

## Chapter 1

# Modeling Preliminaries

### 1.1 Introduction

In this chapter we shall introduce the modeling process. This is a semi formal process by which modeling takes place. In order to do any real modeling however it is still useful to understand a little mathematics and statistics. Precisely how much is still controversial. With software these days it is tempting to think that very little of either is required, however this is like a car driver who knows absolutely nothing of what goes on under the bonnet. Fine as long as everything works, but what if it breaks down? True, we call the experts to fix it, but this text is about model building so it is necessary to be a bit of a mechanic. Basic mathematics is therefore good to have although it is possible to do some modeling without it, see Dyke (1996). In this chapter, we cover the basic mathematics that will enable you to understand the derivations of the fundamental equations in chapter 3. The statistics part has a slightly different function. As well as building models statistics is used for making sense of data, in particular inferring from data. It is difficult to see when this will ever be unnecessary so an introduction to statistics will always be an extremely useful inclusion in a book such as this. There is nothing in this chapter that should be beyond anyone who has studied mathematics in the year they turned 18, and most of it should be accessible to those who dropped mathematics after age 16. The prime motivator for learning anything new is its usefulness, and rest assured everything in this chapter is useful at least once in the rest of the text. Also, when the mathematics and statistics is introduced, it will be done in a way to highlight this motivation, that is whenever possible through the use of relevant examples.

## 1.2 Introduction to Modeling

To the majority, the word modeling still means something to do with photography or, if they have a scientific background, the building of scaled-down replicas that ought to mimic real life situations. In this latter category one thinks of Civil Engineering consultants building models of harbours with attendant breakwaters and jetties, and then subjecting them to a particular wave climate. The way this is done is to build a physical model, usually in a large area reminiscent of an aircraft hanger. In this model, the area of coast or river or estuary (whatever) is built from materials such as concrete, sand and cement. Of course there is a scale, perhaps 1:20 or even larger, which needs to be considered when examining results. If waves are of interest, then there has to be a paddle mechanism included in order to generate them. Exactly how the scale factors can be calculated is the subject of Chapters 2 and 3, but suffice it to say that measurements of quantities such as wave height, current speed and direction, the force on pier or jetty can be made on the model. Appropriate scale factors are then applied and an estimate of the real life wave height, current speed and direction, force on the pier or jetty or whatever can then be made. Up to thirty or so years ago virtually all modeling in coastal engineering or oceanography referred to this kind of activity. These days, modeling invariably means use of the computer and the big once national facilities (e.g. Hydraulics Research in the U.K. Delft Hydraulics in The Netherlands both now privatised) now have much scaled down (no pun intended) the facilities for these physical models but have many sophisticated computer models to replace them. Many would prefer the word enhance rather than replace, as there is still a place for a physical model where the carefully placed strain gauge can give information to reinforce the output from a mathematical model. In most cases, the results from a mathematical model implemented via software on a computer will tell the same story as the results from a physical model, but if there are contradictory results, neither should automatically be believed. Perhaps they are both wrong and the situation is more complicated than either model builder thought. Areas where physical modeling is still dominant is in the building of bridges and in the design of spacecraft. In both of these areas, the final costs are so huge that the expense of building a physical model is less critical than for example estimating dilution rates of a dissolved substance in an environmentally sensitive estuary. In this latter case, mathematical models are now almost always used.

We shall not be discussing physical models in this text. The parallels will be explored a little further in Chapter 2 where this is natural, but thereafter physical models are left behind. Mathematics is often thought of as operating within a very well defined set of axioms using well defined techniques to give precise answers to well defined problems. Pure mathematicians hold this view. However, such a rigid structure is not well suited to contribute to the description of a practical science such as oceanography or coastal engineering. Papers that are very mathematical which may be very interesting in their own right often have only a tenuous link with reality. At the other end of the spectrum there are some very simple mathematical models that embody the essence of oceanographic truth, and we will definitely be meeting some of these. It is this blending of mathematics with the knowledge of oceanographic processes in which the art of successful mathematical modeling lies.

There is no doubt whatever that mathematical modeling has been greatly assisted by the rapid advances in computer power. Nevertheless, mathematical modeling can still certainly take place without it, it is just that computing increases the scope of the modeling. The emphasis in this text is definitely not on computing which may be thought of as a tool that enables modeling to take place. Instead the concentration is on the correct mathematical description of the ocean and coastal physics and to some extent biology. Here we shall treat the terms coastal oceanography, coastal marine physical science and coastal engineering as synonymous. Coastal oceanography is a science that has grown through painstaking observation and progressed through scientists making judgements and deductions from these observations. There are several distinctive features that although not peculiar to coastal engineering, oceanography, meteorology and earth science in general make it particularly amenable to the relatively new art of mathematical modeling. (Yes, although there is a great deal of scientific method and rigour in mathematical modeling, it still remains in many ways very much an art.) First, coastal oceanography as an applied science has to incorporate aspects of physics, chemistry and biology. Indeed it may be argued that the sea provides an ideal vehicle for the study of some (but certainly not all) of the fundamentals of these basic sciences. In order to understand some of the processes that go on in the sea, it is therefore necessary to simplify some aspects and ignore others. This is what occurs in modeling. Secondly there is a very important aspect to modeling called validation. In most sciences and engineering, validation means trying out the model and comparing it with the real situation. In coastal oceanography, the entire history of the science from the accumulated wisdom of

fishermen through the voyages of discovery to modern day scientific expeditions is centred around observations and provides in some respects an ideal scenario for validation. Conditions are however not controlled, as the day has not arrived where weather can be prescribed therefore there is no control over at least one input. This is certainly a disadvantage in some respects, although it does encourage the continuation of the lively debate between the modeler and the observer.

### **1.2.1 *Environmental Issues***

In the last few years many new substances have been developed, from those used in foodstuffs, packaging, building materials to paint additives and the many plastics used everywhere in our daily lives. With the industrial processes that are used to produce these substances has come an awareness that care has to be taken about the disposal of the byproducts of the manufacturing. The environmental lobby has become particularly strong in the last fifteen to twenty years, and it is really only now that we are at last beginning to become aware of the lasting effect of all the foreign material mankind seems to be continually pumping into the earth's environment. Of course, much research needs to be done before our understanding is anywhere near adequate, but now whenever there is a new manufactured chemical or the production of a hitherto unsuspected byproduct there is in most countries strict legislation governing what can and cannot be allowed into the environment. Unquestionably, sometimes the environmental lobby prevents what is an innocent process taking place, but this is much less worrying than permitting a pollution that may unwittingly cause widespread environmental damage.

Some worrying case histories have come to light following the collapse of the Soviet Union. In Poland, Czechoslovakia (now split into the Czech Republic and Slovakia) and other former Warsaw Pact countries cumulative environmental effects have shortened the life expectancy of people living within the range of rivers polluted by the unthinking discharge of industrial waste including that from nuclear industries. Major rivers in Poland were so polluted that they were useless even for industry let alone recreational use. Using them for drinking water was completely out of the question. Before 1991 East Germany (as it was then called) was encouraged to increase industrial production regardless of any impact this might have on the environment. In December 1990, drinking water in Brandenburg did not meet EU (European Union) quality standards. Tankers had to be

used to supply drinking water in some areas. Nitrate levels were up to 25 times EU legal limits. Samples from the Havel river contained high levels of phosphate, ammonium, benzol and zinc. Purification ensued, but even then water was found still to be contaminated with oil, even phenol. Post 1991 the west has become horrifically aware of dead rivers and dead lakes with toxic levels of similar chemicals. The unified Germany is doing its best to clean up the mess left by its communist predecessor, but it takes both time and money. It is a cruel irony that Germany now comprises two pre 1991 countries that were at the opposite ends of the environmental spectrum as far as cleanliness is concerned. A cruelty that has hit the ex West German where it hurts; in his pocket. Some of these more legal and political issues are outlined in section 1.2.4.

