Chapter 4. Algebraic Structures— Groups, Rings, Fields

The preceding chapters have provided us with a great deal of concrete material, which it is now time to consider from a more general point of view. For this purpose, we introduce and study (still on an elementary level) the concepts of groups, rings, and fields, which play a fundamental role in all of algebra.

§1. Sets with algebraic operations

1. Binary operations. Let X be any set. An algebraic binary operation (or composition law) on X is any fixed map $\tau : X \times X \to X$ of the cartesian square $X^2 = X \times X$ to X. Thus, to every ordered pair (a,b) of elements $a, b \in X$ there corresponds a unique element $\tau(a,b)$ of the same set X. We sometimes write $a \tau b$ instead of $\tau(a,b)$, and in fact usually we introduce a special symbol $*, \circ, \cdot, +$, etc. to designate a binary operation on X. In our examples we shall follow this practice, and shall call $a \cdot b$ (or simply ab with no symbol between a and b) the product and $a + b$ the sum of the elements $a, b \in X$. In most cases one of these conventions will be convenient.

It is possible for the same set X to have more than one binary operation defined on it. When we choose one such operation, and want to think of X in conjunction with that particular binary operation $*$, we write $(X, *)$ and say that $*$ determines an algebraic structure on X or that $(X, *)$ is an algebraic system. For example, in addition to the usual operations $+$ (addition) and $\cdot$ (multiplication) on the set $\mathbb{Z}$ of integers, it is easy to give new operations made up from $+$ (or $-$) and $\cdot$: $n \circ m = n + m - nm$, $n * m = -n - m$, etc. We thereby obtain different algebraic structures $(\mathbb{Z}, +)$, $(\mathbb{Z}, \cdot)$, $(\mathbb{Z}, \circ)$, $(\mathbb{Z}, *)$.

Clearly, the imagination can find a boundless expanse in constructing all sorts of binary operations on a set X. But the problem of studying arbitrary algebraic structures is too general to lead to conclusions that have any concrete value. For this reason the problem is studied under certain restrictions to various special types of algebraic structures.

2. Semigroups and monoids. A binary operation $*$ on a set X is called associative if $(a*b)*c = a*(b*c)$ for all $a, b, c \in X$; it is called commutative if $a*b = b*a$ for all $a, b \in X$. We use the same terms to apply to the algebraic structure $(X, *)$. The conditions of associativity and commutativity are independent, i.e., neither implies the other. For example, the operation $*$ on $\mathbb{Z}$ given by $n*m = -n-m$ is obviously commutative, but it is not associative, since, for example, $(1*2)*3 = (-1-2)*3 = -(-1-2)-3 = 0$, while $1*(2*3) = 4$. On the other hand, the set $M_n(\mathbb{R})$ of all $n \times n$ square matrices is associative but not commutative under multiplication (if $n > 1$), as shown in subsection 2 of §3 Ch. 2.

An element $e \in X$ is called a unit element (or a neutral element) relative to a given binary operation $*$ if we have: $e*x = x*e = x$ for all $x \in X$. If e' is another unit element, then it follows immediately from the definition that $e' = e'*e = e$. Thus, an algebraic structure $(X, *)$ can have at most one unit element.

A set X with an associative binary operation is called a semigroup. A semigroup having a unit element is called a monoid (or simply a semigroup with unit).

As with any set, the cardinality of a monoid $M = (M, *)$ is denoted Card M or $|M|$. If it has finitely many elements, we call M a finite monoid of order $|M|$.

We now give some examples of semigroups and monoids.

1) Let Ω be any set, and let $M(\Omega)$ be the set of all transformations of Ω (maps from Ω to itself). It follows from the properties of sets and maps in §5 Ch. 1 that $M(\Omega)$ is a monoid. Of course, we have in mind as our binary operation the composition of maps $\circ$. $M(\Omega)$ has the unit element e_Ω which is the identity map.

Consider the special case when Ω is a finite set of $n = |\Omega|$ elements, which we simply denote by the integers $1, 2, \cdots, n$. Every map $f : \Omega \to \Omega$ is determined by giving an ordered sequence $f(1), f(2), \cdots, f(n)$, where each $f(i)$ is an element of Ω. We allow the possibility that $f(i) = f(j)$ for $i \neq j$. There are precisely n^n possible sequences, i.e., n^n transformations. Thus, $|M(\Omega)| = \text{Card } M(\Omega) = n^n$. For example, take $n = 2$. The four elements e, f, g, h of the monoid $M(\{1,2\})$ and their products (compositions) are completely given by the tables

	1	2
e	1	2
f	2	1
g	1	1
h	2	2

$\cdot$	e	f	g	h
e	e	f	g	h
f	f	e	h	g
g	g	g	g	g
h	h	h	h	h

It is clear from the table to the right that $M(\{1,2\})$ is a non-commutative monoid.

2) Again let Ω be an arbitrary set, and let $\mathcal{P}(\Omega)$ be the set of all subsets (see Exercise 4 of §5 Ch. 1). Since $(A \cap B) \cap C = A \cap (B \cap C)$ and $(A \cup B) \cup C = A \cup (B \cup C)$, it follows that there are two natural associative binary operations defined on $\mathcal{P}(\Omega)$. Obviously, $\phi \cup A = A$ and $A \cap \Omega = A$. So we have two commutative monoids $(\mathcal{P}(\Omega), \cup, \phi)$ and $(\mathcal{P}(\Omega), \cap, \Omega)$, where we denote a monoid as a triple (set, binary operation, unit element). We know that $|\mathcal{P}(\Omega)| = 2^n$ if $|\Omega| = n$.

3) $(M_n(\mathbb{R}), +, 0)$ is a commutative monoid whose neutral element is the zero matrix, and $(M_n(\mathbb{R}), \cdot, E)$ is a non-commutative monoid whose neutral element is the identity matrix. This follows immediately from the properties of matrix addition and multiplication which we encountered in Chapter 2.

4) Let $n\mathbb{Z} = \{mn \mid m \in \mathbb{Z}\}$ be the set of integers divisible by n. It is clear that $(n\mathbb{Z}, +, 0)$ is a commutative monoid, and $(n\mathbb{Z}, \cdot)$ is a commutative semigroup without unit (if $n > 1$).

5) The set $P_n(\mathbb{R})$ of stochastic matrices of order n (see Exercise 7 of §3 Ch. 2) is a monoid under the usual matrix multiplication.

A subset S' of a semigroup S with binary operation $*$ is called a <u>subsemigroup</u> if $x*y \in S'$ whenever $x, y \in S'$. In this case we say that <u>the subset $S' \subset S$ is closed under $*$</u>. If $(M, *)$ is a monoid, and the subset $M' \subset M$ is not only closed under $*$, but also contains the unit element, then we say that M' is a <u>submonoid</u> of M. For example, $(n\mathbb{Z}, \cdot)$ is a subsemigroup of $(\mathbb{Z}, \cdot)$, and $(n\mathbb{Z}, +, 0)$ is a submonoid of $(\mathbb{Z}, +, 0)$. Any submonoid of the monoid $M(\Omega)$ of maps from a set to itself is called a <u>monoid of transformations</u> (of the set Ω).

3. <u>Generalized associativity; powers</u>. Let $(X, \cdot)$ be any algebraic structure. For simplicity, we shall omit the $\cdot$ and write xy instead of $x \cdot y$. Let $x_1, \cdots, x_n$ be an ordered sequence of elements in X. Without changing the order, there are many different ways we can form the product of n elements. Let ℓ_n be the number of ways:

$$\ell_2 = 1 : x_1x_2;$$

$$\ell_3 = 2 : (x_1x_2)x_3,\ x_1(x_2x_3);$$

$$\ell_4 = 5 : ((x_1x_2)x_3)x_4,\ (x_1(x_2x_3))x_4,\ x_1((x_2x_3)x_4),\ x_1(x_2(x_3x_4)),$$

$(x_1x_2)(x_3x_4)$; and so on.

It is clear that we can obtain all ℓ_n possibilities if, for each k, $1 \leq k \leq n-1$,

we run through all possible products of $x_1, \cdots, x_k$ and all possible products of $x_{k+1}, \cdots, x_n$, and then take the product of these two products. It is of great importance that the location of the parentheses does not matter if $(X, \cdot)$ is a semigroup.

THEOREM 1. *If the binary operation on* X *is associative, then the result of applying it successively to* n *elements of* X *does not depend on the location of the parentheses.*

Proof. If $n = 1$ or 2, there is nothing to prove. If $n = 3$, the theorem coincides with the associative law. We now proceed by induction on n. Suppose that $n > 3$ and that the theorem holds when the number of elements is $< n$. We need only show that

$$(x_1 \cdots x_k) \cdot (x_{k+1} \cdots x_n) = (x_1 \cdots x_l) \cdot (x_{l+1} \cdots x_n) \tag{1}$$

for any k and l, $1 \leq k$, $l \leq n - 1$. We have only written out the outside parentheses, since, by the induction assumption, the location of the inner paretheses does not matter. In particular, we can set $x_1 x_2 \cdots x_k = (\ldots((x_1 x_2)x_3)\ldots x_{k-1})x_k$, which is called the *left-normalized* product. We distinguish between two cases:

a) $k = n - 1$. Then $(x_1 \cdots x_{n-1})x_n = (\ldots(x_1 x_2)\ldots x_{n-1})x_n$ is a left-normalized product.

b) $k < n - 1$. By associativity, we have

$$(x_1 \cdots x_k)(x_{k+1} \cdots x_n) = (x_1 \cdots x_k)((x_{k+1} \cdots x_{n-1})x_n) =$$

$$= ((x_1 \cdots x_k)(x_{k+1} \cdots x_{n-1}))x_n =$$

$$= (\ldots((\ldots(x_1 x_2)\ldots x_k)x_{k+1})\ldots x_{n-1})x_n ,$$

i. e., we again obtain a left-normalized product. The right side of the desired equality (1) can be reduced to the same form. □

In §2 Ch. 2 we introduced the summation sign Σx_i. It can be used in any additive commutative monoid. The analogous symbol in a multiplicative monoid is the product sign:

$$\prod_{i=1}^{2} x_i = x_1 x_2 \quad , \quad \prod_{i=1}^{3} x_i = (x_1 x_2) x_3 \quad , \quad \prod_{i=1}^{n} x_i = \left(\prod_{i=1}^{n-1} x_i \right) x_n \quad .$$

By Theorem 1, the parentheses are not needed when writing (or computing) a product $x_1 x_2 \cdots x_n$ of elements in a monoid. The only care that must be exercised is in the order of the elements, and even that is unnecessary if the elements commute with one another. In particular, if $x_1 = x_2 = \ldots = x_n = x$, then, as with ordinary numbers, we let x^n denote the product $xx \ldots x$, which we call the n-th power of the element x. As a consequence of Theorem 1, we have the relations

$$x^m x^n = x^{m+n} \quad , \quad (x^m)^n = x^{mn} \quad , \quad m, n \in \mathbb{N} \quad . \tag{2}$$

If $(M, \cdot, e)$ is a monoid, we further set $x^0 = e$ for every $x \in M$.

In an additive monoid, i.e., in a commutative monoid in which the operation is denoted by $+$, the "power" x^n is written $nx = x + x + \cdots + x$ and is called a multiple of x. Then (2) becomes the following rules for multiples:

$$mx + nx = (m + n)x \quad , \quad n(mx) = (nm)x \quad . \tag{2'}$$

We note another useful fact. If $xy = yx$ in a monoid M, then

$$(xy)^n = x^n y^n \quad , \quad n = 0, 1, 2, \cdots \quad . \tag{3}$$

In particular, this is always the case in a commutative monoid. (3) is proved by induction on n:

$$(xy)^n = (xy)^{n-1}(xy) = (x^{n-1} y^{n-1})(xy) = (x^{n-1} y^{n-1} x)y =$$

$$= (x^{n-1} x y^{n-1})y = (x^{n-1} x)(y^{n-1} y) = x^n y^n \quad .$$

More generally, if we use (3) and induction on m, we obtain

$$x_i x_j = x_j x_i \ , \ 1 \leq i, j \leq m \implies (x_1 \cdots x_m)^n = x_1^n \cdots x_m^n \quad . \tag{4}$$

The analogous rules for multiples are:

$$n(x+y) = nx + ny\,, \qquad n = 0, 1, 2, \cdots\,, \tag{3'}$$

$$n(x_1 + \cdots + x_m) = nx_1 + \cdots + nx_m\,, \qquad n = 0, 1, 2, \cdots\,. \tag{4'}$$

Normally, a monoid which is written $(M, \cdot, e)$ is called a multiplicative monoid, and one which is written $(M, +, 0)$ is called an additive monoid. The additive notation is usually only used for commutative monoids.

4. Invertible elements. An element a in a monoid $(M, \cdot, e)$ is called invertible if there exists an element $b \in M$ such that $ab = e = ba$. (Clearly, in that case b is also invertible.) If we also had $ab' = e = b'a$, then it would follow that $b' = eb' = (ba)b' = b(ab') = be = b$. Thus, we can speak of the inverse element a^{-1} when a is invertible: $a^{-1}a = e = aa^{-1}$.

It is clear that $(a^{-1})^{-1} = a$. The notion of an invertible element in a monoid is a natural generalization of the notion of an invertible matrix in the multiplicative monoid $(M_n(\mathbb{R}), \cdot, E)$.

Since $(xy)(y^{-1}x^{-1}) = x(yy^{-1})x^{-1} = xex^{-1} = xx^{-1} = e$, and similarly $(y^{-1}x^{-1})(xy) = e$, we have: $(xy)^{-1} = y^{-1}x^{-1}$. Hence, the set of all invertible elements in a monoid $(M, \cdot, e)$ is closed under $\cdot$ and is a submonoid of M.

EXERCISES

1. Subsection 2 contains the example of the commutative but non-associative operation $* : n * m = -n - m$ on $\mathbb{Z}$. The algebraic system $(\mathbb{Z}, *)$ has the following identities: $(n * m) * m = n$, $m * (m * n) = n$. Now suppose that we are given an arbitrary algebraic system $(X, *)$ in which $(x * y) * y = x$ and $y * (y * x) = x$ for all $x, y \in X$. Prove that $x * y = y * x$, i.e., $*$ is commutative.

2. Show that

$$M_n^0(\mathbb{R}) = \left\{ A = (a_{ij}) \in M_n(\mathbb{R}) \,\middle|\, \sum_{j=1}^{n} a_{ij} = 0 \ , \ i = 1, 2, \cdots, n \right\}$$

is a semigroup under the usual operation of matrix multiplication. Is $(M_n^0(\mathbb{R}), \cdot)$ a monoid?

3. In a multiplicative monoid M we choose an arbitrary element t and introduce the new operation $* : x * y = xty$. Show that $(M, *)$ is a semigroup, and that $(M, *)$ is a monoid if and only if t is invertible, in which case the neutral element is t^{-1}.

4. Show that the set $\mathbb{Z}$ with the operation $\circ : n \circ m = n + m + nm = (1 + n)(1 + m) - 1$ is a commutative monoid. What is the neutral element in $(\mathbb{Z}, \circ)$? Find all invertible elements in $(\mathbb{Z}, \circ)$.

§2. Groups

1. Definition and examples. Consider the set $GL(n, \mathbb{R})$ of all square $n \times n$ matrices with non-zero determinant. By Theorem 5 of §2 Ch. 3, if $\det A \neq 0$ and $\det B \neq 0$, then $\det AB \neq 0$. Thus, $A, B \in GL(n, \mathbb{R}) \Rightarrow AB \in GL(n, \mathbb{R})$. In addition, $(AB)C = A(BC)$, and there is a special matrix E such that $AE = EA = A$ for all $A \in GL(n, \mathbb{R})$. Finally, every matrix $A \in GL(n, \mathbb{R})$ has an "opposite", i.e., an inverse A^{-1} such that $AA^{-1} = A^{-1}A = E$.

The set $GL(n, \mathbb{R})$ considered with the composition law (binary operation) $(A, B) \mapsto AB$ is called the general linear group of order n over $\mathbb{R}$. Following the terminology in §1, we could define $GL(n, \mathbb{R})$ as simply the submonoid of all invertible elements in the monoid $(M_n(\mathbb{R}), \cdot, E)$. But this submonoid is extremely important in its own right, and is a key example of the following abstract definition.

Definition. A monoid G all of whose elements are invertible is called a group. In other words, the following axioms must hold:

(G1) a binary operation $(x,y) \mapsto xy$ is defined on the set G;

(G2) this operation is associative: $(xy)z = x(yz)$ for all $x,y,z \in G$;

(G3) G has a neutral (unit) element e: $xe = ex = x$ for all $x \in G$;

(G4) every element $x \in G$ has an inverse x^{-1}: $xx^{-1} = x^{-1}x = e$.

Surprisingly, one of the oldest and richest areas of algebra, playing a fundamental role in geometry and in applications of mathematics to the natural sciences, is based on such a simple set of axioms.

A group whose binary operation is commutative is called a commutative group or else an abelian group (in honor of the Norwegian mathematician Abel). The term "group" itself was introduced by the French mathematician Galois, the founder of group theory. The ideas of group theory were "in the air" (as often happens with fundamental mathematical ideas) long before Galois; some of the theorems of group theory were actually proved, although in a more naive form, by Lagrange. The brilliant work of Galois was at first poorly understood, and its importance became fully recognized only after the appearance of Jordan's book "A course in the theory of permutations and algebraic equations" (1870). It was only toward the end of the nineteenth century that group theory "completely left the realm of fantasy. Instead, a logical skeleton was carefully prepared". (F. Klein, "Lectures on the development of mathematics in the nineteenth century").

The symbols $\operatorname{Card} G$, $|G|$, and $(G:e)$ are all used to denote the number of elements in a group (its cardinality). All of the facts about monoids in §1 apply, of course, to groups. However, some new words are introduced. A subset $H \subset G$ is called a subgroup of G if $e \in H$; $h_1, h_2 \in H \Rightarrow h_1h_2 \in H$; and $h \in H \Rightarrow h^{-1} \in H$. A subgroup $H \subset G$ is called proper if $H \neq G$.

We now give some examples of groups.

