ\cdots\circ}^{\text{TYPE I}}_{R}\;\overbrace{\circ\,\circ\,\circ\cdots\;\circ}^{\text{TYPE II}}_{N-R}$$

n_Ω - select n positions to include in sample.

n_E - select r of the O s, then select $n-r$ of the O s.

Or, equivalently (?)

$$n_E = \binom{n}{r}\binom{N-n}{R-r} \qquad (2)$$

$$n_\Omega = \binom{N}{R}$$

Prob is $\dfrac{n_E}{n_\Omega}$

- consider arranging the items in sequence ...

For example

$$\overbrace{\circ\,\circ\,\circ\,\circ\,\circ\cdots\circ}^{\text{IN SAMPLE}}_{n}\;\overbrace{\circ\,\circ\,\circ\,\circ\cdots\cdots\circ}^{\text{NOT IN SAMPLE}}_{N-n}$$

n_Ω - select R positions to colour O

n_E - select r positions from n, then select $R-r$ positions from $N-n$ to colour O.

$\therefore$ Must have

$$\frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} = \frac{\binom{n}{r}\binom{N-n}{R-r}}{\binom{N}{R}}$$

- verify this by direct comparison of factorial terms.

NOTE

These formulae hold if

$0 \leq r \leq n$

$0 \leq r \leq R$

$0 \leq n-r \leq N-R$

ie

$$\mathrm{Max}\{0, N-R+n\} \leq r \leq \mathrm{Min}\{n, R\}$$

EXAMPLE National Lottery

$N = 49$

$R = 6$ Winning numbers

$n = 6$ Your selections

Probability of matching r numbers is

$$\frac{\binom{6}{r}\binom{43}{6-r}}{\binom{49}{6}} \qquad r = 0, 1, \ldots, 6.$$

NOTE

Here we have fixed N, R, n and considered probabilities for various velves of r

However, the "HYPERGEOMETRIC FORMULAE" are general formulae for

$$\{N, R, n, r\}$$

– if we know <u>any</u> three of these values we can calculate probabilities as the other varies.

EXAMPLE Capture / Recapture

N animals

R captured ; tagged ; released.

Subsequently

n animals recaptured

r found to be tagged.

The probability of recapturing r tagged animals is $\dfrac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}}$

Such approaches are used to discover <u>population</u> <u>size</u>

i.e. <u>given</u> r tagged animals recaptured, what is the probability that the population size is equal to N?
(typically N is <u>NOT</u> observed)

- A <u>BAYES</u> <u>THEOREM</u> CALCULATION

- let A_N = "pop^n size is N"
  B_r = "r tagged animals recaptured"
  etc.

THE HYPERGEOMETRIC DISTRIBUTION IN STATISTICS

FISHER'S EXACT TEST

The Hypergeometric probability formula is used in a direct statistical calculation to assess association between two factors. The calculation, and the associated testing procedure, is known as *Fisher's Exact Test*, and relates to a 2×2 table of count data; suppose that two factors, Factor 1 and Factor 2, each have two possible levels and the the count data are dispayed in a table as below:

		FACTOR 2		
		Level 1	Level 2	Total
FACTOR 1	Level 1	a	b	$a+b$
	Level 2	c	d	$c+d$
	Total	$a+c$	$b+d$	N

where $N = a+b+c+d$. Now, we wish to construct a test of association between the two factors, that is, we wish to test whether the classification by Factor 1 is influenced by the classification by Factor 2, or *vice versa*. Our assessment will be carried out under the assumption that the Row and Column Totals are both fixed, that is, that the Row Totals are $a+b$ and $c+d$, and that the Column Totals are $a+c$ and $b+d$; this assumption may be regarded as part of the experimental (data collection) design.

Now, our assessment will be based on the assumption that, if no association between Row and Column classification exists, then we may regard all allocations of observations to cells as equally likely, subject to the constraint that the Row and Column Totals are fixed. Our interest will lie in calculating the probability of observing a particular configuration of table entries. First, we wish to make our assessment symmetric in the labels on the Row and Column levels; that is, we will consider the four tables

$$\begin{matrix} a & b \\ c & d \end{matrix} \qquad \begin{matrix} c & d \\ a & b \end{matrix} \qquad \begin{matrix} b & a \\ d & c \end{matrix} \qquad \begin{matrix} d & c \\ b & a \end{matrix}$$

formed, respectively, by swapping Rows, Columns, and Rows and Columns in the original table, as identical. Without loss of generality, we assume that the original table is the one of the four in which the first Row Total is less than or equal to the Second Row Total ($a+b \leq c+d$), and the first Column Total is less than or equal to the Second Column Total ($a+c \leq b+d$). Denote this table by the vector $[a,b,c,d]$ for the entries to cells 1,2,3 and 4 in the table.