There are many ways in which chemicals can effect the environment. Most of these are local effects and come under the name pollution. There are two effects however that have had an impact over the whole globe since the mid 1980s. In order to reach the public attention these days requires the adoption of a “sound bite” (itself a sound bite incidentally; interesting for those aware of Russell’s Paradox.). The sound bites in question are “global warming” and “ozone depletion”. If the public can get things wrong it usually does, and these effects although very independent have been confused. To sum up global warming, this is the increase of (man made?) gaseous discharge of chemicals such as carbon dioxide, methane and sulphurous oxides into the atmosphere which alters the balance between incoming and outgoing radiation, decreasing the latter so that the earth as a whole warms up. That the earth’s albedo (as this balance between incoming and outgoing radiation is called) is changing is beyond question but whether it is through industrial pollution or the eradication of huge swathes of equatorial rain forests (or some of both) is still controversial. More is said about climate modeling in Chapter 8. In the last few years though, almost everyone (but not the USA) seems convinced that global warming is largely due to man made greenhouse gases. As this is being written, the southern USA is suffering hugely from the after effects of hurricane Katrina. Although hurricanes happen every year and there have been similar sized ones in the past, thankfully missing crucial vulnerable areas, questions will no doubt be asked. Do hurricanes contain more energy on average than they used to? If the answer is yes, could this be yet another manifestation of global warming. (In fact, hurricane activity goes in cycles of around thirty to forty years, and we are ending a quiet period just now – the 1960s was the last active period. This more energetic hurricane activity is coinciding with the

enhanced media awareness of global warming, so the temptation to put two and two together and get five is overwhelming, especially for those who like to give the USA a wake up call.)

There is less controversy about ozone depletion. This is the name given to the disappearance of that layer enveloping the earth largely comprising ozone that is responsible for shielding us from the harmful effects of the sun's ultra-violet rays. The aerosol can is less than 100 years old, and in the last twenty years it has mushroomed in use until the gas used for the propulsion of the spray (CFC or Chlorofluorocarbons) was found to persist long enough to attack and destroy parts of the ozone found in the upper parts of the atmosphere. As has been said briefly above, this ozone protects us from the ultra-violet rays arising out of the sun and overexposure to which can lead to skin cancers. This whole ozone depletion problem has been brought to the attention of the public by the publication of colorful pictures of the increasingly large hole in the ozone layer over the antarctic, and a smaller one discovered over the arctic. The problem in fact has been known for some time; the depletion of stratospheric ozone was first observed as long ago as 1979, but not reported until 1985. A hole was observed in the ozone layer which is largest in the springtime, but at this time no culprit was identified. It was perhaps an oddity that would go away. Not so. By 1988, it was recognised that CFCs had an important role and the Montreal Protocol was signed: 100 countries signed a protocol in 1988 to discuss ways of arresting the depletion of ozone in the upper layers of the atmosphere. By 1990 and 1991 there were two successive years of severe ozone depletion, it was not going away and CFCs were confirmed as being linked to the problem. Although CFCs, first introduced via aerosols in the 1930s were contributing directly to ozone depletion, there was additionally a positive feedback mechanism at work. The CFC derived chlorine lowered air temperatures which in itself made CFC derived chlorine more effective in depleting ozone. By 1993, ozone had depleted to 21% of "normal" values. In 1996, the hole was as big as USA and Canada combined and by 1998 it extended over an area nearly twice the size of the antarctic continent itself. Not until 2015 will there be any sign of recovery. The greenhouse effect interacts with this ozone depletion problem by cooling the upper atmosphere even though the lower atmosphere is warming, which leads to the maintenance of the hole. All this is a salutary lesson in unforeseen environmental consequences to man made chemicals. Besides UV rays (leading to the enhanced risk of skin cancer as well as other health problems) other consequences, as yet unknown, may still await us. Not before time, in December

1995 signatories to the Montreal Protocol met in Vienna and agreed limits on ozone depleting substances (methyl bromide and LDCs, the former to be phased out by 2010, the latter to stabilise at 1995-1998 levels by 2002). Even though ozone depletion has receded from being newsworthy, the hole in the Antarctic ozone layer is likely to be there beyond 2050 no matter what environmental measures are adopted now.

These global problems have helped and continue to help to focus public attention on the environment and how important it is to protect it even though these two large problems are perhaps beyond the individual to influence to any measurable extent. Nevertheless, people are now aware that they must “do their bit” to protect the environment whether this is by being careful about waste disposal (who had heard of a bottle bank or biodiesel thirty five years ago?), or by using their cars less. Nowadays, using unleaded fuel in the family car or supporting so called organically grown foodstuffs is recognised good practice. This text will only touch on such difficult world wide environmental problems, instead the focus will be on smaller scale modeling, and these large problems form the context in which many of the smaller models are embedded. In the next section let us look at the modeling process itself.

### 1.2.2 *The Modeling Process*

In many books on mathematical modeling, the starting place is the description of some kind of idealised modeling process using as a vehicle some equally idealised problem. The trouble with this is that both students and experienced practitioners alike find this less than convincing. The element of trial and error that seems to be involved in the classical modeling process is unrealistic to the practicing oceanographer, whilst to students of ocean science the whole process looks too ideal, not related to actuality. However it is the heuristic trial and error side of modeling that makes it so successful in its mimicry of real life. We therefore must design our modeling process particularly with the coastal ocean scientist in mind.

The singular most important aspect of modeling that has led to its recent popularity is the ready availability of cheap but increasingly powerful computers. In the whole of the 1960s and 1970s and the first half of the 1980s in order to use these computers it was necessary to be able to programme them. In order to programme them it was necessary to learn a high level computer programming language such as ALGOL (in the early days), FORTRAN, PASCAL, C++ or any one of a heap of object-oriented

programming languages. The details of the programming in turn demand a detailed knowledge of mathematics and the numerical methods that are used to translate the mathematics into the discrete mathematics that computers can use. By their very nature, these programs utilising as they do powerful computers are complicated and are based on sophisticated rather than simple mathematics. A requirement for those involved in marine modeling was therefore some knowledge of mathematics including the calculus that is used to describe the dynamic balances in a fluid and the transport of heat and salt, and the techniques of discretising that in turn demand knowledge of numerical methods. Much of this kind of modeling is still of course going on, but unquestionably it is no longer mandatory to be as close to the mathematics.

Many marine scientists are concerned with models on computers because they wish to answer engineering or environmental impact questions, but they lack the mathematical background to formulate and then program models on computers themselves. Software (the modern name for a computer programme that is commercially available) is only a successful product if it is accessible to the majority of likely users. Of course it must also be useful in terms of producing meaningful results. It is widely recognised that not enough marine scientists have the mathematical background to comprehend the details of today's marine models, and it is indeed fortunate that this is no longer necessary. The very computer power that enables the models themselves to be complex also enables so-called "front ends" to be incorporated into models. These front ends act as an interface between the program and the lay user and enables the lay user to use the program constructively without the need to get involved in the programming itself. One common method is to use English commands to enquire of the user what features he or she wishes to incorporate into a model. In effect the user operates with a series of menus, choosing from a set or answering yes or no to simple questions. In this way, a particular problem can be solved by adapting a complex program through menus and without detailed programming knowledge. This is the general philosophy behind expert systems that are now quite widespread especially in the medical field.

The general philosophy behind modeling in marine systems can be expressed succinctly in a flow chart of the type shown in Figure 1.1. In the language of systems, ocean science might be thought of as a mixture of a soft system and hard system. A soft system is one that is ill-posed and usually involves humans, whereas a hard system is one that is controllable, obeys well formed laws and is by and large amenable to exact mathematical