1) In the general linear group $GL(n, \mathbb{R})$, consider the subset $SL(n, \mathbb{R})$ of matrices with determinant 1:

$$SL(n, \mathbb{R}) = \{A \in GL(n, \mathbb{R}) \mid \det A = 1\} \quad .$$

Obviously, $E \in SL(n, \mathbb{R})$. By the results in Ch. 3 about determinants, $\det A = 1$, $\det B = 1 \Rightarrow \det AB = 1$, and $\det A^{-1} = (\det A)^{-1} = 1$. Thus, $SL(n, \mathbb{R})$ is a subgroup of $GL(n, \mathbb{R})$; it is called the special linear group of order n over $\mathbb{R}$. It is also called the unimodular group, although this name is sometimes used for the group of matrices with determinant ± 1.

The group $GL(n, \mathbb{R})$, which contains many interesting groups, has been for mathematicians of several generations a seemingly inexhaustible source of new ideas and unsolved problems.

2) If we replace the real numbers by the rational numbers, we obtain the general linear group $GL(n, \mathbb{Q})$ of order n over $\mathbb{Q}$ and the subgroup $SL(n, \mathbb{Q})$. The group $SL(n, \mathbb{Q})$ contains the interesting subgroup $SL(n, \mathbb{Z})$ of matrices with integer entries and determinant 1. Theorem 1 of §3 Ch. 3, which gives an explicit formula for the entries in an inverse matrix, shows that $SL(n, \mathbb{Z})$ actually is a group. The groups $SL(n, \mathbb{Q})$ and $SL(n, \mathbb{Z})$ occupy an important place in number theory. Figure 11 depicts the partially order set (see subsection 3 of §6 Ch. 1) of these subgroups of $GL(n, \mathbb{R})$.

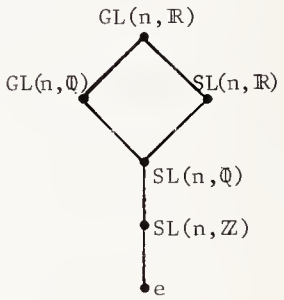


Fig. 11

3) If we set $n = 1$ in examples 1) and 2), we obtain, first of all, the multiplicative groups $\mathbb{R}^* = \mathbb{R}\setminus\{0\} = GL(1, \mathbb{R})$ and $\mathbb{Q}^* = \mathbb{Q}\setminus\{0\} = GL(1, \mathbb{Q})$ of real and rational numbers, respectively. These are obviously infinite groups. Since the only invertible elements in $(\mathbb{Z}, \cdot, 1)$ are 1 and -1, we have $GL(1, \mathbb{Z}) = \{\pm 1\}$. Furthermore, $SL(1, \mathbb{R}) = SL(1, \mathbb{Q}) = SL(1, \mathbb{Z}) = 1$. But if $n = 2$, even the group $SL(2, \mathbb{Z})$ is infinite: it contains, for example, all of the matrices

$$\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ m & 1 \end{pmatrix}, \quad \begin{pmatrix} m & m-1 \\ 1 & 0 \end{pmatrix}, \quad m \in \mathbb{Z} \quad .$$

We further note the infinite additive groups:

$$(\mathbb{R}, +, 0) \ , \ (\mathbb{Q}, +, 0) \ , \ (\mathbb{Z}, +, 0) \quad .$$

4) Let Ω be any set, and let $S(\Omega)$ be the set of all bijective (one-to-one) transformations $f : \Omega \to \Omega$. Using the results of §5 Ch. 1 on set maps (Theorems 1 and 2 and the corollary to Theorem 2), we immediately conclude that $S(\Omega)$ is a group under the natural binary operation of composing maps. $S(\Omega)$ is the submonoid of all invertible elements in the monoid $M(\Omega)$ in example 1) of §1. The group $S(\Omega)$ and its various subgroups are called transformation groups. Such groups are the basic type of groups that arise in applications of group theory. In 1872, F. Klein announced his "Erlangen program", which sought to classify different geometries using the notion of transformation groups. If we take $\Omega = \mathbb{R}^n$, we obtain a very large group $S(\mathbb{R}^n)$, which is difficult to imagine in its entirety. But $S(\mathbb{R}^n)$ contains the subgroup of invertible (bijective) linear transformations $\varphi_A : \mathbb{R}^n \to \mathbb{R}^n$, which we found to be in one-to-one correspondence with the $n \times n$ non-singular matrices A (see §3 Ch. 2). Thus, we have an imbedding of $GL(n, \mathbb{R})$ in $S(\mathbb{R}^n)$. The significance of this imbedding will become clearer after we introduce the important concept of an isomorphism of groups.

2. Systems of generators. Given a subset S of a group G, we try to find a subgroup $H \subset G$ containing S such that every subgroup $K \subset G$ containing S must also contain H. There cannot be two distinct subgroups H, H' which play the role of a minimal subgroup containing S:

$$S \subset H \ , \ S \subset H' \implies H \subset H' \subset H \implies H' = H \quad .$$

Thus, the minimal subgroup H containing S must coincide with the intersection of all subgroups containing S, if we show that this intersection must be a subgroup in G. But we have the following simple result.

THEOREM 1. The intersection $\bigcap_{i \in I} H_i$ of any family $\{H_i \mid i \in I\}$ of subgroups of a group G is a subgroup.

Proof. Let e be the unit element of G. The properties $e \in \bigcap H_i$, $x, y \in \bigcap H_i \Rightarrow xy \in \bigcap H_i$; $x \in \bigcap H_i \Rightarrow x^{-1} \in \bigcap H_i$, which characterize a subgroup, must hold in $\bigcap H_i$, because they hold in each subgroup H_i separately. □

Now take for $\{H_i \mid i \in I\}$ the family of all subgroups containing the given subset $S \subset G$. Then, by Theorem 1 and the remarks before the theorem, the intersection

$$\langle S \rangle = \bigcap_{S \subseteq H} H$$

is precisely the minimal subgroup containing S. We call $\langle S \rangle$ the subgroup generated by S in G, and we call S a set of generators for the subgroup $\langle S \rangle$. At first glance, it seems that $\langle S \rangle$ is defined ineffectively, since we have to find all subgroups containing S. But there is an easier way to determine $\langle S \rangle$, as we see from the following corollary of Theorem 1.

COROLLARY. The subgroup $\langle S \rangle$ coincides with the set T consisting of the unit element e and all possible products

$$t_1 t_2 \cdots t_n \quad , \quad n = 1, 2, 3, \cdots ,$$

where either $t_i \in S$ or $t_i^{-1} \in S$, $1 \leq i \leq n$.

Proof. Since $t_1 \cdots t_n \in T$, $t'_1 \cdots t'_m \in T \Rightarrow t''_1 \cdots t''_{n+m} = t_1 \cdots t_n t'_1 \cdots t'_m \in T$ and $t_1 \cdots t_n \in T \Rightarrow (t_1 \cdots t_n)^{-1} = t_n^{-1} \cdots t_1^{-1} \in T$, it follows that the set T is a subgroup of G. On the other hand, every subgroup H containing all $x_i \in S$ must contain all of the inverses x_i^{-1}, and, hence, must contain all products of the form $t_1 t_2 \cdots t_n$. Hence, $H \supset T$, and T coincides with the intersection of all such subgroups. □

It should be noted that by no means are all of the products $t_1 t_2 \cdots t_n$ distinct elements of $\langle S \rangle$, even if one agrees (as one usually does) to cancel any pair of successive t_i of the form $a\,a^{-1}$ or $a^{-1}a$. In general, when $|S| > 1$, the question of when two products of the form $t_1 t_2 \cdots t_n$ are equal is a difficult one, and we shall only briefly

discuss it in Chapter 7.

Every group G has some set of generators S: for example, we can take S to be the whole group G. For simplicity, we consider a group G which is generated by a finite set of elements (such groups are said to be finitely generated). If we remove from S all "extra" elements, i.e., those which can be written as products of the other elements (and their inverses), we obtain a minimal set M of generators of G. To say that M is minimal means that $\langle M \rangle = G$, but $\langle M' \rangle \neq G$ if M' is obtained from M by removing an element. Let $M = \{g_1, \cdots, g_d\}$. Then we also write $G = \langle g_1, g_2, \cdots, g_d \rangle$ as well as $G = \langle M \rangle$. If $d = 1$, we call G a cyclic group.

3. Cyclic groups. If G is any group and g is an element in G, then, by definition, $\langle g \rangle$ is a cyclic subgroup of G.

Because of Theorem 1 and the properties of the powers of an element in a monoid, we might expect that any cyclic group $\langle a \rangle$ with generator a is an abelian group of the form $\langle a \rangle = \{a^n \mid n \in \mathbb{Z}\}$, or $\langle a \rangle = \{na \mid n \in \mathbb{Z}\}$ if the group operation is written additively (this notation is not meant to imply that all of the elements a^n or na are distinct). This is in fact the case, once we agree to denote $(a^{-1})^k = a^{-k}$ and prove the following fact.

THEOREM 2. For any $m, n \in \mathbb{Z}$,

$$a^m a^n = a^{m+n}, \quad (a^m)^n = a^{mn}$$

(or, in additive notation, $ma + na = (m+n)a$, $n(ma) = (nm)a$).

Proof. If m and n are non-negative, see relations (2) and (2') of subsection 4 §1. If $m < 0$ and $n < 0$, then $m' = -m > 0$, $n' = -n > 0$, and

$$a^m a^n = (a^{-1})^{m'} (a^{-1})^{n'} = (a^{-1})^{m'+n'} = a^{-(m'+n')} = a^{m+n}.$$

If $m' = -m > 0$ and $n > 0$, we have

$$a^m a^n = (a^{-1})^{m'} a^n = (\underbrace{a^{-1} \cdots a^{-1}}_{m'})(\underbrace{a \cdots a}_{n}) = a^{n-m'} \text{ (or } (a^{-1})^{m'-n} \text{, if } m' \geq n) = a^{m+n} .$$

We similarly treat the case when $m > 0$ and $n < 0$. The equality $(a^m)^n = a^{mn}$ is easy to prove using the first equality, just proved, and the definition of the powers of an element. □

The simplest example of a cyclic group is the additive group of integers $(\mathbb{Z}, +, 0)$, which is generated by 1 or by -1. Also, it is easy to see that the matrix $\begin{Vmatrix} 1 & 1 \\ 0 & 1 \end{Vmatrix}$ generates an infinite cyclic subgroup of $SL(2, \mathbb{Z})$. The set $\{1, -1\}$ under multiplication is a cyclic group of order 2.

We can construct an example of a cyclic group of order n by considering all rotations of the plane around a point O which take a regular n-gon P_n with center O to itself. These rotations clearly form a group; the group operation is successively performing the rotations. Our group C_n contains the rotations $\varphi_0, \varphi_1, \cdots, \varphi_{n-1}$ through the angles 0, $2\pi/n, \cdots, (n-1)2\pi/n$, counterclockwise. Here $\varphi_s = \varphi_1^s$, and it is geometrically obvious that $\varphi_s^{-1} = \varphi_1^{n-s}$ and $\varphi_1^n = \varphi_0$, the identity transformation. Thus, $|C_n| = n$, and $C_n = \langle \varphi_1 \rangle$. Note that the cyclic group C_n is a proper subgroup of the group D_n of all symmetries of the n-gon P_n (i.e., rigid transformations of P_n).

Again suppose that G is any group and a is an element of G. There are two possibilities: 1) All powers of a are distinct, i.e., $m \neq n \Rightarrow a^m \neq a^n$. In this case we say that a has <u>infinite order</u>. 2) We have $a^m = a^n$ for $m \neq n$. If, say, $m > n$, then $a^{m-n} = e$, i.e., there is a positive power of a which is the identity element. Let q be the least positive exponent for which $a^q = e$. We then say that a is an element of <u>finite order</u> q. Of course, if G has finite order (i.e., Card $G < \infty$), then all of the elements have finite order.

<u>Warning</u>. The word "order" has many meanings in mathematics. Before we spoke of square matrices of order n (i.e., $n \times n$ matrices), but a non-singular matrix A,

considered as an element of the group $GL(n, \mathbb{R})$, also has an order (perhaps infinite) in the sense just defined. But it will always be clear from the context what meaning of the word "order" we have in mind.

If we think of our example C_n of a cyclic group of order n, the following theorem becomes almost obvious.

THEOREM 3. The order of an element $a \in G$ (where G is any group) is equal to Card $\langle a \rangle$. If a is an element of finite order q, then

$$\langle a \rangle = \{e, a, \cdots, a^{q-1}\} \quad \text{and} \quad a^k = e \iff k = \ell q \ , \ \ell \in \mathbb{Z} \quad .$$

Proof. If a has infinite order, there is nothing left to prove. If a has order q, then, by definition, all of the elements $a, a, a^2, \cdots, a^{q-1}$ are distinct. We claim that any power a^k must coincide with one of these elements, i.e., $\langle a \rangle = \{e, a, \cdots, a^{q-1}\}$. To see this, we use the division algorithm in $\mathbb{Z}$ (see subsection 3 §8 Ch. 1) to write k in the form

$$k = \ell q + r \ , \quad 0 \leq r \leq q - 1 \quad .$$

Then, using the rules in Theorem 2, we obtain

$$a^k = (a^q)^{\ell} a^r = e a^r = a^r \quad .$$

In particular, $a^k = e \Rightarrow r = 0 \Rightarrow k = \ell q$. □

The property of a group being cyclic is very useful. Given a group, we are not always told in advance whether it is cyclic; sometimes this must be proved. An example is the following

PROPOSITION. Let G be any group, and let a and b be two elements that commute with one another. Suppose that a and b have finite orders s and t, and that s and t are relatively prime. Then a and b generate a cyclic subgroup of order st, and

$$\langle a, b \rangle = \langle ab \rangle \quad .$$

Proof. First of all, $D = \langle a \rangle \cap \langle b \rangle = e$, since if $d \in D$ has order q, then, by Theorem 3,

$$d = a^i = b^j \Longrightarrow d^s = (a^s)^i = e\ ,\ d^t = (b^t)^j = e \Longrightarrow q|s\ ,\ q|t\ ,$$

and, since s and t are relatively prime, this means that $q = 1$. Next, if $n = |\langle ab \rangle|$, we have (see relation (3) in §1)

$$a^n b^n = (ab)^n = e \Rightarrow a^n = b^{-n} \in D = e \Rightarrow a^n = e\ ,\ b^n = e \Rightarrow s|n\ ,\ t|n \Rightarrow \text{l.c.m.}(s,t)|n \Rightarrow st|n\ ,$$

since $st = \text{l.c.m.}(s,t) \cdot \text{g.c.d.}(s,t) = \text{l.c.m.}(s,t)$. But $(ab)^{st} = (a^s)^t (b^t)^s = e$ (using Theorem 2), so that $n|st$, and hence $n = st$. It remains to note that

$$\langle a,b \rangle = \{a^i b^j | 0 \leq i \leq s-1\ ,\ 0 \leq j \leq t-1\} \Longrightarrow \text{Card}\,\langle a,b \rangle \leq st\ ,$$

and, since $\langle ab \rangle \subset \langle a,b \rangle$ and $\text{Card}\,\langle ab \rangle = st$, it follows that $\langle a,b \rangle = \langle ab \rangle$. □

We shall return to cyclic groups, but now we examine a richer special type of group, namely, transformation groups, which will be used to illustrate the various group theoretic concepts we introduce.

4. The symmetric group and the alternating group. Let Ω be a finite set with n elements. Since we shall not be concerned with the nature of those elements, we may as well assume that $\Omega = \{1, 2, \cdots, n\}$. The group $S(\Omega)$ (see Example 4 above) of all one-to-one correspondences $\Omega \to \Omega$ is called the n-th symmetric group (or the symmetric group on n elements), and is usually denoted S_n. The elements of S_n, which are usually denoted by small Greek letters, are call permutations.

Written out visually, a permutation $\pi : i \mapsto \pi(i)$, $i = 1, 2, \cdots, n$, is represented as follows:

$$\pi = \begin{pmatrix} 1 & 2 & \cdots & n \\ i_1 & i_2 & \cdots & i_n \end{pmatrix}\ ,$$

where all of the images are given explicitly:

$$\pi : \begin{matrix} 1 & 2 & \cdots & n \\ \downarrow & \downarrow & & \downarrow \\ i_1 & i_2 & \cdots & i_n \end{matrix} ;$$

here $i_k = \pi(k)$, $k = 1, \cdots, n$, are the permuted elements $1, 2, \cdots, n$. As usual, e denotes the identity permutation (even though it is a Latin and not a Greek letter): $e(i) = i$, $\forall\, i$.

Two permutations $\sigma, \tau \in S_n$ are multiplied by the usual rule for composing maps: $(\sigma\tau)(i) = \sigma(\tau(i))$. For example, if

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}, \quad \tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}$$

then we have

$$\sigma\tau = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} = \begin{matrix} 1 & 2 & 3 & 4 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 4 & 3 & 2 & 1 \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 4 & 3 & 2 \end{matrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{pmatrix} .$$

Notice that

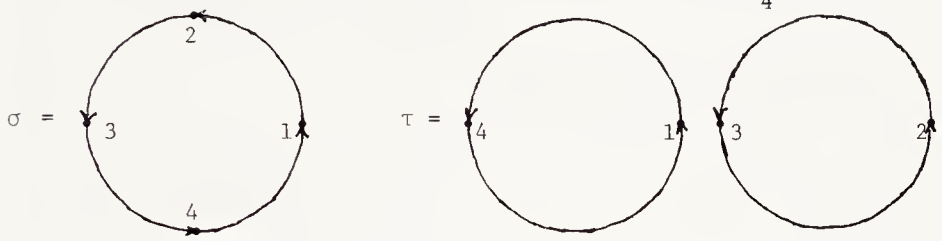
$$\tau\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 1 & 4 \end{pmatrix} ,$$

so that $\sigma\tau \neq \tau\sigma$.

We now find the order of the group S_n. A permutation σ can take the element 1 into any of n possible elements. Once $\sigma(1)$ is fixed, we can choose $\sigma(2)$ to be any of the remaining $n - 1$ elements; thus, there are $n(n-1)$ possible choices for the pair $\sigma(1), \sigma(2)$. Then we can choose $\sigma(3)$ to be any of the $n - 2$ numbers which have not already been taken to be $\sigma(1)$ or $\sigma(2)$. Continuing in this manner, we see that the number of possible choices for the $\sigma(1), \sigma(2), \cdots, \sigma(n)$ is $n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!$ ("n factorial"). Thus,

$$\text{Card } S_n = |S_n| = (S_n : e) = n! .$$

The permutations in S_n can be decomposed into products of simpler permutations. We illustrate by drawing diagrams for the two examples $\sigma, \tau \in S_4$ given above:



We abbreviate the permutation σ, which is a type of permutation called a "cycle" of length 4 (meaning that it "rotates" 4 elements), by writing $\sigma = (1234)$, or, equivalently, $\sigma = (2341) = (3412) = (4123)$. The permutation τ is the product of two "disjoint" cycles of length 2, namely the cycles (14) and (23), and so can be written $\tau = (14)(23)$. Note that $\sigma^2 = (13)(24)$, $\sigma^4 = e$, $\tau^2 = e$.