Using a classical probability approach, we know that the probability of observing the table $[a,b,c,d]$ is n_E/n_{TOT} in the usual notation, where n_E is the number of ways of allocating observations to the table so that the respective cell entries are a,b,c,d, and n_{TOT} is the number of ways of allocating the observations to the table in total, both under the assumption that the Row Totals are $a+b$ and $c+d$, and the Column Totals are $a+c$ and $b+d$.

<u>4.4</u> <u>COMBINATORIAL</u> <u>PARTITIONING</u>

$\binom{n}{r}$ - # ways of sampling r items from n
(unordered, without replacement)

- involves constructing a "PARTITION" of the n items into <u>two</u> subsets

i.e "SAMPLED" size r
and
"REMAINING" size $n-r$

To partition into <u>k</u> subsets

SUBSET 1	r_1 items
SUBSET 2	r_2 items
⋮	⋮
SUBSET k	r_k items

where

$$\sum_{i=1}^{k} r_i = n$$

and

$$r_i \geq 0 \quad \forall i.$$

What is the corresponding number of ways?

Using the multiplication rule :

$$\# \text{ ways} = \binom{n}{r_1}\binom{n-r_1}{r_2}\cdots\binom{n-r_1-\cdots-r_{k-1}}{r_k}$$

① ② ⋯ ⓚ

1) Select the r_1 items for sub 1
2) " " r_2 " " " 2
(from the $n-r_1$ items remaining)

⋮

k) select the r_k items for subset k
(from the $n-r_1-r_2-\cdots-r_{k-1}$ items remaining)

$\therefore$ Number of ways is

$$\binom{n}{r_1, r_2, \ldots, r_k} = \frac{n!}{r_1!\, r_2! \ldots r_k!}$$

EXAMPLE Bridge

\# possible deals (52 cards into 4 hands)

is $\dfrac{52!}{13!\,13!\,13!\,13!}$ $k = 4$, $r_1 = r_2 = r_3 = r_4 = 13$

EXAMPLE

12 Dice are rolled ; what is the probability that the scores {1,2,3,4,5,6} each appears twice ?

SOLUTION

$$P(E) = \frac{n_E}{n_\Omega} = \frac{12!}{6^{12}(2!)^6} \simeq 0.0034$$

n_Ω = # possible rolls = 6^{12}

n_E = # partitions of 12 into 6 lots of 2 $= \dfrac{12!}{2!\,2!\,2!\,2!\,2!\,2!} = \dfrac{12!}{(2!)^6}$

EXAMPLE **POKER HANDS**

5 cards dealt from pack of 52

– relative strength of hand determined by its probability.

Total number of poker hands in single deal of five cards is

$$n_\Omega = \binom{52}{5}$$

(select 5 cards from 52 without replacement, unordered)

"STRONG" HANDS

PAIR	xxabc
TWO PAIR	xxyya
THREE OF A KIND	xxxab
FULL HOUSE	xxxyy

x, y : "scoring" cards

a, b, c : non scoring cards

Special case of "partitioning"

- could use a direct "Conditional probability / Chain rule" approach.

COMBINATORIAL APPROACH

e.g. PAIR: xxabc

$$n_E = \underbrace{\binom{13}{1}}_{\text{denomination for } x} \underbrace{\binom{4}{2}}_{\text{Two cards (suits) from four for the } x} \times \underbrace{\binom{12}{3}}_{\text{3 different denominations for } a,b,c} \underbrace{\binom{4}{1}\binom{4}{1}\binom{4}{1}}_{\text{one card (suit) from four for each of } a,b,c.}$$

$$\therefore P(E) = \frac{n_E}{n_\Omega} = \frac{1098240}{2598960} \approx 0.423$$

TWO PAIR : xxyya

$$n_E = \binom{13}{2}\binom{4}{2}\binom{4}{2} \times \binom{11}{1}\binom{4}{1}$$

need two different denominations for x, y.