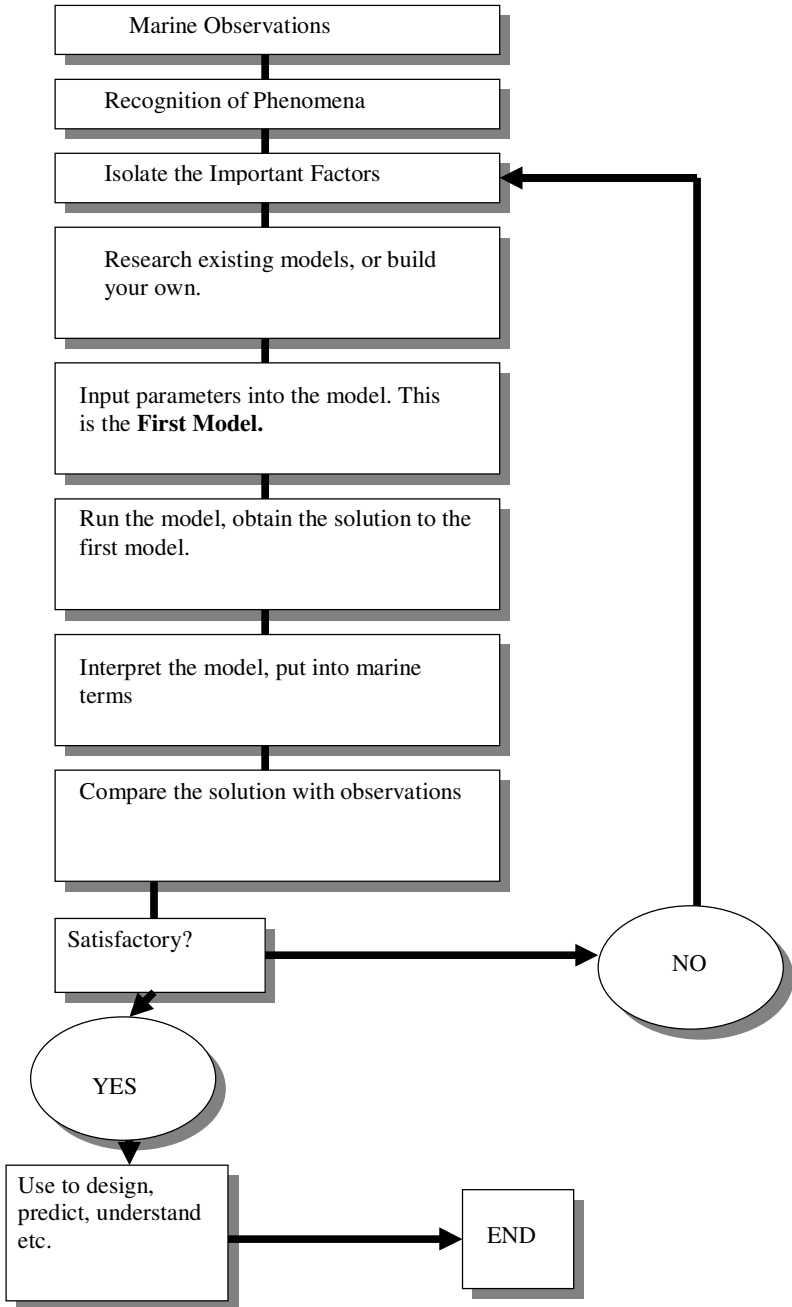


Fig. 1.1 A marine modeling process flow chart

solution. Perhaps a more natural division for marine science is into three classes; natural systems (e.g. biological organisms); artificially designed but physical systems (e.g. engineering devices); and artificially designed but abstract systems (e.g. economic models and models involving scheduling). The kind of modeling indicated by the flow chart of Figure 1.1 mostly fits into the first of these categories, however models of the physics of the sea are very different from biological models. Most models of ocean physics arise from the application of well established laws and lead to well posed mathematical problems. Therefore in systems methodology they are hard systems. Biological models related to the ocean are also in fact hard as although there are no universally recognised laws with the same stature as physical laws, they are well posed and solutions exist. It is only very recently that ocean scientists have brushed with what might be termed soft systems. They are only soft because the sheer complexity of the latest models renders exact predictability difficult. With the development of genetic algorithms the distinction between hard and soft systems is blurring. Related neural network modeling seeks to simulate systems by emulating the way humans tackle problems, through training and learning. These notions are in their infancy and have not yet reached the stage where they can be applied systematically to environmental problems (but see section 2.5.2).

Thinking of modeling itself, it is perhaps tempting to describe it as a well established set of principles and routines. It would be wrong to say that this was far from the truth, but as more variables are considered modeling gets more of an art. If a proposed model has a vast number of free parameters whose values are more or less at the behest of the modeler, then the choice of final model is also very large. There may be many models, all slightly different but all of which fit the observed picture with similar margins of error. The most common instances of models with excessive numbers of parameters are models of ecosystems with many variables and many parameters that describe the exchanges between the variables. In such models, the level of uncertainty associated with outputs must also inevitably increase. This is not a criticism of such models, far from it. It is merely a consequence of having so many free parameters. Note here the difference between *parameter* and *variable*. *Variables* are quantities whose behaviour we wish to model, whereas *parameters* are quantities that can be estimated either directly or indirectly and are there as a direct consequence of the modeling process. Parameters are not natural quantities, variables are. Of course if there is a perceived fault in the outcome of a model, it is usually a matter of conjecture whether any fault lies in what has been left

out of the model, the method the software uses, or the interpretation of the output. Perhaps it is the observations themselves that are wrong and the model is actually working. Usually experience alone tells us the most likely source of the error, and one of the aims of this book is to help you to gain some experience through the eyes of the author.

### 1.2.3 *Engineering Projects and Consultancy*

The prime reason engineering projects are of interest in a book such as this is because they produce waste. The environmental scientist views a factory on the shore as a potential hazard. In contrast, the economist will view the same factory as a potential for growth and employment. Most large factories will use water, perhaps as a coolant or more actively in a chemical process, and the only way to ensure that the factory is economically viable is to discharge the water into the nearby river or estuary. In recent times, the recycling of waste has increased due to advances in recycling technologies. Nevertheless, waste products are still produced which emerge from the factory carried by the waste water. When a new factory is proposed, it is common for various parts of its construction to be put out to tender. This means that various sets of experts will be paid to build sections of it. One section will undoubtedly be the design of the waste disposal, which commonly takes the form of a diffuser to make sure that any waste water remains legal. The more technical aspects of how discharges behave are dealt with in Chapter 5 where diffusers are described. Here we are concerned to ensure that any potentially hazardous dissolved substance present in the waste may remain within environmentally acceptable values as laid down by legislature. In order to do this, the team involved in designing the diffuser often contains someone who is capable of building (or getting hold of) an appropriate mathematical model. This model should be capable of simulating the worst case scenario whereby the levels of the hazard are in some sense maximal. If this maximum level is legally acceptable, then normally the go ahead for the manufacture of the diffuser as designed can be given. Some of the more technical aspects of flumes are covered in Chapter 5.

The environmental questions that arise from not only the building of factories but many other activities such as sport, the modification of harbours, housing developments etc. has meant that environmental consultancy is now big business. As the population increases, and man explores and exploits more of the earth's surface, the environmental scientist assumes the role of a guardian. It is now certainly not optional to consider

most carefully environmental questions every time industrial or other unnatural activity impacts on the world around us. Some of the things that were done in the name of industrial advancement in the UK are now seen as very wrong and make us wince. Unfortunately, like activities are still going on in other less civilised parts of the world. The framework for the control of this is the law which forms part of the next section.

#### **1.2.4 *Public Perception and Legal Issues***

When man entered the industrial age back in the early 19th century, he started releasing chemicals into the environment. However back then the quantities were not large and the sociology and knowledge of environmental chemistry at this time also meant that any discharges were disregarded. More recently, things have certainly changed. The accidental or deliberate release of chemicals into the environment is now a very live issue. When a chemical is so released a whole host of processes ensue. As far as the activity of man is concerned, there are legal processes which will be outlined later. To give a flavour of the scientific processes, first of all the introduced chemical will interact with naturally occurring chemicals in the environment. This interaction could be just mixing, but could also involve chemical reactions with important consequences for the environment. Then there is the physics of how the introduced substance interacts with the environment. In water (river, estuary or sea) this could involve diffusion, sedimentation, or simply the migration of the substance with local currents. In addition the substance could be buoyant and contribute to a surface slick, or sink and interact with the sediment. Finally the introduced chemical could interact with the biology; the plankton, fish, aquatic reptiles, amphibians, mammals and ultimately man. It is this interaction, almost always the most difficult to forecast that of course is the most worrying for the public and lies behind the heightened interest in environmental protection. We have to be so very careful as biological consequences are the most difficult of all to predict. To cite a non marine example, who would have thought that an insecticide (DDT) would cause birds that accidentally (incidentally really as they were more interested in the small mammals and birds that ate the insects that absorbed the DDT) consumed it to lay eggs that had abnormally thin shells. These shells could not withstand the weight of the incubating adult, hence the population of birds (peregrine falcons actually) was decimated. In the medical field we all now know the consequences of a pregnant woman taking the anti morning sickness drug thalidomide. It is

this kind of worry that is behind the reaction of the general public (or do I mean the popular UK press) to genetically modified (GM) crops. A public, it must be remembered which is still smarting and facing uncertainty over BSE (Bovine Spongiform Encephalopathy) and its relationship to new variant CJD (Creutzfeldt Jakob Disease) which although numbers seem to have peaked in 2003 is still potentially a serious worry. Then of course there is the use of the word “nuclear”. A technology that has tremendous potential but the name of which has associate with it the (erroneous) millstone of being linked with bombs and destruction as well as the (not so erroneous) problem of radioactive waste disposal.