Returning to the general case, we call two elements $i, j \in \Omega$ <u>equivalent</u> under a cyclic subgroup $\langle \pi \rangle$ of S_n (or simply π-<u>equivalent</u>) if there is an integer s for which $j = \pi^s(i) = \pi(\ldots\pi(i)\ldots)$. Since S_n is a finite group, all of its subgroups are also finite. By Theorem 3, if $\text{Card}\,\langle \pi \rangle = q$, we may take $0 \leq s < q$. This relation is in fact an equivalence relation, i.e., it is reflexive, symmetric, and transitive (see subsection 2 of §6 Ch. 1), since $i = \pi^0 i = e(i)$; $j = \pi^k(i) \Rightarrow i = \pi^{-k}(j)$ and $j = \pi^s(i)$, $k = \pi^t(j) \Rightarrow k = \pi^{s+t}(i)$. By the usual property of equivalence relations, we obtain a partition

$$\Omega = \Omega_1 \cup \ldots \cup \Omega_p \tag{1}$$

of the set Ω into pair-wise disjoint subsets (equivalence classes) $\Omega_1, \cdots, \Omega_p$, which are called π-<u>orbits</u>. This name makes sense: every $i \in \Omega$ belongs to precisely one orbit, and if Ω_k is the orbit to which i belongs, then Ω_k consists of the images of i under the action of the powers of π: $i, \pi(i), \pi^2(i), \cdots, \pi^{\ell_k - 1}(i)$, where $\ell_k = |\Omega_k|$ is the length of the π-orbit Ω_k. It is obvious that $\ell_k \leq q = \text{Card}\,\langle \pi \rangle$; $\pi^{\ell_k}(i) = i$; and ℓ_k is the least positive integer such that $\pi^{\ell_k}(i) = i$. If we set

$$\pi_k = (i\pi(i) \dots \pi^{l_k-1}(i)) = \begin{pmatrix} i & \pi(i) & \dots & \pi^{l_k-2}(i) \\ \pi(i) & \pi^2(i) & \dots & \pi^{l_k-1}(i) \end{pmatrix} ,$$

we obtain a permutation which is a cycle of length l_k.

Sometimes for clarity we write the cycle $(123 \dots l)$ as $(1,2,3, \cdots , l)$, separating the numbers by commas. The cycle π_k is the permutation which leaves all elements in the set $\Omega \setminus \Omega_k$ fixed and takes j to $\pi(j)$ for all $j \in \Omega_k$. We therefore call π_s and π_t <u>independent</u> or <u>disjoint</u> cycles when $s \neq t$, since they effect disjoint sets of elements. Note that $\pi_k^{l_k} = e$.

Thus, associated with the partition (1) we have a corresponding decomposition of the permutation π into a product of disjoint cycles:

$$\pi = \pi_1 \pi_2 \cdots \pi_p \quad , \tag{2}$$

where all of the cycles commute with one another: $\pi = \pi_1 \pi_2 \cdots \pi_p = \pi_{i_1} \pi_{i_2} \cdots \pi_{i_p}$.

For example, we may arrange the cycles so that $l_1 \geq l_2 \geq \cdots \geq l_m > l_{m+1} = \cdots = l_p = 1$ (i.e., the last $p-m$ cycles correspond to orbits consisting of one element, so that as permutations $\pi_{m+1} = \cdots = \pi_p = e$). Since a cycle of length one acts as the identity, it is natural to omit such cycles in (2), and write

$$\pi = \pi_1 \pi_2 \cdots \pi_m \; ; \; l_k > 1 \; , \; 1 \leq k \leq m \quad . \tag{3}$$

For example, we write the permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 3 & 4 & 5 & 1 & 7 & 6 & 8 \end{pmatrix} \in S_8$$

in the form

$$\pi = (1\ 2\ 3\ 4\ 5)(6\ 7)(8) = (1\ 2\ 3\ 4\ 5)(6\ 7) \quad . \tag{4}$$

It may seem a little unpleasant that, for example, $(1\,2\,3\,4\,5)(6\,7)$ can be considered as a permutation in S_n for any $n \geq 7$, since the total number of elements in the set is not indicated in the notation; but when n is fixed in a given context, there is no ambiguity.

Moreover, we claim that the decomposition of a permutation into a product of cycles is unique. Suppose that we have another decomposition $\pi = \alpha_1 \alpha_2 \cdots \alpha_r$ into a product of disjoint cycles, and let i be an element which is not fixed by π. Then $\pi_s(i) \neq i$ and $\alpha_t(i) \neq i$ for one (and only one) of the $\pi_1, \cdots, \pi_m$ and one and only one of the $\alpha_1, \cdots, \alpha_r$. We have

$$\pi_s^k(i) = \pi^k(i) = \alpha_t^k(i) \quad , \quad k = 0, 1, 2, \cdots \quad .$$

But a cycle is uniquely determined by the action of all of its powers on any one element which it does not leave fixed. Thus, $\pi_s = \alpha_t$. We continue in this way, using induction on m (or r) to show that the cycles occurring in the two decompositions are the same. We have proved:

THEOREM 4. Every permutation $\pi \neq e$ in S_n is a product of disjoint cycles of length ≥ 2. This decomposition is uniquely determined except for the order of the cycles. □

This decomposition (3) is convenient for many reasons. For example, it makes it easy to find the order of a permutation.

COROLLARY 1. The order of a permutation $\pi \in S_n$ (i.e., the order of the cyclic subgroup $\langle \pi \rangle$) is equal to the least common multiple of the lengths of the disjoint cycles in the decomposition of π.

Proof. As noted before, the disjoint cycles in the decomposition of π commute with one another. Hence, by relation (4) in §1,

$$\pi^s = \pi_1^s \cdots \pi_m^s \quad , \quad s = 0, 1, 2, \cdots \quad .$$

Since the cycles $\pi_1, \cdots, \pi_m$ are independent (they act on disjoint sets $\Omega_1, \cdots, \Omega_m$), it follows that $\pi^q = e \Leftrightarrow \pi_k^q = e$ for $k = 1, \cdots, m$. Thus, q is a multiple of all of the orders of the cycles π_k, which, we have seen, coincide with their lengths ℓ_k. If q is the least natural number for which $\pi^q = e$, then $q = \text{Card} \langle \pi \rangle$ and

$q = \text{l. c. m.}(\ell_1, \cdots, \ell_m)$ is the integer defined in subsection 2 of §8 Ch. 1 (see also the proposition at the end of subsection 3). □

As an example, we can immediately say that the permutation (4) has order 10. As another example, suppose we wanted to know the maximum possible order of an element in S_8? If we go through all possible ways that 8 can be written as a sum of positive integers (in non-increasing order), we see that the following numbers occur as orders of elements $\neq e$ in S_8: 2,3,4, 5, 6, 7, 8, 10, 12, 15. An example of a permutation of maximal order 15 is (1 2 3 4 5)(6 7 8).

Definition. A cycle of length 2 is called a transposition.

A transposition has the form $\tau = (ij)$; it leaves fixed all elements besides i and j. Theorem 4 implies the following

COROLLARY 2. Every permutation $\pi \in S_n$ is a product of transpositions.

Proof. By Theorem 4, it suffices to write each cycle as a product of transpositions. But this can be done, for example, by taking

$$(1\,2 \ldots \ell\text{-}1\ \ell) = (1\,\ell)(1\ \ell\text{-}1) \ldots (13)(12) \quad . \quad \square$$

Corollary 2 can be expressed using the notion of a set of generators of a group (see subsection 2):

$$S_n = \langle (12), \ldots, (1n), (23), \ldots, (2n), \ldots, (n\text{-}1\ n) \rangle \quad .$$

Of course, this set of generators is not a minimal set. For example,

$$S_3 = \langle (12), (13), (23) \rangle = \langle (12), (13) \rangle \quad .$$

Notice that we cannot hope for any uniqueness assertion about expressing a permutation as a product of transpositions: in general, transpositions do not commute with one another, and the number of transpositions which appear when we write a permutation as a product of transpositions is not fixed. For example, in S_4 we have

$$(123) = (13)(12) = (23)(13) = (13)(24)(12)(14) \quad .$$

In fact, the non-uniqueness of the transposition decomposition is immediately clear if we note than $\sigma\tau^2 = \sigma$ for any transpositions σ and τ. Nevertheless, there is one thing about the transposition decomposition which is fixed. In order to discover it in the most natural possible way, we consider the action of S_n on functions.

Definition. Suppose that $\pi \in S_n$ and $f(X_1, \cdots, X_n)$ is a function of n variables. Set

$$(\pi \circ f)(X_1, \cdots, X_n) = f\left(X_{\pi^{-1}(1)}, \cdots, X_{\pi^{-1}(n)}\right) . \tag{5}$$

We say that the function $g = \pi \circ f$ is obtained by letting π act on f.

For example, if $\pi = (123)$ and $f(X_1, X_2, X_3) = X_1 + 2X_2^2 + 3X_3^3$, then $(\pi \circ f)(X_1, X_2, X_3) = X_3 + 2X_1^2 + 3X_2^3$.

As in §1 of Ch. 3, a function f is called skew-symmetric if $\tau \circ f = -f$ for every transposition τ, i.e.,

$$f(\ldots, X_j, \ldots, X_i, \ldots) = -f(\ldots, X_i, \ldots, X_j, \ldots) .$$

LEMMA. Let α and β be permutations in S_n. Then

$$(\alpha\beta) \circ f = \alpha \circ (\beta \circ f) .$$

Proof. Using the definition (5), we have

$$((\alpha\beta) \circ f)(X_1, \ldots, X_n) = f\left(X_{(\alpha\beta)^{-1}(1)}, \ldots, X_{(\alpha\beta)^{-1}(n)}\right) =$$

$$= f\left(X_{(\beta^{-1}\alpha^{-1})(1)}, \ldots, X_{(\beta^{-1}\alpha^{-1})(n)}\right) =$$

$$= f\left(X_{\beta^{-1}(\alpha^{-1}1)}, \ldots, X_{\beta^{-1}(\alpha^{-1}n)}\right) =$$

$$= (\beta \circ f)\left(X_{\alpha^{-1}(1)}, \ldots, X_{\alpha^{-1}(n)}\right) = (\alpha \circ (\beta \circ f))(X_1, \ldots, X_n) . \qquad \square$$

THEOREM 5. Let π be a permutation in S_n, and let

$$\pi = \tau_1 \tau_2 \cdots \tau_k \tag{6}$$

be any decomposition of π into a product of transpositions. Then the number

$$\varepsilon_\pi = (-1)^k , \tag{7}$$

which is called the "parity" (or "signature" or "sign") of π is completely determined by π, and does not depend on which decomposition (6) is used, i.e., the parity of k is always the same for a given π. In addition,

$$\varepsilon_{\alpha\beta} = \varepsilon_\alpha \varepsilon_\beta \tag{8}$$

for all $\alpha, \beta \in S_n$.

Proof. Take any skew-symmetric function f in n variables $X_1, \cdots, X_n$. By the lemma, the action of π on f reduces to the composition of the actions of the transpositions $\tau_k, \tau_{k-1}, \cdots, \tau_1$, i.e., to multiplication by -1 k times:

$$\pi \circ f = (\tau_1 \cdots \tau_{k-1}) \circ (\tau_k \circ f) = -(\tau_1 \cdots \tau_{k-1}) \circ f = \cdots = (-1)^k f = \varepsilon_\pi f .$$

Since the left side of this equality depends only on π (not on what choice of decomposition (6) is taken), it follows that the map $\varepsilon : \pi \mapsto \varepsilon_\pi$ defined by (7) must be completely determined by π, provided, of course, that f is not the identically zero function. But we know that there exist skew-symmetric functions which are not identically zero; for example, the Vandermonde determinant $\Delta_n(X_1, \cdots, X_n)$ is such a function.

Next, if we apply the permutation $\alpha\beta$ to f and use the rule in the lemma, we obtain

$$\varepsilon_{\alpha\beta} f = (\alpha\beta) \circ f = \alpha \circ (\beta \circ f) = \alpha \circ (\varepsilon_\beta f) = \varepsilon_\beta (\alpha \circ f) = \varepsilon_\beta (\varepsilon_\alpha f) = (\varepsilon_\alpha \varepsilon_\beta) f ,$$

which gives us (8). □

Definition. A permutation $\pi \in S_n$ is called even if $\varepsilon_\pi = 1$ and odd if $\varepsilon_\pi = -1$.

Thus, every transposition is an odd permutation.

COROLLARY 1. The even permutations in S_n form a subgroup A_n of order $|A_n| = n!/2$ (called the alternating group on n elements).

Proof. By (8), we see that $\varepsilon_{\alpha\beta} = 1$ if $\varepsilon_\alpha = \varepsilon_\beta = 1$, and $\varepsilon_{\pi^{-1}} = \varepsilon_\pi$, since $\varepsilon_e = 1$. It is then easy to see that all of the group axioms are fulfilled in A_n.

We write S_n as the disjoint union $S_n = A_n \cup \overline{A}_n$, where $\overline{A}_n$ is the set of all odd permutations in S_n. The map of S_n to itself defined by the rule

$$\rho_{(12)} : \pi \mapsto (12)\pi$$

is bijective. (Namely, it is injective, since $(12)\alpha = (12)\beta \Rightarrow \alpha = \beta$; then apply Theorem 3 of §5 Ch. 1. Or else, simply observe that $(\rho_{(12)})^2$ is the identity map, from which bijectivity follows.) Since $\varepsilon_{(12)\pi} = \varepsilon_{(12)}\varepsilon_\pi = -\varepsilon_\pi$, we have $\rho_{(12)} A_n = \overline{A}_n$, $\rho_{(12)} \overline{A}_n = A_n$. Thus, the number of even permutations in S_n coincides with the number of odd permutations; hence, $|A_n| = \frac{1}{2}|S_n| = \frac{n!}{2}$. □

COROLLARY 2. Suppose that we write $\pi \in S_n$ as a product of disjoint cycles of length $\ell_1, \ell_2, \cdots, \ell_m$. Then

$$\varepsilon_\pi = (-1)^{\sum_{k=1}^{m} (\ell_k - 1)}.$$

Proof. By Theorem 5, we have $\varepsilon_\pi = \varepsilon_{\pi_1 \cdots \pi_m} = \varepsilon_{\pi_1} \cdots \varepsilon_{\pi_m}$. We also have $\varepsilon_{\pi_k} = (-1)^{\ell_k - 1}$, since π_k can be written as a product of $\ell_k - 1$ transpositions (see the proof of Corollary 2 of Theorem 4). We conclude:

$$\varepsilon_\pi = (-1)^{\ell_1 - 1} \cdots (-1)^{\ell_m - 1} = (-1)^{\sum_{k=1}^{m} (\ell_k - 1)}. \qquad \square$$

We end this section by taking a break from serious things and considering the game "fifteen". Fifteen numbered flat square markers, all of the same size, are placed on a

square board so as to occupy all but one of 16 squares of the same size. The free square can be used

i_1	i_2	i_3	i_4
i_5	i_6	i_7	i_8
i_9	i_{10}	i_{11}	i_{12}
i_{13}	i_{14}	i_{15}	

(a)

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

(b)

Fig. 12

to move the markers horizontally or vertically (without lifting them from the board). Given an arbitrary set-up for the markers (see Fig. 12(a); we may suppose that we start with the free square in the lower right corner), we are required to move them to the set-up in Fig. 12(b). When is this possible? Elementary group theory killed this game when it was at the "height of fashion". Namely, we associate a permutation $\pi \in S_{15}$ to the diagrams 12(a) and (b). It is not hard to see (and it is worthwhile to really convince yourself of this) that it is possible to move the markers to the desired position if and only if the parity ε_π of the permutation π is 1, i.e., if and only if $\pi \in A_{15}$.

EXERCISES

1. Show that if $M = \langle S \rangle$ is the monoid generated by a set S, and if every element $s \in S$ has an inverse in M, then M is a group.

2. Prove that a group is a monoid G in which all equations of the form $ax = b$ or $ya = b$ have a unique solution, if $a, b \in G$.

3. Show that the set $A_1(\mathbb{R})$ of so-called affine transformations $\varphi_{a,b} : x \mapsto ax + b$ $(a, b \in \mathbb{R};\ a \neq 0)$ of the real line $\mathbb{R}$ is a group with multiplication law $\varphi_{a,b}\varphi_{c,d} = \varphi_{ac,\,ad+b}$. The group $A_1(\mathbb{R})$ contains the subgroup $GL(1, \mathbb{R})$ of affine transformations which leave the point $x = 0$ fixed, and the subgroup of "translations" $x \mapsto x + b$.

4. The group $SL(2, \mathbb{Z})$ contains the elements $A = \begin{Vmatrix} 0 & 1 \\ -1 & 0 \end{Vmatrix}$ and $B = \begin{Vmatrix} 0 & 1 \\ -1 & -1 \end{Vmatrix}$ of orders 4 and 3, respectively. Show that $\langle AB \rangle$ is an infinite cyclic subgroup of $SL(2, \mathbb{Z})$. This shows that the product of two elements of finite order in a group is not necessarily of finite order. What happens in an abelian group?