FULL HOUSE : xxxyy

$$n_E = \binom{13}{1}\binom{4}{3} \times \binom{12}{1}\binom{4}{2}$$

(see Solutions to Tutorial sheet for full list).

Table 10.14 Spectacle wearing among juvenile delinquents and non-delinquents who failed a vision test (Weindling *et al.*, 1986)

		Juvenile delinquents	Non-delinquents	Total
Spectacle wearers	Yes	1	5	6
	No	8	2	10
	Total	9	7	16

From last lecture ....

POKER HANDS

PAIR $\quad xx\,abc$

$$n_E = \binom{13}{1}\binom{4}{2}\binom{12}{3}\binom{4}{1}^3$$

-need a, b, c **DIFFERENT**
denominations : 9,8,2
10,6,5
A,Q,9
etc

But, do not count

9, 8, 2

as different from

2, 8, 9 $\quad$ etc

$$\Rightarrow \binom{12}{3}, \text{ not } \binom{12}{1}\binom{11}{1}\binom{10}{1}.$$

Similarly for **TWO PAIRS** $\quad xxyya$

$$\binom{13}{2}, \text{ not } \binom{13}{1}\binom{12}{1}$$

as "AAKK2" $\equiv$ "KKAA2"

OCCUPANCY PROBLEMS

Consider the combinatorial problem of allocating r items (objects,balls) to n boxes (cells): for example, can we enumerate

the number of ways of allocating 6 items to 4 boxes

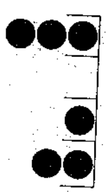


Fig1. Allocation of $n = 6$ balls to $r = 4$ cells

To count the possible number of allocations, we consider the cases of *DISTINGUISHABLE* and *INDISTINGUISHABLE* items separately.

If the items are distinguishable, that is, labelled $1, 2, \ldots r$, we consider *ordered* arrangements. The following two allocations are different:

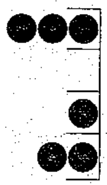


Fig 2a. Allocation 1

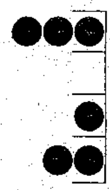


Fig 2b. Allocation 2

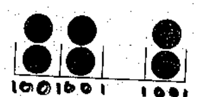
INDISTINGUISHABLE ITEMS

If the items are indistinguishable, that is, here completely identical, and we wish to consider distinct allocation patterns, we must consider *unordered* arrangements; the two allocations in Fig 2a. and Fig 2b are regarded as identical, and identical to the allocation in Fig 1., as the items are not labelled.

For example, consider forming the allocation pattern by dropping the items into the boxes in sequence:

ITEM	1	2	3	4	5	6
SEQUENCE (1)	2	1	2	4	1	4
SEQUENCE (2)	3	2	4	1	3	3
SEQUENCE (3)	2	4	4	1	1	2

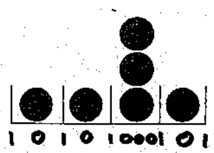
Then, we have that the patterns obtained by sequences (1) and (3) are both of the form



100|001|100|

Fig 3a. Allocation patterns for sequences (1) and (3)

and thus we do not count (1) and (3) as distinct patterns, whereas sequence (2) produces an allocation pattern of



10|0|0001|01|

Fig 3b. Allocation patterns for sequence (2)

which is distinct from the pattern for (1) and (3)

and we can record the two allocations as follows

	ITEM	1	2	3	4	5	6
ALLOCATION 1	BOX LABEL SEQUENCE	2	4	4	1	4	1
ALLOCATION 2	BOX LABEL SEQUENCE	1	2	4	1	4	4

ɪnd, essentially, we have selected r box labels from n **with replacement**, where the box labels are ɪrdered. Hence the number of possible allocations is n^r, by a previous result using the multiplication heorem.