In recent times, the control and management of any waste products of industry have become increasingly governed by legislation. Pollution laws as they are colloquially known tend to be different in different regions of the world. In the USA the laws are usually very strict in terms of health, but less so in environmental protection terms. In Singapore all environmental laws are very strict indeed. In the old eastern block countries (notoriously East Germany as it tried to build its industry up from the ravages of the Second World War) environmental legislation was virtually non-existent and ignored even when it was there. Even in the UK pollution laws can be different in England and Wales than in Scotland and Northern Ireland. Legal principles are there to protect the interests of those who can be threatened or damaged by pollution. The legal process is notoriously expensive and often tortuous, therefore it is in everyone’s interest (apart perhaps from the lawyers) not to go to court. Most cases are thus settled by insurers. Some large cases however need a national forum and usually arise because they are in some way new. Obvious examples are the large tanker accidents the Torrey Canyon (1967) and the Amoco Cadiz (1978) and more recently the Exxon Valdez where liability was contested, not surprisingly given the extent of the environmental disaster in each case. On a smaller scale, the liability for minor slicks caused by the flushing out of “empty” tanks lies squarely with the captain and the only problem is catching him (or her). More relevant to modeling is the legislation that exists to prevent more than particular concentrations of certain chemicals in the sea. This is the maritime equivalent to monitoring lead in car emissions which led to the development of lead free petrol and the catalytic converter. Examples of what is meant in the maritime context is the pollution caused by painting boats with anti fungal paint (tributyl-tin), or the control of heavy metals that arise from chemical process factories (lead (again), mercury and cadmium), enhanced levels of which can occur in the waste water which is discharged

into the nearby river or estuary. Experts in the biochemical effects of toxic chemicals usually formulate levels of chemicals that are deemed reasonable to tolerate, and if these are exceeded prosecutions ensue. As new processes are developed, these are scrutinised and the legislation is modified accordingly. The enforcers of the legislation vary. Sometimes it is a national body such as the River Authority (in England and Wales) or the River Purification Boards (in Scotland), sometimes the transgressor is in breach of a law such as the Environmental Protection Act in which case the police can be involved. Successful prosecution can result in the closing of a factory and the fining or imprisonment of offenders. As authorities become convinced that industrial concerns can consistently meet minimum standards, they are able to issue some kind of licensing agreement. This grants the licensee to manufacture.

In the international sphere, there are international treaties that all signatories obey. Examples of this include the treaty that preserves Antarctica for scientific study, and the Treaty of Rome which set up the fundamental legislature for the European Union. In recent times the law has become more complex in Europe as EU legislation increases. In the maritime field there have always been complications due to offences taking place in international waters, or there being at least three nations involved. One which owns the offending material (usually oil), one which owns the container (oil tanker), or the country of origin of the captain and the country to which the territorial waters in which the act occurred belongs. To this complex picture needs to be added the question of market forces. It is still broadly true that the small guy fighting the multi-million pound company loses.

In the UK case law still plays an important part. That is the court is a very powerful body, and if a case sets a precedent perhaps in the level of compensation paid by a company that has caused bodily harm (for example) this is cited in similar future cases. Some kind of convergence then occurs if there enough cases. This is the power of “common law”.

The big question of global warming is causing meetings with high profile politicians (starting with Kyoto 1997 and Rio 1998, the latest was held in Montreal 2005) but as yet no treaty worth the name has emerged due to the partisan interests of some of the delegates, contradictory differences between environmental desires and local political reality are irreconcilable. A few years ago, the representatives of the Florida Keys community joined forces with other island communities around the world (Fiji, Samoa etc.) in the hope of providing a convincing lobby to the USA industries who are widely recognised as main contributors to the industrial gases that cause

global warming. Then there are the two large emerging economies of China and India that are altering the balance of the world economy and have to be part of any global warming legislation. If the sea level rises continue, then island communities will be no more in a century or two. Despite the current preoccupation with terrorism, this is probably our biggest long term problem and it is a very difficult one to solve.

Given the very long time scales, the signs that mankind can reverse the effects of global warming, or the effects of ozone depletion (for which there is more agreement) are not good. There are some very recent encouraging signs, but not enough. One convincing strategy that will help is to provide accurate models, and to do this the underlying processes have to be understood. This is what this textbook is about.

### 1.3 Mathematical Preliminaries

The basis of most models lies in the mathematical description of the laws obeyed by the thing being modeled. For coastal engineers and marine scientists this is most often the sea, though it could be a marine ecosystem (see Chapter 7). The sea is a fluid (salty water) and so the equations obeyed by the sea are those of fluid mechanics. This presents us with a problem as the mathematical description of a fluid is not something one meets outside quite advanced courses. Moreover, the sea is a fluid with dissolved substances, changes in temperature plus (most of the time) turbulence. So, what we attempt to do here is difficult, but not impossible. It amounts to giving some mathematical background in order that the derivation of the equations in Chapter 3 can be understood. The kind of mathematics required comes under the heading of “vector calculus” or sometimes “advanced calculus”. This presupposes that everyone knows about calculus of course, which might not be the case. Calculus was developed in the 17th century simultaneously (being diplomatic) by Newton in England and Leibniz in Germany. It is Leibniz’ description that is followed nowadays. It is a powerful tool for studying things that change (the word “calculus” derives from the Latin for “a small stone” which were used to help in calculations in ancient times). The question is how much do we need to know about calculus in general. One useful definition is that of the derivative. The usual way to define a derivative is graphically in terms of tangents and this is alluded to when we outline numerical methods in Chapter 4. However, we need calculus to describe rates of change, so it is this definition that wins here. If a quantity  $u(x, y, z, t)$  depends on where it is (the co-ordinates  $(x, y, z)$ )

and when it is (time  $t$ ), then it will change with time and space. The letter  $u$  usually denotes the current in an easterly direction, and it needs no leap of imagination to realise that this current will change in time even if we consider a fixed point in space. Think of the current due to the tide at a particular location as an example. The easterly component of this current will also be different depending on where it is measured. There are thus different rates of change for  $u$ ; four in fact one each for  $x, y, z$ , and  $t$ . We cater for this in words by writing “the rate of change of  $u$  with respect to  $t$ ” or  $x$  or whatever. The rate of change of  $u$  at a particular fixed location with respect to  $t$  is the derivative of  $u$  with respect to  $t$  and is written

$$\frac{\partial u}{\partial t}.$$

It has the definition

$$\frac{\partial u}{\partial t} = \lim_{\Delta t \rightarrow 0} \left\{ \frac{u(x, y, z, t + \Delta t) - u(x, y, z, t)}{\Delta t} \right\}.$$

Note that only  $t$  is varying, this is the reason for the curly style of the “d” and it means that everything except  $t$  is being kept constant. In books on calculus which are normally dauntingly large, there are usually a lot of practise examples and exercises on finding the rates of change, first of functions like  $u(t) = t^2$  where the “d” really is just d and not curly as there is only the one variable. This is the derivative section of the calculus of one variable and usually comes first. Later (usually much later) there will be examples such as  $u(x, y, z, t) = x^2 + y^2 + z^2 - t^2$ . Although it is useful to know how to find rates of change of various functions like polynomials, exponential and trigonometric functions, this is not the point here. This section cannot hope to duplicate large texts on the calculus. If the value  $t^2$  is actually substituted for  $u$  in the right hand side, the limit takes the form

$$\lim_{\Delta t \rightarrow 0} \left\{ \frac{(t + \Delta t)^2 - t^2}{\Delta t} \right\} = \lim_{\Delta t \rightarrow 0} \left\{ \frac{(\Delta t)^2 + 2t\Delta t}{\Delta t} \right\} = 2t.$$

Taking this kind of limit gives all the basic rules of the calculus found in the textbooks and now available on your PC or local area network provided you have paid for the software (MAPLE or MATHEMATICA are very powerful, but there are small calculators that can do symbolic calculus and algebra. These are usually banned from examinations, but this is a book not an examination). So we have introduced what is called differentiation, that is taking this limit and finding rates of change, and it is the meaning of the partial derivative (a rate of change of a quantity with respect to a single

variable, the others remaining constant) that is crucial to the understanding of the balances derived in Chapter 3.