5. Prove that a group G of even order $|G| = 2n$ must contain an element $g \neq e$ of order 2.

6. Prove that $S_n = \langle (12), (13), \cdots, (1n) \rangle$.

7. Prove that $S_n = \langle (12), (1234 \ldots n) \rangle$.

8. Prove that the alternating group A_n, $n \geq 3$, is generated by cycles of length 3; more precisely, prove that

$$A_n = \langle (123), (124), \cdots, (12n) \rangle \quad .$$

9. Find the sign of the permutation

$$\pi = \begin{pmatrix} 1 & 2 & 3 & \ldots & n-1 & n \\ n & n-1 & n-2 & \ldots & 2 & 1 \end{pmatrix} \quad .$$

10. Let $\Omega = \{1, 2, \cdots, n\}$, and let $\Omega \times \Omega$ be the cartesian square. Call a pair $(i, j) \in \Omega \times \Omega$ an <u>inversion relative to the permutation</u> $\sigma \in S_n$ (or simply a <u>σ-inversion</u>) if $i < j$ but $\sigma(i) > \sigma(j)$. Set

$$\operatorname{sgn} \sigma = \prod_{1 \leq i < j \leq n} \frac{\sigma(j) - \sigma(i)}{j - i} \quad .$$

Since $(\sigma(j) - \sigma(i))/(j - i)$ is a non-zero rational number which is negative if and only if (i, j) is a σ-inversion, and since $\sigma : \Omega \to \Omega$ is a bijective map, it follows that $\operatorname{sgn} \sigma = (-1)^k$, where k is the total number of σ-inversions. If $\tau = (ij)$ is a transposition, then $\operatorname{sgn} \tau = -1$. It is easy to see that

$$(\sigma(j)\, \sigma(i))\sigma = \begin{pmatrix} \ldots & \sigma(j) & \ldots & \sigma(i) & \ldots \\ \ldots & \sigma(i) & \ldots & \sigma(j) & \ldots \end{pmatrix} \begin{pmatrix} \ldots & i & \ldots & j & \ldots \\ \ldots & \sigma(i) & \ldots & \sigma(j) & \ldots \end{pmatrix} = \begin{pmatrix} \ldots & i & \ldots & j & \ldots \\ \ldots & \sigma(j) & \ldots & \sigma(i) & \ldots \end{pmatrix} ,$$

so that a σ-inversion (ij) is not an inversion relative to the permutation $\tau\sigma$, where τ is the transposition $(\sigma(j)\,\sigma(i))$. Show that it is possible to find k transpositions $\tau_1, \cdots, \tau_k$ such that $\tau_k\tau_{k-1}\cdots\tau_1\sigma = e$ is the identity. Then $\sigma = \tau_1 \cdots \tau_{k-1}\tau_k$, and $\operatorname{sgn}\sigma = (-1)^k = \varepsilon_\sigma$, so we have two designations for the same thing: sgn and ε. This gives us another useful method for determining the sign of a permutation. Suppose that the set of inversions relative to a permutation π consists of five pairs $(1,5)$, $(2,5)$, $(3,5)$, $(4,5)$, $(6,7)$; then $\operatorname{sgn}\pi = -1$. In practice, this method amounts to counting the number of j in the lower row of a permutation which are greater than i but come before i, as $i = 1, 2, \cdots, n-1$.

11. Prove that a non-empty subset H of a finite (multiplicative) group G is a subgroup if H is closed under multiplication. That is, in that case the requirement that H has an identity element and an inverse h^{-1} for each $h \in H$ is superfluous.

12. Give a possible set of generators for the multiplicative group of positive rational numbers.

13. Prove that the k-th power π^k of the cycle $\pi = (12 \ldots n) \in S_n$ is the product of $d = \text{g. c. d.}(n,k)$ independent cycles, each of length $q = \text{l. c. m.}(n,k) = n/d$.

14. Suppose that $A, B \in M_n(\mathbb{R})$ and $(AB)^m = E$ for some integer m. Is it necessarily true that $(BA)^m = E$?

§3. Morphisms of groups

1. Isomorphisms. As already noted, the three rotations $\varphi_0, \varphi_1, \varphi_2$ counterclockwise through 0°, 120°, 240° take the equilateral triangle P_3 to itself. But there are also three reflections about an axis of symmetry, which we denote ψ_1, ψ_2, ψ_3; in Fig. 13, the axes of symmetry are $1\text{--}1'$, $2\text{--}2'$, and $3\text{--}3'$. For each of the six transformations of P_3 there is a corresponding permutation of the set of vertices of the triangle. We have the correspondence

$$\varphi_0 \sim e\,, \qquad \varphi_1 \sim (123)\,, \quad \varphi_2 \sim (132)$$

$$\psi_1 \sim (23)\,, \quad \psi_2 \sim (13)\,, \quad \psi_3 \sim (12)$$

Since this exhausts all of S_3, we can say that the group D_3 of all symmetry transformations of an equilateral triangle is very much like the symmetric group S_3.

In the same sense, the two groups C_n (the cyclic group of order n, see the example in subsection 3 of §2) and $\langle(12\ldots n)\rangle \subset S_n$ are very similar to one another. These examples, along with some general thought about the nature of groups, inevitably lead to a very natural question about the most essential properties of groups. At first glance, complete information is contained in the multiplication table for a group G, sometimes called the Cayley table:

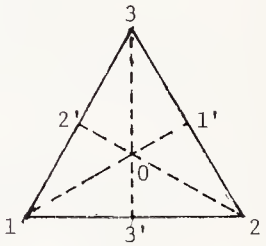


Fig. 13

	g_1	g_2	$\cdots$	g_n	$\cdots\cdot$
g_1	g_1g_1	g_1g_2	$\cdots$	g_1g_n	$\cdots\cdot$
g_2	g_2g_1	g_2g_2	$\cdots$	g_2g_n	
$\vdots$	$\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot$				
g_n	g_ng_1	g_ng_2	$\cdots$	g_ng_n	$\cdots\cdot$
$\vdots$	$\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot\cdot$				

And, in fact, many properties of a group can be seen from the Cayley table, i.e., from the $n \times n$-matrix $M = (m_{ij})$, where $n = (G:e)$, with entries $m_{ij} = g_ig_j \in G$. We note, for example, that in each row or column of M every element of G occurs exactly once (see the proof of Theorem 2 below). A group G is abelian if and only if M is symmetric, i.e., $m_{ij} = m_{ji}$. There are other such properties, but we soon see that it is rather hard to compare two multiplication tables, say for groups G and G', because the appearance of M depends on how the elements of the group are indexed; moreover, the situation becomes more complicated when the groups are infinite.

The best and most profound approach to comparing two groups is based on the notion of an isomorphism.

Definition. Two groups G and G' with operations $*$ and $\circ$ are called isomorphic if there exists a map $f : G \to G'$ such that:

(i) $f(a * b) = f(a) \circ f(b)$ for all $a, b \in G$.

(ii) f is bijective.

Such a map f is called an isomorphism between G and G'. We use the notation $G \cong G'$ when G and G' are isomorphic.

We give the simplest properties of an isomorphism.

1) The identity goes to the identity. To see this, let e be the identity of G. Since $e * a = a * e = a$, we have $f(e) \circ f(a) = f(a) \circ f(e) = f(a)$ for all $a \in G$; hence, $f(e) = e'$ is the identity in G'. Note that, in addition to property (i), we have used property (ii), since surjectivity of f is needed to be able to write any element of G' in the form $f(a)$. □

2) $f(a^{-1}) = f(a)^{-1}$. Namely, by 1), we have $f(a) \circ f(a^{-1}) = f(a * a^{-1}) = f(e) = e'$, where e' is the identity in G'. Thus,

$$f(a)^{-1} = f(a)^{-1} \circ e' = f(a)^{-1} \circ (f(a) \circ f(a^{-1})) =$$

$$= (f(a)^{-1} \circ f(a)) \circ f(a^{-1}) = e' \circ f(a^{-1}) = f(a^{-1}) . \quad \square$$

3) The inverse map $f^{-1} : G' \to G$ (which exists by property (ii)) is also an isomorphism.

By the corollary of Theorem 2 of §5 Ch. 1, we need only verify property (i) for f^{-1}. Let $a', b' \in G'$. Then, since f is bijective, we can write $a' = f(a)$, $b' = f(b)$ for some $a, b \in G$. Because f is an isomorphism, we have $a' \circ b' = f(a) \circ f(b) = f(a * b)$. This gives $a * b = f^{-1}(a' \circ b')$, and, since $a = f^{-1}(a')$ and $b = f^{-1}(b')$, it follows that $f^{-1}(a' \circ b') = f^{-1}(a') * f^{-1}(b')$. □

A simple verification shows that our correspondence $\sim$ between the groups D_3 and S_3 in the example above is actually an isomorphism.

The function $f = \ln$ is an isomorphism of the multiplicative group of positive real numbers with the additive group of all real numbers. The basic property of the logarithm $\ln ab = \ln a + \ln b$ is precisely property (i) in the definition of an isomorphism. The inverse mapping is $x \mapsto e^x$.

We now prove two general theorems which illustrate the role of isomorphisms in group theory.

THEOREM 1. Any two cyclic groups of the same order (in particular, any two infinite cyclic groups) are isomorphic.

Proof. First, if $\langle g \rangle$ is an infinite cyclic group, then all powers g^n are distinct, and we obtain an isomorphism $f : \langle g \rangle \to (\mathbb{Z}, +)$ by setting $g^n \mapsto n$. f is obviously bijective, and the property $f(g^m g^n) = f(g^m) + f(g^n)$ follows from Theorem 2 of §2.

Now suppose that $G = \{e, g, \dots, g^{q-1}\}$ and $G' = \{e', g', \dots, g'^{q-1}\}$ are two cyclic groups of order q (where we are using multiplicative notation for the group operation in both groups). We define a bijective map by setting

$$f : g^k \longmapsto (g')^k, \qquad k = 0, 1, \dots, q-1 .$$

Setting $n + m = \ell q + r$, $0 \le r \le q - 1$, for any $n, m = 0, 1, \dots, q-1$ and reasoning as in the proof of Theorem 3 of §2, we have

$$f(g^{n+m}) = f(g^r) = (g')^r = (g')^{n+m} = (g')^n (g')^m = f(g^n)\, f(g^m) . \qquad \square$$

THEOREM 2 (Cayley). Any finite group of order n is isomorphic to some subgroup of the symmetric group S_n.

Proof. Let G be a group with $n = |G|$. We may take S_n to be the group of all bijective maps of the set G to itself, since the nature of the elements permuted by the elements of S_n is immaterial.

For any $a \in G$, consider the map $L_a : G \to G$ given by the formula

$$L_a(g) = ag \quad .$$

If $e = g_1, g_2, \cdots, g_n$ are all elements of G, then $a, ag_2, \cdots, ag_n$ are the same elements but in a different order (recall the Cayley table!). (To see why these elements are distinct, we have:

$$ag_i = ag_j \Rightarrow a^{-1}(ag_i) = a^{-1}(ag_j) \Rightarrow (a^{-1}a)g_i = (a^{-1}a)g_j \Rightarrow g_i = g_j \quad .)$$

Hence, L_a is a bijective map (permutation), whose inverse will be $L_a^{-1} = L_{a^{-1}}$. Of course, L_e is the identity permutation. If we use the associativity of the group operation, we obtain: $L_{ab}(g) = (ab)g = a(bg) = L_a(L_b g)$, i.e., $L_{ab} = L_a \circ L_b$.

Thus, the set $L_e, L_{g_2}, \cdots, L_{g_n}$ forms a subgroup -- call it H -- in the group of all bijective maps of G to itself, i.e., in S_n. The group G is isomorphic to the subgroup H using the correspondence $a \mapsto L_a$, which, by what was said in the last paragraph, has all of the properties of an isomorphism. □

Despite its simplicity, Cayley's theorem has an important meaning for group theory. It shows the existence of a sort of "universal object" -- the family $\{S_n \mid n = 1, 2, \ldots\}$ of symmetric groups -- in which all finite groups (considered up to isomorphism) live. The phrase "up to isomorphism" is so typical, not only of group theory, but of all mathematics, which tends to consider at once all objects having common properties; without such abstraction and generalization, the whole point of the subject would be lost.

If $G' = G$ in the definition of an isomorphism, we have the concept of an isomorphism $\varphi : G \to G$ of a group G to itself. Such an isomorphism is called an automorphism of G. For example, the identity map $e_G : g \mapsto g$ (henceforth denoted simply 1) is an automorphism. But, in general, a group G also has non-trivial automorphisms. Property 3) of isomorphisms shows that the inverse of an automorphism is also an automorphism. Furthermore, if φ and ψ are automorphisms of G, then

$(\varphi \circ \psi)(ab) = \varphi(\psi(ab)) = \varphi(\psi(a)\,\psi(b)) = (\varphi \circ \psi)(a) \cdot (\varphi \circ \psi)(b)$ for any $a, b \in G$. Hence, the set $\mathrm{Aut}(G)$ of all automorphisms of a group G forms a group, in fact, a subgroup of the group $S(G)$ of all bijective maps $G \to G$.

2. Homomorphisms. The group of automorphisms $\mathrm{Aut}(G)$ of a group G contains a very special subgroup, which is denoted $\mathrm{Inn}(G)$ and is called the group of inner automorphisms. Its elements are the maps

$$I_a : g \mapsto a g a^{-1} \quad .$$

A simple exercise shows that I_a really has all of the properties required of an automorphism, and that $I_a^{-1} = I_{a^{-1}}$, $I_e = 1$ is the identity automorphism, and $I_a \circ I_b = I_{ab}$ (because

$$(I_a \circ I_b)(g) = I_a(I_b(g)) = I_a(bgb^{-1}) = abgb^{-1}a^{-1} = abg(ab)^{-1} = I_{ab}(g) \ .)$$

This last fact about I_a shows that the map

$$f : G \to \mathrm{Inn}(G), \quad f(a) = I_a \quad \text{for} \quad a \in G \ ,$$

from the group G to the group $\mathrm{Inn}(G)$ of inner automorphisms of G satisfies property (i) in the definition of an isomorphism: $f(a) \circ f(b) = f(ab)$. However, property (ii) is not necessarily satisfied. For example, if G is an abelian group, then $a g a^{-1} = g$ for all $a, g \in G$, i.e., $I_a = I_e$ for all $a \in G$, and $\mathrm{Inn}(G)$ only consists of the identity I_e. The example of this map f makes it natural to introduce the following

Definition. A map $f : G \to G'$ from the group $(G, *)$ to the group $(G', \circ)$ is called a homomorphism if

$$f(a * b) = f(a) \circ f(b), \qquad \forall\, a, b \in G$$

(in other words, property (ii) in the definition of an isomorphism is omitted).

By the kernel of the homomorphism f, we mean the set

$$\mathrm{Ker}\, f = \{g \in G \mid f(g) = e', \text{ where } e' \text{ is the identity of } G'\} \quad .$$

A homomorphism from a group to itself is called an endomorphism.

In the definition of a homomorphism, f need be neither injective nor surjective. We can make f into a surjective map by replacing G by $\operatorname{Im} f \subset G'$, which is obviously a subgroup of G'. So the "main" difference between a homomorphism and an isomorphism is the presence of a non-trivial kernel $\operatorname{Ker} f$, which is, one might say, a measure of the non-injectivity of f. If $\operatorname{Ker} f = \{e\}$, then $f : G \to \operatorname{Im} f$ is an isomorphism.

Note that

$$f(a) = e',\ f(b) = e' \Rightarrow f(a * b) = f(a) \circ f(b) = e' \circ e' = e'$$

and

$$f(a^{-1}) = f(a)^{-1} = (e')^{-1} = e' \quad .$$

Hence $\operatorname{Ker} f$ is a subgroup of G. Let $H = \operatorname{Ker} f \subset G$. Then (we are now omitting the $*$ and $\circ$):

$$f(ghg^{-1}) = f(g)\, f(h)\, f(g)^{-1} = f(g)\, e'\, f(g)^{-1} = e', \qquad \forall\, h \in H,\ g \in G,$$

i.e., $ghg^{-1} \in H$; hence, $gHg^{-1} \subset H$. If we replace g by g^{-1} here, we obtain $g^{-1}Hg \subset H$, so that $H \subset gHg^{-1}$. Thus,

$$gHg^{-1} = H, \qquad \forall\, g \in G \quad .$$

A subgroup which has this property is called a normal subgroup (sometimes an invariant subgroup or a normal divisor). We have thereby proved

THEOREM 3. The kernel of a homomorphism is always a normal subgroup. □

We shall see the importance of this fact somewhat later. For now, we note that far from every subgroup is normal. For example, in S_3 the cyclic subgroup $\langle (123) \rangle = A_3$ is normal, but $\langle (12) \rangle = \{e, (12)\}$ is not normal.

3. Glossary. Examples. The terms "surjective map" (map "onto"), "injective map" (imbedding), "bijective map" (one-to-one correspondence), which can be used for maps of any sets (with or without operations), are often replaced by other terms when used for groups

(the same happens for other algebraic systems). We use the terms epimorphism (homomorphism "onto"), monomorphism (homomorphism whose kernel is the identity element), and isomorphism (a one-to-one homomorphism, i.e., a homomorphism which is both an epimorphism and a monomorphism). There is a tendency to replace the term homomorphism by the word morphism. These words are useful to know when reading mathematical literature, but the reader can, if he chooses, get along using only the terms isomorphism and homomorphism with the prepositions "into" and "onto".

We now give some further examples of group homomorphisms.

1) The additive group of integers $\mathbb{Z}$ maps homomorphically onto the finite cyclic group $\langle g \rangle$ of order q if we set $f : n \mapsto g^n$ (see Theorem 2 of §2). Here $\operatorname{Ker} f = \{\ell q \mid \ell \in \mathbb{Z}\}$. Namely, it is clear that $\{\ell q\} \subset \operatorname{Ker} f$, and the reverse inclusion follows from Theorem 3 of §2.

2) The map $f : \mathbb{R} \to T = SO(2)$ of the additive group of real numbers onto the group T of rotations of the plane about the origin, which is given by $f(\lambda) = \Phi_\lambda$ (Φ_λ is counterclockwise rotation through an angle of $2\pi\lambda$) is a homomorphism. Since rotation through a multiple of 2π coincides with rotation through an angle of zero, i.e., the identity map, we have: $\operatorname{Ker} f = \mathbb{Z}$. We also say that f gives a homomorphism of $\mathbb{R}$ onto the unit circle S^1, since there is a one-to-one correspondence between $SO(2)$ and S^1, namely, Φ_λ corresponds to the point with polar coordinates $(1, 2\pi\lambda)$.

3) The general linear group $GL(n)$, consisting of matrices A with entries in $\mathbb{R}$ and non-zero determinant, maps homomorphically onto the multiplicative group $\mathbb{R}^*$ of non-zero real numbers, if we set $f = \det$. The condition that f is a homomorphism: $f(AB) = f(A)\,f(B)$, is merely another way of stating Theorem 5 of §2 Ch. 3. By definition, $SL(n) = \operatorname{Ker} f$.