If we require

r_1 items in BOX 1
r_2 items in BOX 2
$\vdots \quad \vdots$
r_n items in BOX n

then we must **partition** the box label sequence to contain r_1 1s, r_2 2s,...,r_n ns. Hence the number (possible allocations is given by the partition formula

$$\frac{r!}{r_1!r_2!\ldots r_n!} \quad \text{where} \quad \sum_{i=1}^{n} r_i = r$$

EXAMPLE

If r identical dice are rolled, with $n = 6$ possible scores for each die, the total number of distinct score patterns is

$$\binom{n+r-1}{r} = \binom{6+r-1}{r} = \binom{5+r}{r}$$

For example, with $r = 4$, we could have

DICE NUMBER				SCORE PATTERN					
1	2	3	4	1	2	3	4	5	6
6	1	6	2	1	1	0	0	0	2
3	2	3	4	0	1	2	1	0	0
6	2	1	6	1	1	0	0	0	2

and the number of distinct patterns is

$$\binom{9}{4} = 126$$

4

OCCUPANCY PROBLEMS : EXAMPLES

EXAMPLE 1 Allocate n items to n boxes. Evaluate the probability of event E that no box is empty.

SOLUTION: Probability is

$$P(E) = \frac{n_E}{n_\Omega} = \frac{n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1}{n \times n \times n \times \ldots \times n \times n \times n} = \frac{n!}{n^n}$$

as we allocate by sampling n boxes with and without replacement for numerator and denominator respectively (each of which are a product of n terms).

To enumerate the number of possible allocation patterns, we utilize a binary sequence representation. We code an allocation pattern by reading from left to right, and writing a 1 for a box edge, and 0 for an item, so that the pattern in Fig3a. is coded

1 0 0 1 0 0 1 1 0 0 1

and the pattern in Fig3b is coded

1 0 1 0 1 0 0 0 1 0 1

The number of possible allocation patterns is equal to the number of binary sequences that correspond to them, and these sequences are composed as follows; they contain $n+1$ 1s (for the box edges) and r 0s (for the items), but also they begin with a 1, and end with a 1. The number of sequences like this is therefore equal to the number of ways of arranging a sequence of $(n+1)+r-2 = n+r-1$ binary digits containing precisely $n-1$ 1s and r 0s. This number is

$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$

from combination/binomial coefficient definition, and this is therefore the total number of distinct allocation patterns.

EXAMPLE 2 Allocate r items to n boxes. Evaluate the probability of event E that no box is contains more than one item.

SOLUTION: Probability is

$$P(E) = \frac{n_E}{n_\Omega} = \frac{n \times (n-1) \times (n-2) \times \ldots \times (n-r+1)}{n \times n \times n \times \ldots \times n} = \frac{n!/(n-r)!}{n^r} = \frac{(n)_r}{n^r}$$

as we allocate by sampling r boxes **with** and **without** replacement for numerator and denominator respectively (each of which are a product of r terms).

NOTE: we can re-write $P(E)$ using a conditional probability/chain rule argument corresponding to a sequence of selection probabilities:

$$P(E) = 1 \times \left(1 - \frac{1}{n}\right) \times \left(1 - \frac{2}{n}\right) \times \ldots \left(1 - \frac{r-1}{n}\right)$$

where each term is the conditional probability of choosing a currently empty box, given the allocations at that instant.

SPECIAL CASE : THE BIRTHDAY PROBLEM

In a group of r people, what is the probability that no two people have the same birthday ? Assuming that all of the $n = 365$ days in the year are equally likely to be a birthday, then we identify in EXAMPLE 2 the "boxes" as days, and "items" as people, and evaluate the probability as

$$\frac{(n)_r}{n^r} = \frac{(365)_r}{365^r}$$

which we can evaluate numerically

r	5	10	20	22	23	50
Probability	0.973	0.883	0.589	0.524	0.493	0.030

EXAMPLE 3 Allocate r items to n boxes. Evaluate the probability of event E that box 1 contains precisely k items.

SOLUTION: For $0 \leq k \leq r$, probability is

$$P(E) = \frac{n_E}{n_\Omega} = \frac{\binom{r}{k}(n-1)^{r-k}}{n^r} = \binom{r}{k}\left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{r-k}$$

For the numerator, we first select the k items from r (**without** replacement) to place in box 1, and then select $(r-k)$ boxes from the remaining $(n-1)$ (**with** replacement) to house the remaining $(r-k)$ items. For the denominator, we merely select r boxes from n with replacement.

CHAPTER 4 - SUMMARY

- permutations/combinations
- sampling with/without replacement
- Hypergeometric Probabilities
- Partitioning
- Occupancy Problems

(- binary sequence representations)