The inverse of addition is subtraction, the inverse of multiplication is division so the inverse of differentiation is integration. The mechanics of subtraction are more difficult than the mechanics of addition, the mechanics of division are more difficult than the mechanics of multiplication, so the mechanics of integration are more difficult than the mechanics of differentiation. Indeed, sometimes it is impossible to find certain integrals whereas it is always possible to find derivatives (provided they exist – there are strange functions that do not have derivatives, but we shall not encounter them very much in this text.) Do we need to bother about integration for our modeling? The answer is not very much, we really only need to understand the symbols, not to find actual integrals. Although actually doing the integration is a good aid to understanding. It is the difference between playing with a dog (doing) and seeing its picture (just looking). The symbol for the integral of a function  $u(t)$  is

$$\int u(t)dt.$$

The symbol derives from the letter “S” which is elongated and stylised but stands for “summation”. There are unfortunately several different kinds of integration though they are all the inverse of differentiation in some sense, and all involve summing over some domain or other. The integral of  $2t$  will therefore be  $t^2$  although we add a constant  $c$ , as the derivative of a constant is zero. (The derivative of  $t^2 + c$  is  $2t$  no matter what the value of  $c$ .) The physical interpretation of the first type of integration usually encountered by students involves determining the area under a curve. This is the easiest to understand, but unfortunately it is not the one we need here. In this text we take a curve and divide it up into infinitesimally small straight bits, then sum all these infinitely many bits. The integral of the density from one end of this curve to the other is then its mass and is called a *line integral*. The integrals we actually need in Chapter 3 involve summing over surfaces and volumes rather than just lengths. The area or volume is divided into small chunks and summed over, but the principle is the same and this time the result of the integral is the mass of the surface or volume. The actual rigorous definition stems from the one dimensional definition of the line integral. The integral written

$$\int_{P_1}^{P_2} \rho(x, y, z)ds$$

can be interpreted as follows. Suppose a thin wire is in the form of a curve, and this curve starts at point  $P_1$  and ends at point  $P_2$ . The  $ds$  represents a very small (infinitesimal) arc length which is so tiny that the density is constant along it. There is a limiting process at work here, and in this limit we can say that  $ds$  has in fact zero length so that at each *point* along the wire it has a density  $\rho(x, y, z)$  that depends upon  $x, y$  and  $z$  and is therefore a function of position given by the cartesian coordinates  $(x, y, z)$ . The integral is the total mass of the wire. If  $\rho = \text{constant}$ , then the mass is simply this constant times the length of the wire, so the integral is adding up all the little arc lengths  $ds$ . In this case the mass is the density times the length as expected. (In reality we need to multiply by the very small and constant cross sectional area of the wire to achieve “mass = volume  $\times$  density” of course.) Rather than specify the end points as  $P_1$  and  $P_2$ , the curve is often represented by the letter  $C$  and the integral is written

$$\int_C \rho(x, y, z) ds.$$

If the curve is closed, that is it is a loop without an end there is a special symbol

$$\oint_C \rho(x, y, z) ds.$$

Now you may be wondering how on earth such objects are evaluated. That is, given some functional form of  $\rho$  say  $x^2 + y^2 + z^4/a^2$  or some such, how do we find these integrals? Fortunately, we do not really need to know, but let's just play with the dog a little and do a single simple example. The trick used involves working on the equation of the shape that describes the curve of the wire, then parameterising it in terms of mathematics. For example, a circle of unit radius lying in the  $(x, y)$  plane centre the origin can be parameterised by  $x = \cos \theta$ ,  $y = \sin \theta$ ,  $z = 0$  with  $\theta$  taking the range  $0 \leq \theta < 2\pi$ . If this last bit is gobbledygook, it doesn't really matter. If you really want to work through this example just spend a little time with a textbook to get the appropriate technical details. It is reinforced that such technical details are not actually essential here. Just feel happy if you know about such things, because they do help with analogy and insight. The actual calculation for this example is as follows. If the density is  $x^2 + y^2 + z^4/a^2$  on the wire itself, this can be written in terms of  $\theta$ ; the notation is displayed as Figure 1.2. Of course  $z = 0$ , but we can substitute  $x = \cos \theta$  and  $y = \sin \theta$ . Additionally the arc length of the unit circle is simply  $ds = 1.d\theta$ . The limits of the integral will be  $\theta = 0$  and  $\theta = 2\pi$ ,

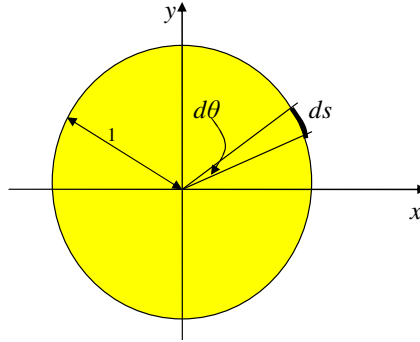


Fig. 1.2 The unit circular shaped wire

geometrically the same point of course, but then the wire is a closed curve, a unit circle. The integral (mass of the wire) is thus:

$$\int_0^{2\pi} (\cos^2 \theta + \sin^2 \theta) d\theta$$

which actually has the value  $2\pi$  using  $\sin^2 \theta + \cos^2 \theta = 1$ . If the curve was a little more twisted or had the odd kink or was actually three dimensional like a coiled spring, then you can see that the technicalities of evaluating such integrals can get overwhelming.

The important feature to glean is that the density is evaluated *on the curve* and that it is the *shape of the curve* and not the density that determines the limits of the integral. This is useful to remember in Chapter 3 and in particular when Finite Elements are described in Chapter 4. So, in this way it is possible to calculate this kind of integral. What about integrals that find the weight of surfaces and volumes? The principle is exactly the same. For a surface, we integrate twice and for a volume we integrate three times, the number of times coinciding with the dimension of the object. In mathematical terms, a surface needs *two* parameters to describe it, so when the density is expressed as  $\rho(x, y, z)$ , it becomes  $\rho(\theta, \phi)$  *on the surface* because the surface itself will be described by a set of equations that relate  $x, y$  and  $z$  to  $\theta$  and  $\phi$ . For example the surface of a sphere  $x^2 + y^2 + z^2 = a^2$  is parameterised by  $x = a \cos \phi \sin \theta$ ,  $y = a \sin \phi \sin \theta$  and  $z = a \cos \theta$ . This means that if  $\rho(x, y, z)$  retains the previous algebraic form of  $x^2 + y^2 + z^2/a^2$  then on the sphere  $x^2 + y^2 + z^2 = a^2$  we have

$$\rho(x, y, z) = \rho(\theta, \phi) = a^2 \sin^2 \theta \cos^2 \phi + a^2 \sin^2 \theta \sin^2 \phi + a^2 \cos^4 \theta.$$

Additionally, the small element of area (patch) on the surface of the sphere is an infinitesimal rectangle of sides  $a d\theta$  and  $a \sin \theta d\phi$ . Hence

$$dS = a^2 \sin \theta d\theta d\phi \text{ with } 0 \leq \theta \leq \pi.$$

So the mass of the sphere is given by the *double* integral

$$\int \int \rho(\theta, \phi) a^2 \sin \theta d\theta d\phi$$

which is

$$\int_0^\pi \int_0^{2\pi} (a^3 \sin^3 \theta \cos^2 \phi + a^3 \sin^3 \theta \sin^2 \phi + a^3 \cos^4 \theta) \sin \theta d\theta d\phi.$$