4) Consider the cyclic group $C_2 = \langle -1 \rangle = \{1, -1\}$ of order 2. If we want, we can give this group by writing its Cayley table:

C_2 :

.	1	-1
1	1	-1
-1	-1	1

The map $S_n \to C_2$ given by our function $\varepsilon = \operatorname{sgn} : \pi \mapsto \varepsilon_\pi$ is a homomorphism. Here $\operatorname{Ker} \varepsilon = A_n$, by the definition of the alternating group.

5) An infinite group can be isomorphic to a proper subgroup. For example, the additive group $(\mathbb{Z}, +)$ contains the proper subgroup $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$, where $n > 1$ is a fixed natural number. It is easy to check that the map $g_n : \mathbb{Z} \to n\mathbb{Z}$ given by $g_n(k) = nk$ is an isomorphism. Incidentally, note that $\mathbb{Z}$ and $n\mathbb{Z}$ are infinite cyclic groups with generator 1 or -1 and n or $-n$, respectively; hence g_n and the map $k \mapsto -kn$ are all possible isomorphisms $\mathbb{Z} \to n\mathbb{Z}$.

6) The group $\operatorname{Aut}(G)$, and even a single element $\varphi \in \operatorname{Aut}(G)$ which is not the identity, can be a source of important information about the group G. Here is an example of this. Let G be a finite group having an automorphism φ of order 2 (i.e., $\varphi^2 = 1$) which has no fixed points:

$$a \neq e \Longrightarrow \varphi(a) \neq a \quad .$$

Suppose that $\varphi(a)\, a^{-1} = \varphi(b)\, b^{-1}$ for any $a, b \in G$. If we multiply this equality on the left by $\varphi(b)^{-1}$ and on the right by a, we see that $\varphi(b)^{-1} \varphi(a) = b^{-1} a$, i.e., $\varphi(b^{-1} a) = b^{-1} a$, which implies that $b^{-1} a = e$, and so $b = a$. Thus, as a runs through the elements of G, so does $\varphi(a) a^{-1}$; equivalently, any element $g \in G$ can be written in the form $g = \varphi(a) a^{-1}$. But then $\varphi(g) = \varphi(\varphi(a))\, \varphi(a^{-1}) = \varphi^2(a)\, \varphi(a^{-1}) = a \varphi(a)^{-1} =$ $= (\varphi(a) a^{-1})^{-1} = g^{-1}$. Thus, φ is precisely the map $g \mapsto g^{-1}$. With this in mind, we obtain $ab = \varphi(a^{-1}) \varphi(b^{-1}) = \varphi(a^{-1} b^{-1}) = (a^{-1} b^{-1})^{-1} = ba$, i.e., the group G turns out to be abelian! In addition, $(G : e)$ is an odd number, since G consists of e and disjoint pairs of elements $g_i, g_i^{-1} = \varphi(g_i)$.

7) The following example shows how much one can alter the group operation without

changing the group itself, i.e., only changing it to an isomorphic group (see also Exercise 3 of §1). Let G be any group, and let t be a fixed element of G. Introduce a new operation on the set G:

$$(g, h) \longmapsto g * h = gth \quad .$$

We immediately verify that $(g_1 * g_2) * g_3 = g_1 * (g_2 * g_3)$, i.e., the operation $*$ is associative. In addition, $g * t^{-1} = t^{-1} * g = g$, and $g * (t^{-1} g^{-1} t^{-1}) = (t^{-1} g^{-1} t^{-1}) * g = t^{-1}$; this means that $(G, *)$ is a group with identity element $e_* = t^{-1}$. The inverse of an element g in $(G, *)$ is $g_*^{-1} = t^{-1} g^{-1} t^{-1}$. The map $f : g \mapsto gt^{-1}$ gives an isomorphism between $(G, \cdot)$ and $(G, *)$.

All of these examples illustrate the general principle that studying the morphisms of a group G gives much information about G itself.

4. Cosets of a subgroup. It is clear from the definition of a homomorphism $f : G \to G'$ that all of the elements of the set

$$a \operatorname{Ker} f = \{ab \mid b \in \operatorname{Ker} f\}, \qquad a \in G \quad ,$$

are mapped to the same element $f(a)$ in G': $f(ab) = f(a) f(b) = f(a)e' = f(a)$ if $b \in \operatorname{Ker} f$. Conversely, if $f(g) = f(a)$, then $f(a^{-1} g) = f(a^{-1}) f(g) = f(a)^{-1} f(g) = e'$, so that $a^{-1} g = b \in \operatorname{Ker} f$ and $g = ab \in a \operatorname{Ker} f$. This fact shows the usefulness of partitioning G into subsets of the form $a \operatorname{Ker} f$. We now forget about homomorphisms for a moment, and study such partitions in their own right.

Definition. Let H be a subgroup of G. A left coset of H in G is a set of the form gH, consisting of all elements of the form gh, where g is a fixed element of G and h runs through all elements of the subgroup H. The element g is called a coset representative for gH.

We similarly define a right coset Hg. (Sometimes the terminology "right" and "left" is reversed, i.e., used to refer to the position of H rather than of g; the

important thing is to be consistent, whichever convention is adopted.) If $H = \operatorname{Ker} f$ is the kernel of a homomorphism, then $gH = Hg$, because H is normal in G (see subsection 2). Note that the subgroup H itself is a coset: $H = He = eH$. But none of the other cosets can be a subgroup, because, if gH were a subgroup, then we would have $e \in gH$, so that $e = gh$, $g = h^{-1}$, and $gH = h^{-1}H = H$.

THEOREM 4. Two left cosets of H in G must either coincide or be disjoint. The partition of G into left cosets of H gives an equivalence relation on G.

Proof. Suppose that the cosets g_1H and g_2H have an element in common: $a = g_1h_1 = g_2h_2$. Then $g_2 = g_1h_1h_2^{-1}$, and any element g_2h of the coset g_2H has the form $g_1h_1h_2^{-1}h = g_1h'$, where $h' = h_1h_2^{-1}h \in H$. Thus, $g_2H \subset g_1H$. We similarly prove that every element in g_1H is contained in g_2H. Hence, $g_1H = g_2H$.

Since each element $g \in G$ is contained in the coset gH, we may conclude that G is a union of disjoint left cosets of H:

$$G = \cup\, g_iH \quad .$$

By the general principle in §6 of Chapter 1, this partition induces an equivalence relation on G, which is defined in the obvious way:

$$a \sim b \iff a^{-1}b \in H \quad .$$

If we want, we can verify reflexivity, symmetry, and transitivity of this relation directly: $a \sim a$, because $a^{-1}a = e \in H$; $a \sim b \Rightarrow a^{-1}b = h \Rightarrow b^{-1}a = h^{-1} \in H \Rightarrow b \sim a$; $a \sim b$, $b \sim c \Rightarrow b^{-1}a = h_1$, $c^{-1}b = h_2 \Rightarrow c^{-1}a = c^{-1}bh_1 = h_2h_1 \in H \Rightarrow a \sim c$. □

The analogous theorem holds for right cosets.

The partition into cosets arises naturally in permutation groups. Let $G = S_n$ be the symmetric group, acting on the set $\Omega = \{1, 2, \dots, n\}$. If we consider the set H of elements $\pi \in S_n$ for which $\pi(n) = n$, then we easily see that H is a subgroup of S_n, which can be identified with S_{n-1}. Let $\tau_0 = e$, and let $\tau_i = (in)$ be the

transposition taking n to i $(i = 1, 2, \cdots, n-1)$. It is clear that

$$S_n = \bigcup_{k=0}^{n-1} \tau_k S_{n-1} \quad .$$

Here is the partition of S_3 into left and right cosets of the subgroup $\langle(12)\rangle = S_2$:

$$S_3 = \{e, (12)\} \cup \{(13), (123)\} \cup \{(23), (132)\} \ ;$$

$$S_3 = \{e, (12)\} \cup \{(13), (132)\} \cup \{(23), (123)\} \quad .$$

We see that the set of left cosets gS_2 is not the same as the set of right cosets S_2g'. Nevertheless, the sets $\{gH\}$ and $\{Hg'\}$ are always in bijective correspondence, as follows:

$$x = gh \in gH \longleftrightarrow x^{-1} = h^{-1}g^{-1} \in Hg^{-1} \quad .$$

In fact, if, say, $h_1 g_1^{-1} = h_2 g_2^{-1}$, then $g_1 = g_2 h_2^{-1} h_1$, and $g_1 H = g_2 H$. In particular, <u>if $\{e, x, y, z, \cdots\}$ is a set of left (respectively, right) coset representatives, then $\{e, x^{-1}, y^{-1}, z^{-1}, \cdots\}$ is a set of right (respectively, left) coset representatives. Both sets have the same cardinality.</u>

We denote the set of all left cosets of H in G by the symbol G/H (or $(G/H)_\ell$ if we have to consider simultaneously the set $(G/H)_r$ of right cosets). We call the cardinality of this set $\mathrm{Card}\, G/H$ the "index of the subgroup H in G", which we denote $(G:H)$. (This agrees with the notation $(G:e)$ introduced before for the order $|G|$ of G, i.e., the number of cosets of the trivial subgroup $\{e\}$.) Since we have the one-to-one correspondence $a \mapsto ga$ between H and gH (see the proof of Cayley's theorem), it follows that $\mathrm{Card}\, gH = (H:e)$. We thus have the following simple formula:

$$(G:e) = (G:H)\,(H:e) \quad ,$$

which implies the classical

THEOREM 5 (Lagrange). <u>The order of a finite group is divisible by the order of any subgroup.</u> □

COROLLARY. The order of any element divides the order of the group. A group whose order is a prime p is always cyclic; such a group is unique up to isomorphism.

Proof. The order of an element $g \in G$ is the same as the order of the cyclic subgroup generated by g (Theorem 3 of §2), and so divides $|G|$ by Theorem 5. Next, if $|G| = p$ is a prime, and if H is a non-trivial subgroup (i.e., $H \neq \{e\}$), then Theorem 5 implies that $|H| = p$, and hence $H = G$. Thus, G coincides with the cyclic subgroup generated by any element $g \neq e$. Since all cyclic groups of a given order are isomorphic (Theorem 1), we have the uniqueness assertion in the corollary. □

Lagrange's theorem leads one to want to find a subgroup of order m for every m dividing $n = |G|$. But this cannot always be done. For example, the reader can verify that the alternating group A_4, which has order 12, has no subgroup of order 6.

But in some groups the "converse Lagrange theorem" does hold. For example:

THEOREM 6. Every subgroup of a cyclic group is cyclic. The subgroups of the infinite cyclic group $(\mathbb{Z}, +)$ are precisely the (infinite cyclic) groups $(m\mathbb{Z}, +)$, $m \in \mathbb{N}$; and the subgroups of a cyclic group of order q are in one-to-one correspondence with the (positive) divisors d of q.

Proof. Let $A = \langle a \rangle$ be a cyclic group. For variety, let's use additive notation. Thus, every element has the form ka, where $k \in \mathbb{Z}$ or else $k = 0, 1, \ldots, q-1$ if A is a finite group of order q (see Theorem 3 of §2). Let B be a non-zero subgroup of A. If $ka \in B$ for some $k \neq 0$, then we also have $-ka \in B$. Among all elements $ka \in B$ with positive k, let ma be the element for which m is minimal. If we write any $k > 0$ in the form $k = \ell m + r$, $0 \leq r < m$, we see that $ka \in B$ implies $ra = ka - \ell(ma) \in B$; hence, $r = 0$. Thus, $B = \langle ma \rangle$ is a cyclic group.

All infinite cyclic groups are isomorphic (by Theorem 1). So without loss of generality we may take $(\mathbb{Z}, +)$ to be the model of an infinite cyclic group. It has generator 1 or -1, so that, by the last paragraph, any subgroup of $(\mathbb{Z}, +)$ is determined by a

natural number m and has the from

$$m\mathbb{Z} = \langle m \cdot 1 \rangle = \{0, \pm m, \pm 2m, \cdots\} \quad .$$

It is obvious that all of these subgroups are infinite.

Now suppose that $\langle a \rangle = \{0, a, \cdots, (q-1)a\}$, $qa = 0$. We found that $B = \{0, ma, 2ma, \cdots\}$, where $m \in \mathbf{N}$ and $sa \in B$ for $s \in \mathbf{N}$ only if s is a multiple of m. We claim that m divides q. In fact, let $q = dm + r$, $0 \leq r < m$. Then

$$0 = qa = d(ma) + ra \quad ,$$

so that $ra = -d(ma) \in B$. Since m is minimal, we have $r = 0$, and so $q = dm$. Thus,

$$B = \{0, ma, 2ma, \cdots, (d-1)ma\} = mA$$

is a subgroup of A of order d. As m runs through all positive divisors of q, so does d, and we obtain exactly one subgroup for each order d dividing q. □

COROLLARY. In a cyclic group $\langle a \rangle$ of order q, the subgroup of order $d \mid p$ is precisely the set of elements $b \in \langle a \rangle$ such that $db = 0$.

Proof. If $dm = q$, then $b \in B = mA$, and $db = 0$. Conversely, suppose that $b = \ell a \in \langle a \rangle$ and $db = 0$. The condition $d\ell a = 0$ implies that $d\ell = qk = dmk$, so that $\ell = mk$, and $b = \ell a = k(ma) \in mA$. □

5. The monomorphism $S_n \to GL(n)$. Recall that a monomorphism of groups $G \to G'$ is an injective homomorphism from G to G'.

For example, here is a monomorphism $f : S_3 \to GL(3)$:

$$e \longmapsto \begin{Vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{Vmatrix}, \quad (12) \longmapsto \begin{Vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{Vmatrix}, \quad (13) \longmapsto \begin{Vmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{Vmatrix},$$

$$(23) \longmapsto \begin{Vmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{Vmatrix}, \quad (123) \longmapsto \begin{Vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{Vmatrix}, \quad (132) \longmapsto \begin{Vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{Vmatrix}.$$

The reader can check that f is really a monomorphism, and that for each $\pi \in S_3$ the determinant of $f(\pi)$ is ± 1, depending on the signature of the permutation π.

THEOREM 7. <u>There exists a monomorphism</u> $f : S_n \to GL(n)$ <u>such that the matrix</u> $f(\pi)$, $\pi \in S_n$, <u>has determinant</u> $|f(\pi)| = \varepsilon_\pi$.

<u>Proof.</u> We shall write an $n \times n$ matrix (a_{ij}) as a sequence of columns: $(a_{ij}) = (A^{(1)}, A^{(2)}, \dots, A^{(n)})$. In particular, let

$$E^{(1)} = \begin{Vmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{Vmatrix}, \quad E^{(2)} = \begin{Vmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{Vmatrix}, \dots, \quad E^{(n)} = \begin{Vmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{Vmatrix}$$

be the columns of the identity matrix E. We define the map $f : S_n \to GL(n)$ by setting

$$\pi \longmapsto f(\pi) = (E^{(\pi(1))}, E^{(\pi(2))}, \dots, E^{(\pi(n))}). \qquad (1)$$

Thus, $f(\pi)$ is an $n \times n$ matrix in which each row and each column has one 1 and the other entries zero. It is easy to see that $f(\pi) \in GL(n)$.

Let σ and τ be any permutations, and let $\pi = \sigma\tau$ be their product. By definition, the non-zero entries in the i-th row of $f(\sigma) = (a_{is})$ and in the j-th column of $f(\tau) = (b_{k\ell})$ are $a_{i, \sigma^{-1}(i)} = 1$ and $b_{\tau(j), j} = 1$, respectively. Thus, in the matrix $f(\sigma) f(\tau) = (c_{ij})$, the condition $c_{ij} \neq 0$ is equivalent to $\sigma^{-1}(i) = \tau(j)$, i.e., $i = \sigma\tau(j) = \pi(j)$; but this means that $f(\sigma) f(\tau) = f(\sigma\tau)$. Consequently, f is a homomorphism.

The property $\operatorname{Ker} f = \{e\}$ is obvious, since it is clear from (1) that $f(\pi) = E \Rightarrow \pi = e$. Thus, f is a monomorphism.

Finally, since f, det, and ε all preserve products, i.e.,

$$|f(\sigma\tau)| = |f(\sigma)\ f(\tau)| = |f(\sigma)||f(\tau)|, \qquad \varepsilon_{\sigma\tau} = \varepsilon_\sigma\, \varepsilon_\tau\ ,$$

and since any π can be decomposed into a product of transpositions, it suffices to prove the equality $|f(\pi)| = \varepsilon_\pi$ when $\pi = (ij)$ is a transposition. But then $f(\pi)$ is obtained from E by interchanging the i-th and j-th columns; hence, $|f(\pi)| = -|E| = -1 = \varepsilon_{(ij)}$. □

Matrices of the form $f(\pi)$, $\pi \in S_n$, are called permutation matrices. The restriction of the monomorphism f to A_n is a monomorphism into $SL(n, \mathbb{R})$. Given any finite group G, the composition $f \circ L$ of the map $L : G \to S_n$ (see Theorem 2) and the map $f : S_n \to GL(n)$ gives a monomorphism $G \to GL(n)$.

Using Theorem 7, we can easily prove the so-called theorem on the complete expansion of the determinant:

THEOREM 8. The determinant

$$\det A = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

can be written as a sum of $n!$ terms, each of which is a product of n entries in A:

$$\det A = \sum_{\pi \in S_n} \varepsilon_\pi\, a_{\pi(1),1}\, a_{\pi(2),2} \cdots a_{\pi(n),n}\ . \tag{2}$$

Proof. Let $|A^{(1)}, \cdots, A^{(j-1)}, E^{(i)}, A^{(j+1)}, \cdots, A^{(n)}|$ denote the determinant obtained from $|A|$ by replacing the column $A^{(j)}$ with the i-th column $E^{(i)}$ of the identity matrix. Using the formula for expanding a determinant along the j-th column, we see that the cofactor A_{ij} of the element a_{ij} in $|A|$ can be written as the following n-th order determinant:

$$A_{ij} = |A^{(1)}, \dots, A^{(j-1)}, E^{(i)}, A^{(j+1)}, \dots, A^{(n)}| \ ,$$

and, so, expanding $|A|$ along the j-th column, we have:

$$\det A = \sum_i a_{ij} A_{ij} = \sum_i a_{ij} |A^{(1)}, \dots, A^{(j-1)}, E^{(i)}, A^{(j+1)}, \dots, A^{(n)}| \ .$$

If we first do this for $j = 1$, and then evaluate each of the n determinants in the sum by repeating the method for $j = 2$, and so on, we end up with an expression for $\det A$ containing first n, then $n^2, \dots$, finally n^n determinants:

$$\det A = \sum_{i_1} a_{i_1,1} |E^{(i_1)}, A^{(2)}, \dots, A^{(n)}| =$$

$$= \sum_{i_1} a_{i_1,1} \sum_{i_2} a_{i_2,2} |E^{(i_1)}, E^{(i_2)}, A^{(3)}, \dots, A^{(n)}| =$$

$$= \sum_{i_1, i_2} a_{i_1,1} a_{i_2,2} |E^{(i_1)}, E^{(i_2)}, \dots, A^{(n)}| =$$

$$= \dots = \sum_{i_1, i_2, \dots, i_n} a_{i_1,1} a_{i_2,2}, \dots, a_{i_n,n} |E^{(i_1)}, E^{(i_2)}, \dots, E^{(i_n)}| \ .$$

Here $i_1, i_2, \dots, i_n$ run through all sets of n numbers in $\{1, 2, \dots, n\}$ (with repetitions allowed). In other words, taking all n^n maps $\pi : \{1, 2, \dots, n\} \to$ $\to \{1, 2, \dots, n\}$ (see Example 1 in subsection 2 of §1), where $\pi(1) = i_1, \dots, \pi(n) = i_n$, we write $\det A$ in the form

$$\det A = \sum_\pi a_{\pi(1),1} a_{\pi(2),2} \cdots a_{\pi(n),n} |E^{(\pi(1))}, E^{(\pi(2))}, \dots, E^{(\pi(n))}| \ .$$

It remains for us to note that, if $\pi(i) = \pi(j)$ for any $i \neq j$, then the determinant $|E^{(\pi(1))}, \dots, E^{(\pi(n))}|$ has two identical columns, and hence vanishes. Consequently, the terms in the above summation are only non-zero when π is bijective, i. e., when it is a permutation. But in that case, by Theorem 7, we have $|E^{(\pi(1))}, \dots, E^{(\pi(n))}| =$ $= |f(\pi)| = \varepsilon_\pi$. □

Remark. Of course, Theorem 8 can be proved directly by induction on n, but

the appearance of the sign ε_π becomes easier to understand when one adopts the point of view of group theory. Theorem 7 is also of independent interest, in addition to its use in proving Theorem 8.

Theorem 8 can be taken as the starting point for the theory of determinants (as is often done). Namely, after defining $\det A$ by (2), we would then prove all of the properties, including the formula for expanding $\det A$ along the elements of the first (or j-th) column, which was our point of departure in Chapter 3.

EXERCISES

1. Prove that, up to isomorphism, there are only a finite number $\rho(n)$ of groups of given order n.

2. Using Exercise 7 of §2, show that every finite group can be imbedded in a finite group with two generators (i.e., there exists a monomorphism into such a group).

3. Prove that a subgroup of index 2 must be normal.

4. Using Exercise 3, try to prove that, up to isomorphism, S_3 is the only non-abelian group of order 6.

5. Try to show that all of the subgroups of the alternating group A_4 are depicted in Fig. 14. Here we let V_4 denote the so-called Klein four-group:

$V_4 = \{e, (12)(34), (13)(24), (14)(23)\}$;

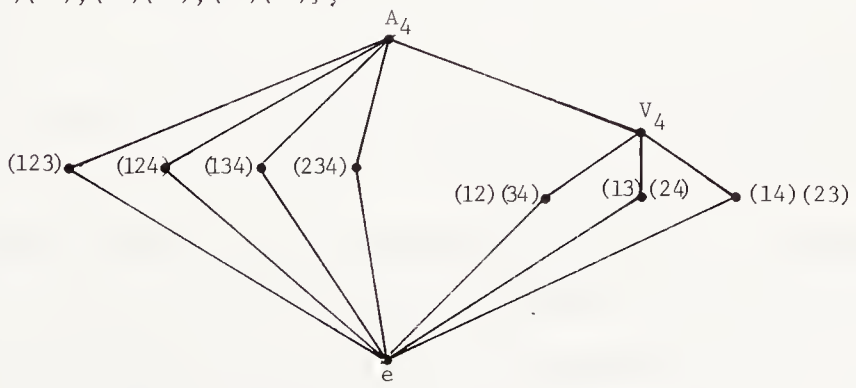


Fig. 14

the other vertices correspond to cyclic groups whose generators are given in the diagram.

6. Show that any group of order 4 is abelian, and is isomorphic to one of the permutation groups: $U = \langle(1234)\rangle$ or the Klein four-group V_4. Alternately, it is isomorphic to one of the matrix groups

$$L_1 = \left\{ \begin{Vmatrix} 1 & 0 \\ 0 & 1 \end{Vmatrix}, \begin{Vmatrix} 0 & 1 \\ -1 & 0 \end{Vmatrix}, \begin{Vmatrix} -1 & 0 \\ 0 & -1 \end{Vmatrix}, \begin{Vmatrix} 0 & -1 \\ 1 & 0 \end{Vmatrix} \right\} \subset GL(2, \mathbb{R}) ,$$

$$L_2 = \left\{ \begin{Vmatrix} 1 & 0 \\ 0 & 1 \end{Vmatrix}, \begin{Vmatrix} 1 & 0 \\ 0 & -1 \end{Vmatrix}, \begin{Vmatrix} -1 & 0 \\ 0 & 1 \end{Vmatrix}, \begin{Vmatrix} -1 & 0 \\ 0 & -1 \end{Vmatrix} \right\} \subset GL(2, \mathbb{R}) .$$

Write out explicitly isomorphisms $U \to L_1$, $V_4 \to L_2$.

§4. Rings and fields

1. The definition and general properties of rings. The algebraic structures $(\mathbb{Z}, +)$ and $(\mathbb{Z}, \cdot)$ were our first examples of monoids, and then $(\mathbb{Z}, +)$ turned out to be an additive abelian group (actually, a cyclic group). But usually one thinks of these two structures together, as a single structure known as a ring. There are important relations of arithmetic which involve combining the additive and multiplicative structures of $\mathbb{Z}$, the most basic of which is the distributive law $(a + b)c = ac + bc$. This law seems trivial to us only because we are so accustomed to using it. If we try, for example, to combine the algebraic structures $(\mathbb{Z}, +)$ and $(\mathbb{Z}, \circ)$ where $n \circ m = n + m + nm$, we find that it is not so easy to find a basic relationship between the two binary operations. Before proceeding to more examples, we give the precise definition of a ring.

Definition. Let R be a non-empty set with two binary operations $+$ (addition) and $\cdot$ (multiplication) which satisfy the following conditions:

(R1) $(R, +)$ is an abelian group;

(R2) $(R, \cdot)$ is a semi-group;

(R3) The operations of addition and multiplication are related by the left and right distributive laws (in other words: multiplication is distributive with respect to addition):

$$(a + b)c = ac + bc \quad , \qquad c(a + b) = ca + cb$$

for all $a, b, c \in R$.

Then $(R, +, \cdot)$ is called a ring.

We call $(R, +)$ the additive group of the ring, and we call $(R, \cdot)$ its multiplicative semi-group. If $(R, \cdot)$ is a monoid, then we say that $(R, +, \cdot)$ is a ring with unit.

It is customary to let 1 denote the unit element of a ring with unit. Sometimes, the existence of a unit element is stipulated in the definition of a ring, but we shall not do this, and shall maintain the distinction between a ring and a ring with unit.

Sometimes in applications and in general ring theory (which now exists in a very developed form), one considers algebraic systems in which the axiom (R2) is either completely dispensed with or else replaced with a weaker axiom, depending on the concrete problem. In such cases one speaks of non-associative rings. But here we shall only deal with the usual (associative) rings. Thus, we shall freely use Theorem 1 of §1, which allows us to disregard parentheses in a product $a_1 a_2 \cdots a_k$ of any number k of elements of a ring.

A subset L of a ring R is called a subring if

$$x, y \in L \Rightarrow x - y \in L \quad \text{and} \quad xy \in L \quad ,$$

i.e., if L is a subgroup of the additive group and a subsemigroup of the multiplicative semigroup of the ring.

It is clear that the intersection of any family of subrings of R is a subring (the argument is the same as in the case of groups), and so it makes sense to speak of the subring $\langle T \rangle \subset R$ generated by the subset $T \subset K$. By definition, $\langle T \rangle$ is the intersection of all subrings of R which contain T. If T itself is a subring, then

$\langle T \rangle = T$.

A ring is said to be commutative if $xy = yx$ for all $x, y \in K$ (unlike for groups, the word "abelian" is not usually used for commutative rings!).

This notion of a ring is very broad. Even the class of commutative rings, which at first glance seems to be rather specialized, has been the object of intensive study for many decades, and the theory of commutative rings is now interwoven with algebraic geometry -- a beautiful mathematical discipline which overlaps with algebra, geometry and topology.

Examples. 1) $(\mathbb{Z}, +, \cdot)$, the ring of integers with the usual operations of addition and multiplication. The set $m\mathbb{Z}$ of integers divisible by m is a subring of $\mathbb{Z}$ (which does not have a unit element if $m > 1$). Similarly, $\mathbb{Q}$ and $\mathbb{R}$ are rings with unit, and the natural inclusions $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}$ give a chain of subrings of the ring $\mathbb{R}$.

2) The properties of addition and multiplication in $M_n(\mathbb{R})$, which were introduced and studied in detail in Chapter 2, show that $M_n(\mathbb{R})$ is a ring with unit element $1 = E$. It is called the full matrix ring over $\mathbb{R}$, or else the ring of $n \times n$ matrices over $\mathbb{R}$. This is one of the most important examples of a ring. Since matrices do not generally commute with one another if $n > 1$, $M_n(\mathbb{R})$ is a non-commutative ring for $n > 1$. The rings $M_n(\mathbb{Q})$ and $M_n(\mathbb{Z})$ of $n \times n$ matrices with entries in $\mathbb{Q}$ and $\mathbb{Z}$, respectively, are contained as subrings in $M_n(\mathbb{R})$. Actually, $M_n(\mathbb{R})$ is full of a wide variety of subrings. Many of them will arise in various contexts later in the book. We further note that it is possible to consider the ring $M_n(R)$ of $n \times n$ matrices with entries in any commutative ring R, since the sum or product of two matrices in $M_n(R)$ will again have entries in R, and the distributive laws for $M_n(R)$ follow from the distributive laws for R. All of this is a direct consequence of the rules for matrix operations, which were summarized at the end of subsections 1 and 3 of §3 Ch. 2.

3) In many areas of mathematics the concept of a ring of functions plays a vital role. Let X be any set, and let R be any ring. Let $R^X = \{X \to R\}$ denote the set of all functions (i.e., all maps) $f : X \to R$, considered along with the two binary operations

pointwise sum $f + g$ and pointwise product fg, which are defined as follows:

$$(f + g)(x) = f(x) \oplus g(x) \quad ,$$

$$(fg)(x) = f(x) \odot g(x)$$

($\oplus$ and $\odot$ are the addition and multiplication operations in R). Here multiplication is obviously not the composition of functions which in the case of linear maps led us to the ring M_n. Rather, pointwise multiplication reflects the point of view in calculus, where $X = \mathbb{R}$, $R = \mathbb{R}$, and, for example, the product of $\tan$ and $\sin$ is $\tan \cdot \sin : x \mapsto \tan x \sin x$, and not $\tan \circ \sin : x \mapsto \tan(\sin x)$.

It is easy to check that R^X satisfies all of the ring axioms. For example, the distributive law in R gives

$$[f(x) \oplus g(x)] \odot h(x) = f(x) \odot h(x) \oplus g(x) \odot h(x)$$

for any three functions $f, g, h \in R^X$ and any $x \in X$; by the definition of the operations in R^X, this gives $(f + g)h = fh + gh$. If 0 and 1 denote the zero element and the unit element in R, then the zero element and the unit element in R^X are the constant functions

$$0_X : x \longmapsto 0 \quad , \quad 1_X : x \longmapsto 1 \quad .$$

If R is commutative, then so is R^X.

The ring R^X contains many subrings, which can be defined by various special properties of functions. For example, let $X = [0, 1]$ be the closed interval in $\mathbb{R}$, and let $R = \mathbb{R}$. Then the ring $\mathbb{R}^{[0,1]}$ of all real-valued functions defined on $[0, 1]$ contains the subring $\mathbb{R}^{[0,1]}_{bd}$ of all bounded functions, the subring $\mathbb{R}^{[0,1]}_{cont}$ of all continuous functions, the ring $\mathbb{R}^{[0,1]}_{diff}$ of all differentiable functions, and so on, since all of these properties are preserved under addition (and subtraction) and multiplication of functions.

To every element $a \in R$ there corresponds the constant function a_X defined by

$a_X(x) = a$ for all $x \in X$, and the imbedding which takes each $a \in R$ to $a_X \in R^X$ allows us to consider R as a subring of R^X.

4) Every additive abelian group $(A, +)$ has the structure of a ring with zero multiplication if we set $xy = 0$ for all $x, y \in A$.

Many of the properties of rings are simply reformulations of properties of groups or, more generally, of sets with an associative operation. For example, $a^m a^n = a^{m+n}$ and $(a^m)^n = a^{mn}$ for all non-negative integers m and n and for all $a \in R$ (compare with relation (2) in §1). Other properties, which are more specific properties of rings resulting from the ring axioms, are suggested by the properties of $\mathbb{Z}$. We note a few of them. First of all,

$$a \cdot 0 = 0 \cdot a = 0 \quad \text{for all} \quad a \in R \quad . \tag{1}$$

Namely, $a + 0 = a \Rightarrow a(a + 0) = aa \Rightarrow a^2 + a \cdot 0 = a^2 \Rightarrow a^2 + a \cdot 0 = a^2 + 0 \Rightarrow a \cdot 0 = 0$ and similarly $0 \cdot a = 0$).

Now suppose for a moment that $0 = 1$. We obtain: $a = a \cdot 1 = a \cdot 0 = 0$ for all $a \in R$, i.e., R only contains the element 0. Thus, $0 \neq 1$ in a non-trivial ring R. Next, we have

$$(-a) \cdot b = a \cdot (-b) = -(a \cdot b) \quad , \tag{2}$$

since, for example, (1) and the distributive laws imply

$$0 = a \cdot 0 = a(b - b) = ab + a(-b) \Longrightarrow a(-b) = -(ab) \quad . \tag{3}$$

Since $-(-a) = a$, it follows from (2) that $(-a)(-b) = ab$ (in particular, $(-1)(-1) = 1$) and $-a = (-1) \cdot a$.

The distributive laws imply the following general distributive law:

$$(a_1 + \cdots + a_n)(b_1 + \cdots + b_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \quad , \tag{4}$$

which one easily derives by induction, first (with $m = 1$) using induction on n, and

then using induction on m. Now using (1), (2) and (3), we obtain

$$n(ab) = (na)b = a(nb)$$

for all $n \in \mathbb{Z}$ and all $a, b \in R$.

Finally, we note the binomial formula of Newton

$$(a + b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} \quad , \tag{5}$$

which holds for all $a, b \in R$, but only if R is a commutative ring. To prove (5), one uses (4) and proceeds just as in §7 Ch. 1, where we considered the special case $R = \mathbb{Z}$.

2. Congruences. The ring of residue classes. According to Theorem 6 of §2, the only non-zero subgroups of the group $(\mathbb{Z}, +)$ are the groups $m\mathbb{Z}$, where m runs through the set $\mathbb{N}$ of natural numbers. But the set $m\mathbb{Z}$ is obviously closed under multiplication as well as addition, and all of the ring axioms are satisfied. Thus, we have the following assertion: every non-zero subring of $\mathbb{Z}$ has the form $m\mathbb{Z}$, where $m \in \mathbb{N}$.

We now try to use the subring $m\mathbb{Z} \subset \mathbb{Z}$ to construct a non-zero ring having only finitely many elements. To do this we introduce the

Definition. Two integers n and n' are said to be congruent mod m (in words: modulo m) if they have the same remainder when divided by m, i.e., if $n - n'$ is divisible by m. In that case we write $n \equiv n' \pmod{m}$ or simply $n \equiv n' (m)$.

In this way $\mathbb{Z}$ is partitioned into equivalence classes of numbers congruent to one another mod m; these classes are called residue classes mod m. Each residue class has the form

$$\{r\}_m = r + m\mathbb{Z} = \{r + mk \mid k \in \mathbb{Z}\} \quad ,$$

so that we may write

$$\mathbb{Z} = \{0\}_m \cup \{1\}_m \cup \ldots \cup \{m - 1\}_m \quad . \tag{6}$$

Note that the residue classes are the cosets of the subgroup $m\mathbb{Z}$ in the additive group $\mathbb{Z}$, and the partition (6) is the one given in Theorem 4 of §2.

By definition, $n \equiv n'\ (m) \Leftrightarrow n - n'$ is divisible by m. But the notation $n \equiv n'\ (m)$ is more convenient than $m \mid (n - n')$, because it is possible to operate with congruences in the same way as with equations. Namely, if $k \equiv k'\ (m)$ and $\ell \equiv \ell'\ (m)$, then $k \pm \ell \equiv k' \pm \ell'\ (m)$ and $k\ell \equiv k'\ell'\ (m)$. In particular, $k \equiv k'\ (m) \Rightarrow ks \equiv k's\ (m)$ for any $s \in \mathbb{Z}$.

Thus, given two residue classes $\{k\}_m$ and $\{\ell\}_m$, we can define their sum, difference or product independently of the choice of coset representative. In other words, we have operations $\oplus$ (addition) and $\odot$ (multiplication) defined on the set $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ of residue classes mod m:

$$\begin{aligned} \{k\}_m \oplus \{\ell\}_m &= \{k + \ell\}_m \ , \\ \{k\}_m \odot \{\ell\}_m &= \{k\ell\}_m \ . \end{aligned} \tag{7}$$

Since these operations are defined using the usual operations on integers, it follows that $(\mathbb{Z}_m, \oplus, \odot)$, like $\mathbb{Z}$, is a commutative ring with unit $\{1\}_m = 1 + m\mathbb{Z}$. This ring is called the ring of residue classes mod m. If one is dealing with a fixed m, one often writes $\bar{k}$ instead of $\{k\}_m$, so that

$$\begin{aligned} \bar{k} \oplus \bar{\ell} &= \overline{k + \ell} \ , \\ \bar{k} \odot \bar{\ell} &= \overline{k\ell} \ . \end{aligned}$$

It is often especially convenient to forget about bars and curly brackets entirely, and simply choose the so-called reduced system of residues mod m, namely, the set $\{0, 1, \ldots, m - 1\}$ and work with this fixed set of representatives. For example, if we use this convention, we have: $-k = m - k$, $2(m - 1) = -2 = m - 2$.