This will not be evaluated, but a few comments are worth making. The first is about notation. The limits of integration ensure that the surface is swept out just once. Think of the sphere as the earth, then  $\theta$  is the co-latitude (90– latitude) so the range of  $\theta$  sweeps out a line of longitude from north to south pole (180 degrees or  $\pi$  radians). The second integral's limits takes this semicircle and sweeps it around the globe a full circle (360 degrees or  $2\pi$  radians) to form the sphere. The second comment is that the order of integration is dictated by the order in which the surface is described. The convention is to work from the left so we integrate with respect to  $\theta$  first (holding  $\phi$  constant), then we integrate with respect to  $\phi$ . By the second integration, all the  $\theta$ s have disappeared. This order is also indicated by the order of the  $d\theta$  and  $d\phi$  at the end of the expression. We could evaluate this now of course, but by not doing so emphasises that it is not the mathematical technicalities that are important but understanding what the notation means. (In fact for this integral the  $\phi$  integral is very easy as we use  $\cos^2 \phi + \sin^2 \phi = 1$  and for those interested, the final result is  $68\pi a^3/15$ ). The extension to three dimensions is straightforward; we have three integrals to evaluate. Here it is hardly worth parameterising and integrating first with respect to  $x$  then  $y$  then  $z$  can be done directly. In general this will look like

$$\int_{f(x_0,y,z)}^{f(x_1,y,z)} \int_{g(y_0,z)}^{g(y_1,z)} \int_{z_0}^{z_1} \rho(x,y,z) dx dy dz.$$

Again the first integral is respect to  $x$  holding  $y$  and  $z$  constant. Once the integration has been done and the limits (the  $f$  functions) have been inserted, all  $x$ s have disappeared and only  $y$  and  $z$  remain. The second integral is with respect to  $y$ , and similarly when the integration is done and the limits inserted (the  $g$  functions) only  $z$  remains. Finally, the third

integration takes place and the limits are  $z_0$  and  $z_1$  and we get a number answer. The order is the order of  $dx$ ,  $dy$  and  $dz$ . Some books work the integrals from the inside out, but this is considered old fashioned (and poor practise) by mathematicians these days.

That is enough calculus. The other branch of mathematics that needs our attention is vectors. A vector is the name given to a quantity that has direction as well as magnitude. Quantities that only have magnitude and do not have a direction are called scalars. Here are some examples of vectors: force, ocean current and wind. All of these have both direction and magnitude. On the other hand, density, temperature and salinity are scalars because they only have magnitude. In the above discussion of calculus we only integrated scalars (specifically the density) so what about vectors? It turns out that they too can be differentiated and integrated, but first we need notation. Vectors are indicated in one of three ways. If a single letter is to indicate a vector then it is either underlined,  $\underline{a}$  or it is written in boldface  $\mathbf{a}$ . Underlining vectors really belongs in the classroom, so we shall use boldface characters to indicate vector quantities. Geometric vectors are denoted by  $\overrightarrow{AB}$  (the vector joining point  $A$  to point  $B$ ) but these are not used in this text. The third way of indicating vectors is by giving its three components. Components are scalar quantities and represent how the vector quantity is made up. For example we might write  $\mathbf{u} = (u, v, w)$  to indicate that the current  $\mathbf{u}$  has components  $u, v$  and  $w$ . The double use of the letter  $u$  is not confusing as one is a scalar (the component) and the other a vector. Components are simply the proportion of the vector in each of three perpendicular directions. Commonly,  $x$  is east,  $y$  is north and  $z$  is up. So a wind can be specified by how much is due east, how much is due north and how much is vertical (usually very small). Therefore a north east wind (conventionally coming from the north-east) will have equal  $x$  and  $y$  components, and a zero  $z$  component. In mathematical terms the letters  $\mathbf{i}, \mathbf{j}$  and  $\mathbf{k}$  are used to denote vectors in the  $x, y$  and  $z$  directions with magnitude one (called unit magnitude); the vectors of unit magnitude are called *unit vectors* and are a very handy device. We certainly need them in Chapter 3 when deriving the Coriolis term due to the rotation of the earth. It means that we have an alternative and easier way of writing vectors in terms of their components. For example

$$\mathbf{u} = (u, v, w) = u\mathbf{i} + v\mathbf{j} + w\mathbf{k}$$

and the right hand side can be manipulated more easily than the middle triple provided we know some basic rules for adding and multiplying the

unit vectors. To add or subtract two vectors, we simply add or subtract the components; nothing could be more straightforward. Here is a simple example:

$$\mathbf{i} + 2\mathbf{j} + 3\mathbf{k} + \mathbf{i} + 3\mathbf{j} - \mathbf{k} = (1 + 1)\mathbf{i} + (2 + 3)\mathbf{j} + (3 - 1)\mathbf{k} = 2\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}.$$

Again there is no time to really explore the algebra of vectors here, they can be used very fruitfully in geometry and a modern application is in visualization and computer animation. To multiply vectors is less straightforward. In fact there are two different types of product, the scalar product that gives rise to a scalar and the vector product that gives rise to a vector. The definitions here are given in terms of components for two reasons. First they are the only ones used, second they are the easiest to grasp.

The scalar or “dot” product of two vectors  $a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$  and  $b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$  is  $a_1b_1 + a_2b_2 + a_3b_3$ . In other words (symbols in fact)

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1 + a_2b_2 + a_3b_3.$$

One usually does lots of elementary examples in much the same way as when one meets quadratic equations for the first time, but this is not done here for reasons of space as well as relevance. For us, it is more important to know the definition and one or two important facts. One of these is that if the dot product of a pair of vectors is zero then either one of the vectors is zero or they are at right angles to each other. The vector or “cross” product is written  $\mathbf{a} \times \mathbf{b}$  and has the definition

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

For those who are not familiar with determinants, the right hand side multiplies out to

$$(a_2b_3 - b_2a_3)\mathbf{i} + (a_3b_1 - b_3a_1)\mathbf{j} + (a_1b_2 - b_1a_2)\mathbf{k}.$$

The determinant is merely a convenient notation that displays the symmetry. There are only a few important properties of cross products needed here. First of all  $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$ , secondly the direction of  $\mathbf{a} \times \mathbf{b}$  is perpendicular to both  $\mathbf{a}$  and  $\mathbf{b}$  so as to form a right handed system. Using the right hand, if the thumb is aligned with  $\mathbf{a}$  and the forefinger aligned with  $\mathbf{b}$  then the cross product  $\mathbf{a} \times \mathbf{b}$  is in the direction of the second finger if this is held at right angles to thumb and forefinger. The other properties and all the algebraic examples usually associated with this branch of

mathematics will be glossed over. Finally mention needs to be made of the triple products. The scalar triple product of three vectors  $\mathbf{a}$ ,  $\mathbf{b}$  and  $\mathbf{c}$  is

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix},$$

and the result is preserved if either  $\mathbf{a}$ ,  $\mathbf{b}$  or  $\mathbf{c}$  are cyclicly permuted; so the above result is the same as  $\mathbf{b} \cdot (\mathbf{c} \times \mathbf{a})$  and  $\mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$ . It represents the volume of a parallelepiped with three adjacent sides as the three vectors, (a parallelepiped is a solid whose plane faces are parallelograms). The final object needed in this whistle stop tour of vector algebra is the vector triple product, and this is best introduced through the formula

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = (\mathbf{c} \cdot \mathbf{a})\mathbf{b} - (\mathbf{a} \cdot \mathbf{b})\mathbf{c}.$$

With this vector product not only is the order important, but the placement of the parentheses also changes its value (mathematicians say that the expression is not associative). The only time this vector triple product is met is if the centripetal acceleration due to the earth's rotation ever needs to be calculated (see Chapter 3). The term is  $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r})$  where  $\boldsymbol{\Omega}$  is the angular velocity of the earth, and  $\mathbf{r}$  is the vector pointing out perpendicularly from the earth's axis to the surface. Using the formula above, and that the dot product of two vectors that are at right angles is zero gives

$$\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{r}) = -\Omega^2 \mathbf{r}.$$