Thus, there is such a thing as finite rings. Here are three examples of the rings $\mathbb{Z}_m$, with the addition and multiplication tables given separately:

$\mathbb{Z}_2$:

+	0	1
0	0	1
1	1	0

·	0	1
0	0	0
1	0	1

$\mathbb{Z}_3$:

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

·	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

$\mathbb{Z}_4$:

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

·	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

The residue rings $\mathbb{Z}_m$ have been of interest to number theorists for a long time, and in algebra they served as a point of departure for various important generalizations.

3. Ring homomorphisms and ideals. By (7), the map $f : n \mapsto \{n\}_m$ has the following properties: $f(k+\ell) = f(k) \oplus f(\ell)$, $f(k\ell) = f(k) \odot f(\ell)$. This suggests that we should call f a homomorphism of the rings $\mathbb{Z}$ and $\mathbb{Z}_m$ and make the following general definition.

Definition. Let $(R, +, \cdot)$ and $(R', \oplus, \odot)$ be two rings. A map $f : R \to R'$ is called a homomorphism if it preserves both operations, i.e., if

$$f(a + b) = f(a) \oplus f(b) \quad ,$$

$$f(ab) = f(a) \odot f(b) \quad .$$

Of course, in this case $f(0) = 0'$ and $f(na) = nf(a)$ for $n \in \mathbb{Z}$.

The kernel of a homomorphism f is the set

$$\operatorname{Ker} f = \{a \in R \mid f(a) = 0'\} \quad .$$

It is clear that $\operatorname{Ker} f$ is a subring of R. But $\operatorname{Ker} f$ is much more than just a subring. Namely, for all $x \in R$ we have $x \cdot \operatorname{Ker} f \subset \operatorname{Ker} f$ (since for all $k \in \operatorname{Ker} f$ we have $f(xk) = f(x) \odot f(k) = f(x) \odot 0' = 0'$) and similarly $(\operatorname{Ker} f) \cdot x \subset \operatorname{Ker} f$. Thus, if we denote $L = \operatorname{Ker} f$, we have $RL \subset L$ and $LR \subset L$. A subring L having these two properties is called a (two-sided) ideal of the ring R. Thus, the kernel of a homomorphism is always an ideal.

As in the case of groups (see the glossary in subsection 3 of §3), a homomorphism $f : R \to R'$ is called a monomorphism if $\operatorname{Ker} f = 0$, an epimorphism if its image (the set of all $a' \in R'$ of the form $f(a)$ for $a \in R$) is all of R', and an isomorphism if it is

both a monomorphism and an epimorphism. If R and R' are isomorphic, we write $R \cong R'$.

The map $f : n \mapsto \{n\}_m$ is clearly an epimorphism $\mathbb{Z} \to \mathbb{Z}_m$ with kernel $\operatorname{Ker} f = m\mathbb{Z}$. When we constructed $\mathbb{Z}_m$ we implicitly used the fact that $m\mathbb{Z}$ is an ideal of the ring $\mathbb{Z}$.

It so happens that every non-zero subring in the ring $\mathbb{Z}$ is an ideal, but that is unusual. For example, in the matrix ring $M_2(\mathbb{Z})$ the set

$$\left\{ \begin{pmatrix} \alpha & \beta \\ 0 & \delta \end{pmatrix} \middle| \alpha, \beta, \delta \in \mathbb{Z} \right\}$$

is a subring but not an ideal.

The example of $m\mathbb{Z}$ suggests a method for constructing ideals (not necessarily all ideals) in an arbitrary commutative ring R: if a is an arbitrary fixed element of R, then the set aR is always an ideal in R. This is because

$$ax + ay = a(x + y) \quad , \quad (ax)y = a(xy) \quad .$$

We say that aR is the principal ideal generated by the element $a \in R$.

Note that if we insist on defining rings to be rings with unit, then ideals are not generally subrings. In that case, we define an ideal to be a subgroup of the additive group of the ring which is taken to itself under left or right multiplication by any element of the ring. If we are using this definition of a ring, then we also insist that $f(1) = 1'$ in the definition of a homomorphism. Of course, even with our broader definition of a homomorphism, this condition $f(1) = 1'$ holds whenever f is an epimorphism.

Isomorphic rings are identical in their algebraic properties, and only properties which are preserved under isomorphisms are of real mathematical interest. It is for this reason, for example, that we allow ourselves to think of the ring $\mathbb{Z}_m$ either as a set of residue classes or as an arbitrary chosen set of representative numbers from the residue classes.

4. The concept of quotient group and quotient ring. Normal subgroups of a group and ideals in a ring have a common origin -- they are kernels of homomorphisms. This common element is reflected in the notion of forming a quotient, which we shall briefly discuss. We shall return to this theme in the second part of the book.

Let us start with groups. The equivalence relation $\sim$ on a group G which is obtained from the partition of G into cosets of a normal subgroup H has a remarkable property. Namely, if a and b are arbitrary elements of G, and if $a \sim c$ and $b \sim d$, then, by definition (see the proof of Theorem 4 in §3), we have $a^{-1}c = h_1 \in H$, $b^{-1}d = h_2 \in H$, and so

$$(ab)^{-1}cd = b^{-1}a^{-1}cd = b^{-1}(a^{-1}c)d = b^{-1}h_1b(b^{-1}d) = h_1'h_2 \in H$$

which means that $ab \sim cd$. Here we used the fact that H is normal in G: $b^{-1}h_1b = h_1' \in H$. Thus,

$$a \sim c, b \sim d \Longrightarrow ab \sim cd \quad .$$

This says that the multiplicative operation on G induces a multiplication of the quotient set $G/\sim$ (see subsection 3 of §6 Ch. 1), which we agreed to denote G/H.

It makes sense to speak of the product of any two subsets A and B of a group G, where we mean the set AB of all products ab with $a \in A$ and $b \in B$. Because G is associative, we have

$$(AB)C = \{(ab)c\} = \{a(bc)\} = A(BC) \quad .$$

A subset $H \subset G$ is a subgroup if and only if $H^2 = H$ and $H^{-1} = \{h^{-1} | h \in H\} \subset H$.

From this point of view, the coset aH is the product of the one-element set $\{a\}$ and the subgroup H. The product of two cosets aH and bH is the set $aH \cdot bH$, which does not necessarily have to be a coset of H. For example, the partition of S_3 into cosets of $H = \{e, (12)\}$, which we considered in subsection 4 of §3, shows that

$$H \cdot (13)H = (13)H \cup (23)H \quad .$$

However, if H is a normal subgroup of G, then the product of two cosets aH, bH will turn out to be a coset of H. Namely, since $gH = Hg$ for all $g \in G$, it follows that

$$aH \cdot bH = a(Hb)H = a(bH)H = abH^2 = abH \quad ,$$

and the same reasoning as used above shows that the coset abH does not depend on the representatives a and b of the cosets aH and bH. The properties

$$aH \cdot bH = abH \quad ,$$

$$H \cdot aH = aH \cdot H = aH \quad ,$$

$$a^{-1}H \cdot aH = aH \cdot a^{-1}H = eH = H$$

show that we have the following

THEOREM 1. *If H is a normal subgroup of G, then the operation $aH \cdot bH = abH$ gives the quotient set G/H the structure of a group, which is called the quotient group of G by H. The coset H is the identity element in G/H, and $a^{-1}H$ is the inverse of aH.* □

If G is a finite group, then the order of the quotient group G/H is given by the formula

$$|G/H| = \frac{|G|}{|H|} = (G:H) \quad ,$$

which should come as no surprise in view of Lagrange's theorem (subsection 4 of §3).

If G is an abelian group whose binary operation is written additively, then the operation induced on the quotient group G/H is written

$$(a + H) + (b + H) = (a + b) + H \quad .$$

In this case, G/H is often called the group G modulo H. If we apply this to the pair $G = \mathbb{Z}$ and $H = m\mathbb{Z}$, we also use the expression "the group $\mathbb{Z}$ modulo m".

We now proceed to the idea of a quotient ring R/L, where R is a ring and L

is an ideal. We base the construction on the additive group of the ring R. Thus, the elements of R/L are the cosets $a + L$, which we call the residue classes of R modulo the ideal L. They are added by the usual rule:

$$\begin{aligned}(a + L) \oplus (b + L) &= (a + b) + L \quad , \\ \ominus (a + L) &= -a + L \quad .\end{aligned} \tag{8}$$

We take the product of these residue classes to be

$$(a + L) \odot (b + L) = ab + L \quad . \tag{9}$$

We have to be careful that this multiplication is correctly defined, i.e., that it does not depend on the choice of coset representatives. Suppose that $a' = a + x$, $b' = b + y$, where $x, y \in L$. Then

$$a'b' = ab + ay + xb + xy = ab + z \quad ,$$

where $z = ay + xb + xy$ is an element of L, because L is a two-sided ideal. Hence, $a'b'$ is in the same coset as ab, and this means that the product (9) is correctly defined. For brevity we write $\bar{a} = a + L$, so that

$$\bar{a} \oplus \bar{b} = \overline{a + b} \quad , \quad \bar{a} \odot \bar{b} = \overline{ab} \quad .$$

In particular, $\bar{0} = L$ and $\bar{1} = 1 + L$ (if R has a 1). We should also check that the set $\bar{R} = R/L = \{\bar{a} \mid a \in R\}$ with the operations $\oplus$ and $\odot$ satisfy all the ring axioms, but this is fairly obvious, because the operations on $\bar{R}$ are defined in terms of the operations on elements of R. For example, the distributive law is verified as follows:

$$(\bar{a} \oplus \bar{b}) \odot \bar{c} = \overline{(a + b)} \odot \bar{c} = \overline{(a + b)c} = \overline{ac} + \overline{bc} = \overline{ac} \oplus \overline{bc} = \bar{a} \odot \bar{c} \oplus \bar{b} \odot \bar{c} \quad .$$

All of this shows that the map

$$\pi : a \longmapsto \bar{a}$$

is an epimorphism of rings $R \to R'$ with kernel $\operatorname{Ker} \pi = L$. Starting with the special case of the quotient ring $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ and the epimorphism $\mathbb{Z} \to \mathbb{Z}_m$, we have found

an analogous situation occurring in arbitrary rings.

It is worth mentioning, although it goes beyond our immediate goal (which is to explain the construction of $\mathbb{Z}_m$ from a general algebraic point of view), that the quotient rings of R by its ideals essentially exhaust all possible images of R under homomorphisms. Namely, if $f : R \to R'$ is a homomorphism and $f(R)$ is the image of R under f, then, if we consider $f(R) \subset R'$ in place of R', we obtain an epimorphism. In order to simplify notation, suppose that f was an epimorphism from the very start, i.e., that $f(R) = R'$. According to the general principle in subsection 3 of §6 Ch. 1, f determines an equivalence relation O_f on R; in our present situation, O_f is the partition of R into cosets $a + \operatorname{Ker} f = C_a$. The map f leads to a bijective correspondence f' between the elements a' of R' and the classes C_a, namely, $f' : C_a \mapsto a'$ if $a' = f(a)$. Under this correspondence f' we have

$$f'(C_a + C_b) = f'(C_{a+b}) = f(a+b) = f(a) + f(b) = f'(C_a) + f'(C_b) ,$$

$$f'(C_a \cdot C_b) = f'(C_{ab}) = f(ab) = f(a) \cdot f(b) = f'(C_a) \cdot f'(C_b) ,$$

so that the bijective map f' is an isomorphism (for simplicity, we have denoted addition and multiplication in the rings R, $R/\operatorname{Ker} f$, and R' in the same way).

We have proved the following

THEOREM 2 (the fundamental theorem on ring homomorphisms). Any ideal L of a ring R determines a ring structure on the quotient set R/L (using the formulas (8) and (9)), and R/L is a homomorphic image of R under a map having kernel L. Conversely, any homomorphic image $R' = f(R)$ of the ring R is isomorphic to the quotient ring $R/\operatorname{Ker} f$. □

Remark. The right side of (9) is not, in general, the same thing as the product of the residue classes $a + L$ and $b + L$ in the set-theoretic sense. For example, if $R = \mathbb{Z}$ and $L = 8\mathbb{Z}$, then $24 \in 16 + 8\mathbb{Z}$ cannot be written in the form $(4 + 8s)(4 + 8t)$, since the latter is always divisible by 16.

5. Types of rings. Fields. In our familiar rings $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$, whenever $ab = 0$ we must have either $a = 0$ or $b = 0$. But the ring of square matrices M_n does not have this property. For example, using the notation E_{ij} (see the proof of Theorem 3 of §3 Ch. 2), we have $E_{ij} E_{k\ell} = 0$ if $j \neq k$, although, of course, $E_{ij} \neq 0$ and $E_{k\ell} \neq 0$. One might think that this is because of the unpleasant phenomenon of non-commutativity in M_n, but that has nothing to do with it. As we saw in subsection 2, in the ring $\mathbb{Z}_4$ we have $2 \odot 2 = 0$ (despite the well-known platitude "twice two is four"!).

Here are two more examples.

Example 1. Pairs of numbers (a,b) (where we may take a, b in $\mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$) with addition and multiplication defined by the formulas

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2) ,$$

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1 a_2, b_1 b_2) ,$$

obviously make up a commutative ring with unit $(1,1)$. Here we encounter the same phenomenon: $(1,0) \cdot (0,1) = (0,0) = 0$.

Example 2. In the ring $\mathbb{R}^{\mathbb{R}}$ of real-valued functions (see Example 3 in subsection 1), the functions $f : x \mapsto |x| + x$ (which is 0 when $x \leq 0$) and $g : x \mapsto |x| - x$ (which is 0 when $x \geq 0$) have the property that their product is the zero function, even though $f \neq 0$ and $g \neq 0$.

Definition. If $ab = 0$ for $a \neq 0$ and $b \neq 0$ in the ring R, then a is called a left zero divisor and b is called a right zero divisor (if R is commutative, then there is no distinction, and we speak simply of zero divisors). The zero element in R can be considered the trivial zero divisor. If there are no zero divisors (except 0), then R is called a ring without zero divisors. If R is a commutative ring with unit $1 \neq 0$ not having zero divisors, then R is called an integral domain.

THEOREM 3. A non-trivial commutative ring R with unit is an integral domain if and only if the law of cancellation holds:

$$ab = ac \ , \ a \neq 0 \Longrightarrow b = c$$

for all $a, b, c \in R$.

Proof. If R has the cancellation law, then whenever $ab = 0 = a \cdot 0$, we then have either $a = 0$ or else $a \neq 0$, in which case $b = 0$. Conversely, if R is an integral domain, then $ab = ac$, $a \neq 0 \Rightarrow a(b-c) = 0 \Rightarrow b - c = 0 \Rightarrow b = c$. □

If R is a ring with unit, it is natural to consider the set of invertible elements: an element a is called invertible (or a unit, not to be confused with the alternative use of this word for the element 1) if there exists an element a^{-1} such that $a a^{-1} = 1 = a^{-1} a$. More precisely, one might want to speak of right invertible or left invertible elements (if $ab = 1$ or $ba = 1$ can be solved for b, respectively). But if R is commutative or is without zero divisors, then right (or left) invertibility implies invertibility. Namely, suppose, for example, that $ab = 1$ in a ring without zero divisors. Then $aba = a$, so that $a(ba - 1) = 0$. Since $a \neq 0$, we must have $ba - 1 = 0$, i.e., $ba = 1$.

As an example, we already know that in the ring $M_n(\mathbb{R})$ or $M_n(\mathbb{Q})$ the invertible elements are precisely the matrices with non-zero determinant.

An invertible element a cannot be a zero divisor: $ab = 0 \Rightarrow a^{-1}(ab) = 0 \Rightarrow$ $\Rightarrow (a^{-1}a)b = 0 \Rightarrow 1 \cdot b = 0 \Rightarrow b = 0$ (similarly $ba = 0 \Rightarrow b = 0$).

THEOREM 4. The set $U(R)$ of all invertible elements in a ring R with unit is a group under multiplication.

Proof. Since the set $U(R)$ contains the identity element, and multiplication is associative in R, it remains for us to see that $U(R)$ is closed under multiplication, i.e., we must check that the product ab of two invertible elements is invertible. But this is obvious, since $b^{-1}a^{-1}$ can be seen to be the inverse of ab (for example, $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = a \cdot 1 \cdot a^{-1} = aa^{-1} = 1$). □

As an example, it is easy to see that $U(\mathbb{Z}) = \{\pm 1\}$ is a cyclic group of order 2.

If we replace the ring axiom (R 2) by the much stronger axiom (R 2') : the set R^* of non-zero elements in R forms a group under multiplication, then we obtain a very interesting class of rings, called division rings or skew fields. Thus, a division ring must be a ring without zero divisors in which every non-zero element is invertible. A commutative division ring -- in which multiplication has essentially all of the properties of addition (i.e., the non-zero elements form an abelian group) -- is called a field. Thus, to repeat:

Definition. A field K is a commutative ring with unit $1 \neq 0$ in which every non-zero element is invertible. The group $K^* = U(K)$ is called the multiplicative group of the field.

A field is sort of a hybrid of two abelian groups -- the additive and the multiplicative -- which are connected by the distributive law. A product of the form ab^{-1} is usually written in the form of a fraction a/b. Thus, the fraction a/b, which only makes sense when $b \neq 0$, is the unique solution of the equation $bx = a$. Operations with fractions are subject to the rules:

$$
\begin{aligned}
&\frac{a}{b} = \frac{c}{d} \iff ad = bc, && b, d \neq 0, \\
&\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, && b, d \neq 0, \\
&-\frac{a}{b} = \frac{-a}{b} = \frac{a}{-b}, && b \neq 0, \\
&\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}, && b, d \neq 0, \\
&\left(\frac{a}{b}\right)^{-1} = \frac{b}{a}, && a, b \neq 0 .
\end{aligned}
\tag{10}
$$

These are the usual rules of "grammar school arithmetic", but, rather than taking them on faith and memorizing them, the reader should derive them from the axioms of a field; but this is not hard. For example, here is a derivation of the second rule in (10). Let $x = a/b$ and $y = c/d$ be the solutions of the equations $bx = a$ and $dy = c$. It follows from

these equations that $dbx = da$, $bdy = bc \Rightarrow bd(x + y) = da + bc \Rightarrow t = x + y = (da + bc)/bd$ is the unique solution of the equation $bdt = da + bc$.