So we have gone through calculus and skated through vector algebra. The mathematics required to fully understand Chapter 3 combines these in the form of vector calculus. As a simple example, the integral

$$\int_C \mathbf{u} \cdot d\mathbf{r}$$

represents the (infinitesimal) quantity  $\mathbf{u} \cdot d\mathbf{r}$  summed along the curve  $C$ . In fluid mechanics this quantity is called the circulation. It combines the scalar product and the notion of a line integral. The evaluation of such objects can be simplified through the use of vector identities that relate the three vector derivatives called "grad", "div" and "curl". These are defined now. First we have grad  $\phi$  or in full the gradient of  $\phi$ ; this is

$$\nabla \phi = \mathbf{i} \frac{\partial \phi}{\partial x} + \mathbf{j} \frac{\partial \phi}{\partial y} + \mathbf{k} \frac{\partial \phi}{\partial z}$$

and is a vector. The operator  $\nabla$  is called the gradient operator and calculates a spatial gradient of a scalar such as temperature or density. A little

thought will tell you that such gradients must have a direction as well as magnitude, so that they are vectors is not surprising. The next definition is for “div”, or the divergence; this is

$$\nabla \cdot \mathbf{u} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z}$$

where we have used the components  $(u, v, w)$  of the vector  $\mathbf{u}$ . The physical meaning of this is as follows. The vector  $\mathbf{u}$  has a magnitude and direction at every point of a three dimensional domain, much like a wind or sea current. The quantity  $\nabla \cdot \mathbf{u}$  is the amount of  $\mathbf{u}$  being created at each point. Integrated over a volume (domain) it represents how much is being created inside this domain. For the sea (where  $\mathbf{u}$  is current) or the air (where  $\mathbf{u}$  is wind) this is zero because mass can neither be created nor destroyed. We meet this in Chapter 3. Finally we define the “curl” of a vector. The name is not short for anything this time. The definition is:

$$\nabla \times \mathbf{u} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ u & v & w \end{vmatrix}.$$

Written out without using determinants it is:

$$\nabla \times \mathbf{u} = \mathbf{i} \left( \frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \right) + \mathbf{j} \left( \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \right) + \mathbf{k} \left( \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right).$$

It is not intuitively obvious what this might represent. However, if  $\mathbf{u}$  is the current then it is the fluid equivalent of angular momentum in mechanics and represents vorticity or twisting motion. Perhaps this can best be inferred in a two dimensional sense by looking at the  $\mathbf{k}$  or  $z$  component which is

$$\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}.$$

If the northerly current  $v$  increases as we travel east in the  $x$  direction and  $u = 0$  then there could be an anti-clockwise tendency, which following the right hand rule implies a direction upwards (out of the paper in Figure 1.3), that is in the  $z$  direction consistent with the definition of curl. Maybe an actual example helps: let us calculate the curl of a vector that is given by the expression

$$\mathbf{u} = x^2 \mathbf{j}.$$

This corresponds to a vector that is entirely in the  $y$  direction (northwards)

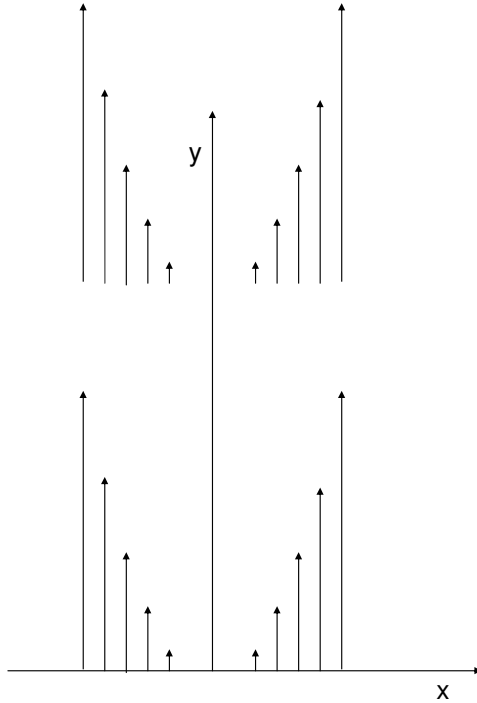


Fig. 1.3 The field  $x^2\mathbf{j}$  expressed pictorially

but increases in magnitude as we travel east provided  $x$  is positive. The curl of this vector (or vector field as it is correctly called) is

$$\nabla \times \mathbf{u} = \mathbf{k} \left( \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) = \mathbf{k} \frac{\partial v}{\partial x} = 2x\mathbf{k}.$$

We see that this is in the  $z$  direction, i.e. upwards and also increases with  $x$  so long as this is positive. If  $x$  is negative then the curl reverses direction and points downwards. This is because although  $\mathbf{u}$  increases with  $x$  positive it also increases with  $x$  negative. Figure 1.3 indicates what is happening here; it displays two manifestations of this vector field, though in reality *every* point of the plane (in fact every point of three dimensional space) will have an arrow the length of which is  $x^2$  and the direction of which is northwards, but this cannot easily be drawn. Note that along the  $y$  (or  $\mathbf{j}$ ) axis the arrows have zero length because  $x = 0$  there.

Now there is a result that is important but has nothing to do with calculating these vector derivatives. To get to grips with it we need the

definitions of grad and curl. Suppose the current vector  $\mathbf{u}$  is of the form of the gradient of some scalar, i.e.

$$\mathbf{u} = \nabla\phi.$$

This means that the components of  $\mathbf{u}$  are

$$u = \frac{\partial\phi}{\partial x} \quad v = \frac{\partial\phi}{\partial y} \quad w = \frac{\partial\phi}{\partial z}.$$

If these expressions for  $u, v$  and  $w$  are substituted into  $\nabla \times \mathbf{u}$  each component becomes zero, and we deduce that

$$\nabla \times \mathbf{u} = \mathbf{0}.$$

Fluids that have velocity vectors that have zero curl are called *irrotational* and have an important place in fluid mechanics. Although irrotational fluids are in some sense idealised their properties do help in the general understanding of how a fluid behaves. It is tempting to believe that the mathematician's concentration on delving into the properties of irrotational fluid mechanics stems from the elegance of the mathematics rather than any practical application. There may be some truth in this, but we will see in Chapter 3 how useful irrotational fluid mechanics can be in studying surface water waves. Advances in aerodynamics could not have occurred without irrotational flow theory which in this application explains simply and elegantly how flow around an aerofoil generates lift; enough lift for aeroplanes to fly. Its role in coastal engineering and near-shore oceanography is less prominent, but it is there. It turns out that if we have an irrotational flow with its zero curl, then the velocity vector  $\mathbf{u}$  must be of the form  $\nabla\phi$  ("going the other way round" to what's above), but this is more awkward to prove – it is not proved here. Physically, the absence of vorticity in a fluid is closely linked with the neglect of friction. If a current is adjacent to a coast, then frictional forces will act and we expect the current to decrease in magnitude as the coast is approached. This is tantamount to injecting vorticity at the coast. If the current is allowed to slip freely against the coast without reduction (think of a straight coast here with a vertical wall) then the vorticity will be zero. These notions will be picked up again first in Chapter 3 then again in later applications.

The final topic in this brief run through relevant mathematics is again conceptual rather than technical. In fluid mechanics, we deal with volumes of fluid that are representative of the fluid as a whole. These volumes are arbitrary and we consider them in the light of getting the fluid to obey physical laws such as Newton's second law of motion or the conservation

of mass. In order to do this, the properties of the fluid are summed (integrated) over the volume, and this integral can represent a useful quantity such as the momentum or mass of the arbitrary piece of fluid. There are relations between such a volume integral (as it is called) and the integral over the surrounding closed surface. The relationship from which others can be derived is called Gauss' Flux Theorem named after perhaps one of the cleverest mathematicians of all time and the founder of the Göttingen school of mathematics in Germany, Carl Friedrich Gauss (1777 - 1855). In terms of the current  $\mathbf{u}$  it takes the form

$$\int_V \nabla \cdot \mathbf{u} dV = \int_S \mathbf{u} \cdot d\mathbf{S}.$$

This has a simple physical interpretation; it states that the amount of fluid being created inside the volume  $V$  must be equal to the flux of fluid across the surface  $S$  of the same volume. Some may worry that  $d\mathbf{S}$  is a vector rather than the scalar that  $S$  is, but  $d\mathbf{S}$  is an infinitesimally small therefore flat bit of area and it is a vector because it has a direction, namely the direction of the normal to it conventionally drawn out of the volume. The dot product between  $\mathbf{u}$  and  $d\mathbf{S}$  indicates the component of  $\mathbf{u}$  out of the surface and integrating this over the entire surface therefore gives the flux.