A subfield of a field is a subring which itself is a field. For example, the field $\mathbb{Q}$ of rational numbers is a subfield of the field $\mathbb{R}$ of real numbers. If F is a subfield of a field K, then we also say that K is an extension field of F. It follows from the definition of a subfield that the zero and identity elements in K will also be contained in F, and will be the zero and identity elements of F. If in the field K we take the intersection F_1 of all subfields which contain F and also contain a given element $a \in K$ not in F, then F_1 will be the minimal field containing the set $\{F, a\}$ (the reasoning is the same as for groups, see subsection 2 of §2). In this case we say that F_1 is the extension of F obtained by adjoining a to F, and we write $F_1 = F(a)$. Similarly, we speak of the subfield $F_1 = F(a_1, \cdots, a_n)$ of K obtained by adjoining to F the n elements $a_1, \cdots, a_n \in K$.

For example, if we adjoin $\sqrt{2}$ to the subfield $\mathbb{Q} \subset \mathbb{R}$, the resulting field $\mathbb{Q}(\sqrt{2})$ is easily seen to be the set of numbers of the form $a + b\sqrt{2}$, $a, b \in \mathbb{Q}$, since $(\sqrt{2})^2 = 2$ and $1/(a + b\sqrt{2}) = (a/(a^2 - 2b^2)) - (b/(a^2 - 2b^2))\sqrt{2}$ if $a + b\sqrt{2} \neq 0$. We have a similar situation for $\mathbb{Q}(\sqrt{3})$, $\mathbb{Q}(\sqrt{5})$, and so on.

Two fields are said to be isomorphic if they are isomorphic considered as rings. By definition, if f is an isomorphism, then $f(0) = 0$ and $f(1) = 1$ (where we are using the same notation for the zero and identity elements in both rings). There is no such thing as a non-trivial homomorphism of fields with non-zero kernel, since $\operatorname{Ker} f \neq 0 \Rightarrow f(a) = 0$, $a \neq 0 \Rightarrow f(1) = f(aa^{-1}) = f(a)\, f(a^{-1}) = 0 \cdot f(a^{-1}) = 0 \Rightarrow f(b) = f(1 \cdot b) = f(1)\, f(b) = 0 \cdot f(b) = 0$, $\forall b \Rightarrow \operatorname{Ker} f = K$.

The automorphisms of a field K, i.e., the isomorphisms from K to itself, are connected with the deepest properties of fields, and are a powerful instrument for studying these properties. This subject is called Galois theory.

The notion of a field extension is a reflection of the tendency of mankind, from time

immemorial, to try to increase the supply of number systems to work with. This lengthy historical process can be roughly summarized by the diagram:

$$\{\text{one}\} \rightsquigarrow \{\text{one plus one is two}\} \rightsquigarrow \mathbb{N} \rightsquigarrow \{\mathbb{N}, 0\} \rightsquigarrow \mathbb{Z} \rightsquigarrow \mathbb{Q} \rightsquigarrow \mathbb{Q}(\sqrt{2}) \rightsquigarrow \mathbb{R}.$$

It has continued right up to modern times, and has led to the study of a very extensive network of fields, which range far from the usual number systems of everyday use. Not all of the steps in constructing field extensions are purely algebraic. For example, we go from the rational numbers to the real numbers using the concepts of continuity and completeness (limits of Cauchy sequences), and this is normally done in an axiomatic calculus course. It turns out that there is a completely analogous construction of the so-called p-adic number fields, which we shall not deal with here. The resulting theory of p-adic analysis is the worthy offspring of three areas of mathematics -- number theory, algebra, and analysis.

6. The characteristic of a field. In subsection 2 we constructed the finite ring $\mathbb{Z}_m$ of residue classes

$$\overline{0}, \overline{1}, \overline{2}, \ldots, \overline{m-1}$$

with addition and multiplication operations $\overline{k} + \overline{\ell} = \overline{k+\ell}$, $\overline{k} \cdot \overline{\ell} = \overline{k\ell}$ (we are no longer using the special symbols $\oplus$ and $\odot$). If $m = st$, $s > 1$, $t > 1$, then $\overline{s} \cdot \overline{t} = \overline{m} = \overline{0}$, i.e., $\overline{s}$ and $\overline{t}$ are zero divisors in $\mathbb{Z}_m$.

Now suppose that $m = p$ is a prime number. We claim that $\mathbb{Z}_p$ is a field (having p elements). In the cases $p = 2, 3$, this is immediately clear from the multiplication tables in subsection 2. In the general case, it is sufficient to show that, for every $\overline{s} \in \mathbb{Z}_p^*$ there exists an inverse element $\overline{s'}$ (where we only allow numbers s and s' which are not divisible by p). To see this, we consider the elements

$$\overline{s}, \overline{2s}, \ldots, \overline{(p-1)s} \quad . \tag{11}$$

They are all non-zero, since $s \not\equiv 0 \pmod{p} \Rightarrow ks \not\equiv 0 \pmod{p}$ for $k = 1, 2, \cdots, p-1$; this is because p is a prime number. For the same reason, all of the elements in (11) are distinct: $\overline{ks} = \overline{\ell s}$ for $k < \ell$ would imply that $\overline{(\ell - k)s} = \overline{0}$, which is false. Thus,

the sequence of elements (11) is the same, except for the order in which they are written, as the sequence

$$\overline{1}, \overline{2}, \cdots, \overline{p-1} \quad .$$

In particular, there exists s', $1 \leq s' \leq p-1$, for which $\overline{s's} = 1$. But this means that $\overline{s'} \cdot \overline{s} = \overline{1}$, i.e., $\overline{s'}$ is an inverse for $\overline{s}$. We have proved

THEOREM 5. The ring of residue classes $\mathbb{Z}_m$ is a field if and only if m is a prime number. □

COROLLARY (Fermat's Little Theorem). For any integer a not divisible by the prime p,

$$a^{p-1} \equiv 1 \pmod{p} \quad .$$

Proof. The multiplicative group $\mathbb{Z}_p^*$ has order $p-1$. By Lagrange's theorem (§3), $p-1$ is divisible by the order of any element in $\mathbb{Z}_p^*$, in particular, by the order of $\overline{a}$. Thus, $\overline{1} = (\overline{a})^{p-1} = \overline{a^{p-1}}$, i.e., $\overline{a^{p-1} - 1} = \overline{0}$. □

An alternative proof of this corollary can be obtained by replacing s by a in (11) and multiplying together all of the elements in the sequence; that product must be congruent to the product of $1, 2, \ldots, p-1$.

Even though the fields $\mathbb{Z}_2, \mathbb{Z}_3, \mathbb{Z}_5, \ldots$ seem so different from the familiar field $\mathbb{Q}$, they have one basic property in common with $\mathbb{Q}$: they have no smaller subfields.

Let K be a field. As we already noticed, the intersection of any family of subfields of K is itself a subfield of K.

Definition. A field which does not contain any proper subfield is called a prime field.

THEOREM 6. Every field K contains precisely one prime field K_0. This prime field is isomorphic either to $\mathbb{Q}$ or to $\mathbb{Z}_p$ for some prime p.

Proof. If K contained two fields K' and K'', their intersection would be a

field (non-empty, since both K' and K'' contain 0 and 1) which is distinct from K' and K''. But this is impossible if K' and K'' are prime fields. Thus, the prime field K_0, which is the intersection of all fields contained in K, is unique.

Since K_0 contains 1, it must contain all multiplies $n \cdot 1 = 1 + \cdots + 1$. It follows from the general properties of addition and multiplication in rings (see the end of subsection 1) that

$$s \cdot 1 + t \cdot 1 = (s + t) \cdot 1, \quad (s \cdot 1)(t \cdot 1) = (st) \cdot 1; \quad s, t \in \mathbb{Z} . \tag{12}$$

Hence, the map f of the ring $\mathbb{Z}$ to K which is defined by $f(n) = n \cdot 1$ is a homomorphism. Its kernel is an ideal in $\mathbb{Z}$: $\operatorname{Ker} f = m\mathbb{Z}$. If $m = 0$, then f is an isomorphism, in which case the fractions $(s \cdot 1)/(t \cdot 1)$, $s, t \in \mathbb{Z}$, which make sense because K is a field, form a field K_0 which is isomorphic to $\mathbb{Q}$. This field is clearly the prime field in K.

If, on the other hand, $m > 0$, then the map f^* defined by setting

$$f^* : \bar{k} = \{k\}_m \longmapsto f(k) ,$$

is obviously an imbedding $\mathbb{Z}_m \to K$. By Theorem 5, this is only possible if $m = p$ is a prime number. Thus, $f^*(\mathbb{Z}_p)$ is the prime subfield of K. □

Definition. We say that a field K has characteristic zero if its prime field K_0 is isomorphic to $\mathbb{Q}$. We say that K has characteristic p if $K_0 \cong \mathbb{Z}_p$. We write $\operatorname{char} K = 0$ or $\operatorname{char} K = p > 0$, respectively.

Instead of $\mathbb{Z}_p$ the notation $\mathbb{F}_p$ or $GF(p)$ (GF for Galois field) is often used to denote the "abstract" field of p elements. It turns out that there exists a finite field $GF(q)$ with $q = p^n$ elements for any prime p and positive integer n. We shall return to this interesting question in Chapter 9, but for now we merely give the example of the field of four elements $\{0, 1, \alpha, \beta\}$:

GF(4):

+	0	1	α	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

·	0	1	α	β
0	0	0	0	0
1	0	1	α	β
α	0	α	β	1
β	0	β	1	α

(At this point we are not concerned with what α and β are.) The reader should check, for example, that the distributive law holds.

A field K of characteristic zero has the property that 1 has infinite order in the additive group of K. If K has characteristic p, then any non-zero element has order exactly p in the additive group:

$$px = x + \ldots + x = p(1 \cdot x) = 1 \cdot x + \ldots + 1 \cdot x = (1 + \ldots + 1)x = (p \cdot 1)x = 0 \quad .$$

7. A remark on linear systems. The time has come to cast a thoughtful glance at the theory of systems of linear equations and determinants developed in the earlier chapters. In those chapters the coefficients in the equations and the corresponding matrix entries were numbers (rational or real), but the exact nature of rational and real numbers was never used. There is nothing to stop us now from allowing the coefficients and matrix entries to be elements of any given field K. Of course, then the results must be formulated in terms of the field K: the components of a solution to the linear system and the values of the det function will lie in K. Gauss' method for solving systems of linear equations, the theory of determinants, Cramer's rule, and so on, all remain valid for an arbitrary field K.

Example 1. Suppose we are given a homogeneous system of linear equations $AX = 0$ with square matrix

$$A = (a_{ij}) = \begin{Vmatrix} 1 & 2 & 3 & 4 \\ -10 & 13 & 14 & 15 \\ 12 & -9 & 14 & 15 \\ 12 & 13 & -8 & 15 \end{Vmatrix}$$

and column of unknowns $X = [x_1, x_2, x_3, x_4]$. A direct computation shows that $\det A = 2^3 \cdot 11^3$. Hence, if we consider a_{ij}, x_k to be in K, where K is any field of

characteristic zero or of characteristic $p \neq 2, 11$ (i.e., the integers $1, 2, 3, 4, -10, \cdots, 15$ are replaced by their residue classes in characteristic p), then the system is determined, and so only has the trivial solution $X = 0$.

If $\text{char } K = 2$ (for example, if $K = \mathbb{Z}_2$), then we conclude from the congruence

$$\begin{Vmatrix} 1 & 2 & 3 & 4 \\ -10 & 13 & 14 & 15 \\ 12 & -9 & 14 & 15 \\ 12 & 13 & -8 & 15 \end{Vmatrix} \equiv \begin{Vmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{Vmatrix} \pmod 2$$

that the rank of the system is equal to two, and so the system has two independent solutions $X_1 = [1,0,1,0]$ and $X_2 = [0,1,0,1]$. To avoid confusion we should write $X_1 = [\bar{1},\bar{0},\bar{1},\bar{0}]$, $X_2 = [\bar{0},\bar{1},\bar{0},\bar{1}]$, but we now have had enough experience to adopt the simpler notation, ignoring the bars denoting residue class.

If $\text{char } K = 11$, then it follows from the congruence

$$\begin{Vmatrix} 1 & 2 & 3 & 4 \\ -10 & 13 & 14 & 15 \\ 12 & -9 & 14 & 15 \\ 12 & 13 & -8 & 15 \end{Vmatrix} \equiv \begin{Vmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{Vmatrix} \pmod{11}$$

that the system has three independent solutions

$$X_1 = [9,1,0,0], \quad X_2 = [8,0,1,0], \quad X_3 = [7,0,0,1] \quad .$$

As we see, the answer to our questions about the system depends on our field K, although the procedure is the same in all cases. Thus, one of the advantages of generalizing beyond $\mathbb{R}$ and $\mathbb{Q}$ to an arbitrary field is the avoidance of duplication of the same arguments. But there are even more important advantages.

When we spoke about the general linear group, we have so far meant the group of all non-singular matrices with coefficients in $\mathbb{Q}$ or $\mathbb{R}$. We now let $M_n(K)$ denote the ring of $n \times n$ matrices with entries in an arbitrary field K, and we define the general linear group $GL(n,K)$ to be the subset of all non-singular matrices $A \in M_n(K)$ (i.e., matrices with $\det A \neq 0$). If we vary K, for example, if we take $K = \mathbb{F}_p$, many

important groups arise in a natural way (see Ch. 7).

Fields of the type $\mathbb{R}$, $\mathbb{Q}$, $\mathbb{Q}(\sqrt{2})$, etc. are usually called number fields. The field $\mathbb{F}_p$ is an example of a field which is not a number field. We do not think of $\mathbb{F}_p$ as a "number field", even though its elements can be identified with the elements of the set $\{0, 1, \ldots, p-1\}$.

In §2 Ch. 1 we posed the problem (Problem 3) of using finite fields in coding theory. Here is a small example related to this theme.

Example 2. In order to transmit the word PEACE, in principle it is sufficient to use four elementary message units $P = (0,0)$, $E = (1,0)$, $A = (0,1)$, $C = (1,1)$, which we interpret as row-vectors in the vector space $\mathbb{F}_2^2$ over the field $\mathbb{F}_2 \cong \mathbb{Z}_2 = \{0,1\}$ containing two elements. But suppose some static arises during the transmission, some interference which occasionally switches a 0 and 1. This could cause the receiver to pick up, for example, the message APACE. According to a fundamental theorem of Shannon, the static can always be overcome at the cost of increasing the length of the elementary message units (i.e., lengthening the time of transmission). Suppose, for example, that we know that there can be no more than one distortion in every elementary message unit of length five. Then take a subset S_0 of so-called code vectors in the vector space $S = \mathbb{F}_2^5$, say: $S_0 = \{P = (0,0,1,1,0), E = (1,0,0,1,1), A = (0,1,1,0,1), C = (1,1,0,0,0)\}$. From the table

Code Vectors	P	E	A	C
	00110	10011	01101	11000
Possible vectors obtained from the code vectors as a result of the distortions	00010	00011	00101	01000
	00100	10001	01001	10000
	00111	10010	01100	11100
	01110	10111	01111	11001
	10110	11011	11101	11010

it is clear that the sets of distorted vectors in the different columns do not overlap, and, hence, accurate decoding is possible, i.e., the true message can be re-established.

We have obtained a code S_0 which can correct one error. If we use $\mathbb{F}_2^n$ for n sufficiently large, we can construct a similar code which can be used to transmit the Latin alphabet, and hence any text, accurately. If we want to avoid unnecessarily long and cumbersome decoding, we should choose S_0 carefully. There are many techniques for doing this, including some purely algebraic approaches based on using the finite field $\mathbb{F}_q$.

EXERCISES

1. Developing the idea in Example 2) of §1, show that the set $\mathcal{P}(\Omega)$ with the operations

$$A + B = (A \cup B) \setminus (A \cap B), \quad AB = A \cap B; \quad A, B \in \Omega,$$

is a ring with unit in which all of the elements of the additive group have order two.

2. Prove that any ring in which $x^2 = x$ for every element x must be commutative. Is this true if $x^3 = x$ for every x?

3. Are the fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{5})$ isomorphic?

4. Do the non-invertible elements of the following rings form an ideal: 1) $\mathbb{Z}_{16}$; 2) $\mathbb{Z}_{24}$?

5. Show that the image of a commutative ring under an epimorphism is a commutative ring.

6. Show that, if R is a ring with unit and L is an ideal, then the quotient ring R/L also has a unit element.

7. Show that any finite integral domain is a field.

8. Let p be a prime number, and let R be a commutative ring with unit such that $px = 0$ for all $x \in R$. Show that then

$$(x + y)^{p^m} = x^{p^m} + y^{p^m}, \quad m = 1, 2, \ldots .$$

9. Prove that a ring consisting of five elements is either a ring with zero multiplication or is isomorphic to $\mathbb{Z}_5$.

10. The set $T = \{ \left\| \begin{smallmatrix} a & c \\ 0 & b \end{smallmatrix} \right\| \mid a, b \in \mathbb{Z} \}$ of upper triangular matrices is a subring in $M_2(\mathbb{Z})$. Prove this, and give a description of all of the ideals of the ring T.

11. A non-zero element x in a ring R is called <u>nilpotent</u> if $x^n = 0$ for some $n \in \mathbb{N}$. Show that:

(i) if R is a ring with unit and x is nilpotent, then the element $1 - x$ is invertible;

(ii) the ring $\mathbb{Z}_m = \mathbb{Z}/m\mathbb{Z}$ contains nilpotent elements if and only if m is divisible by the square of a natural number greater than 1.

12. Prove that, in a commutative ring with unit R of infinite cardinality, there cannot be a finite number $n \geq 1$ of non-invertible non-zero elements.

13. Let R be any associative ring with unit 1, and let $a, b \in R$. Show that

$$(1 - ab)c = 1 = c(1 - ab) \Longrightarrow (1 - ba)d = 1 = d(1 - ba) ,$$

where $d = 1 + bca$, i.e., if $1 - ab$ is invertible in R, then so is $1 - ba$. What does the element $1 + adb$ equal?

14. Show that the matrices $\left\| \begin{smallmatrix} a & b \\ -b & a \end{smallmatrix} \right\|$ with $a, b \in \mathbb{Z}_3$ form a field of 9 elements, and that the multiplicative group of this field is cyclic of order 8.

15. Can the code S_0 in Example 2 at the end of the section correct two errors?