There is also a relation between the vorticity summed over an open surface  $S$  and the flow around its bounding curve. This is called Stokes theorem, after George Gabriel Stokes (1819 - 1903) who was born in Ireland and became one of the fathers of fluid mechanics, particularly water wave theory. Stated baldly this is

$$\int_S \nabla \times \mathbf{u} \cdot d\mathbf{S} = \int_C \mathbf{u} \cdot d\mathbf{s}.$$

This is valid for any surface with a bounding curve, but it is often used for plane surfaces in which case it becomes the non-vectorial Green theorem in the plane which is perhaps a bit easier to swallow:

$$\int_S \left( \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) dS = \int_C (u dx + v dy).$$

George Green (1793 - 1841) was a self educated Nottinghamshire miller who did some very original mathematics, graduated eventually at age 42 and whose remarkable story reads like some kind of fairy tale until his tragically early death from influenza at 47. This is one of several formulae attributed to him and can be derived by directly integrating the left hand side, but what does it mean? The right hand side is called the circulation of the fluid around  $C$ , the left hand side is the flux of the vorticity through the surface

*S.* The upshot of Stokes theorem for a fluid is that vorticity tends to be conserved in a fluid. A theorem called the Kelvin circulation theorem states that the circulation does not change in time if the curve  $C$  moves with the flow, so Stokes theorem leads to the same conclusion for the flux of vorticity. (Lord Kelvin was Sir William Thomson (1824 - 1907) another Irishman and another of the pioneers of fluid mechanics and much else besides, for example electromagnetism; but a terrible teacher by all accounts.) This leads on to considerations of vortex shedding and quantifying lift around bodies such as aerofoils so taking us well into fluid mechanics and beyond the scope of what we can do here. If you are interested in such things, the book by David Acheson, Acheson (1990) is really good. However, instead of pursuing this we shall now turn to statistics.

## 1.4 Statistical Preliminaries

The origins of statistics lie in ancient times, but in the last century and a half “Statistics” has become a scientific study in its own right, separate from mathematics. It owes this status to the pioneers Florence Nightingale (1820 - 1910) (yes that one – she was a very able mathematician), Sir Francis Galton (1822 - 1911) an explorer and meteorologist who became the father of eugenics, selecting parents to “improve” physical and mental abilities of the child, a very dirty word nowadays, and in particular Karl Pearson (1857 - 1936). Briefly, before these eminent people, statistics was permutations and combinations together with mathematical probability and some fitting of data to predetermined lines, with only the Reverend Thomas Bayes (1702 - 1761) making a foray into inference in the 18th century. It was really Pearson who invented modern statistics as the science of making sense of large data sets. Later Ronald Fisher (1890 - 1962) developed the alternative view of being able to deduce and infer from sampling, working directly with the data, and it was the brilliant Russian Andrey Nikolaevich Kolmogorov (1903 - 1987) who actually showed the rigour of this new statistics and gave it mathematical form. In this text advanced mathematical structure will be avoided and the basic ideas will be introduced in a very applied way.

There is no doubt that statistics plays an increasingly important role in marine science and coastal engineering; one could even say a pivotal role. The principal difficulty in writing a text such as this is to cater for the wide variety of previous experience amongst the readership. The safest path to take is to assume very little previous knowledge. Those

who have managed to get through the last section with a modicum of understanding will already be mathematically quite sophisticated, in which case do pick and choose from what follows. We shall start with the revision of what statisticians call *measures of central tendency*, which means ways of assessing where the middle of a set of data is. The simplest form of data is a list of numbers, although data are also often produced in the form of frequency tables. We shall deal with both.

**Example** We wish to find the mode, median and mean of the following list of numbers:

5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4.

**Solution** First of all, do not worry about the definitions of these words, this will be addressed later; instead, we put the numbers in ascending order as follows:

2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 8, 8, 8.

The mode is the number that appears the most times. So the mode = 5. The median is the number which is in the middle of the distribution, which as the number of numbers is even is a bit tricky, so we'll come back to this. Finally, the mean of the numbers is the sum of the numbers divided by 20 (there are 20 numbers in all), so the mean is computed as  $96/20 = 4.8$ . In this example, there is a clear mode since there are six 5's, and fewer of each of the other numbers (in general there is often a tie). There is an even number (20) of numbers, therefore the median is the average of the tenth and eleventh numbers. Since both of these are 5, so is the median. The mean is, uniquely, 4.8.

Next let us consider something a little more usual in scientific applications, that is, a situation where the numbers are grouped into classes and we have what is called a frequency distribution. Frequency distributions are usually given in tabular form. Table 1.1 gives the numbers of zooplankton of various lengths as measured by a student marine biologist (this is adapted from research data and considerably simplified). The frequency polygon associated with these data is shown in Figure 1.4.

The median of these data is still the middle number, but this is troublesome to find when the data take this form. There is a formula that we will give later, but it is best to draw the graph. The median is then given by the value taken on the horizontal scale when a vertical line precisely divides the area under the frequency polygon into two equal halves. The mode is the peak of the frequency polygon (there may be more than one).

Table 1.1 A Frequency Table.

Length of zooplankton (mm)	Number of zooplankton
0.01-0.50	4
0.51-1.00	10
1.01-1.50	15
1.51-2.00	13
2.01-2.50	7
2.51-3.00	1

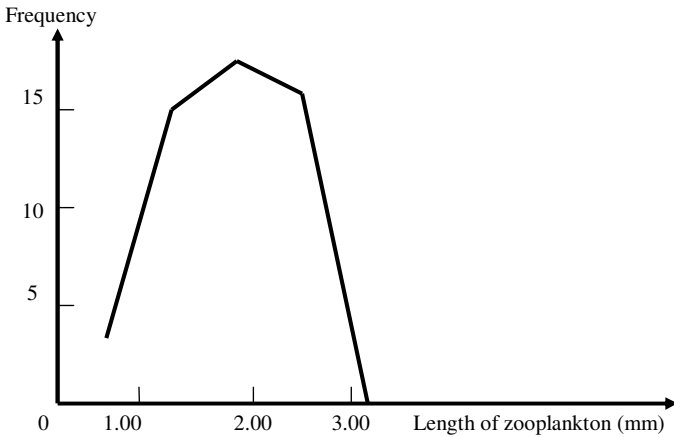


Fig. 1.4 A frequency polygon

The mean is the quantity  $\mu$ , which is given by the formula:

$$\mu = \frac{\sum f_i n_i}{\sum n_i},$$

where the letters  $f$  and  $n$  denote the frequency of occurrence of the number, and the number respectively. The subscript is there to designate that there are many numbers ( $i$  would run from one to six in our example) and the  $\sum$  sign denotes that summation over all  $i$  is to occur. The mean and median are always uniquely defined, but the same cannot be said of the mode. The mode is, straightforwardly, the class that contains the largest number but this might not be unique. The median either has to be determined graphically, or by a rather messy formula derived from its definition as being the “middle”. If the median occurs in a particular class, and the lower boundary of this class is  $L$ , then the median itself is determined from

the formula:

$$\text{Median} = L + \left( \frac{N/2 - (\sum f)}{f_{\text{median}}} \right) c,$$

where  $N$  is the total number of items in the data,  $\sum f$  is the sum of frequencies of all classes *lower* than the median class,  $f_{\text{median}}$  is the frequency of the median class, and  $c$  is the size of the median class interval. Given grouped data, it is easy enough to spot in which class the median lies; all the above formula represents is a mathematically precise way of dividing the area of this class to ensure that the median line so derived cuts the total area under the frequency polygon precisely in half. The results for this particular data set are mode = 1.255 mm, mean = 1.375 mm and median = 1.3969 mm. In this problem we meet several features that are typical in the handling of data. The mode is simply the mid-point of the interval (1.01-1.50) that contains the greatest number of animals. The mean follows by applying the formula remembering that in this instance the number of animals is multiplied by the length of zooplankton corresponding to the *middle* of the range (for example,  $4 \times 0.255$  is the first entry in the numerator,  $10 \times 0.755$  is the second, etc.). Finally, the median is calculated using the given formula with  $L = 1.01$ ,  $c = 0.49$ ,  $\sum f = 14$  and  $f_{\text{median}} = 15$ . This gives the idea of how these measures of central tendency can be calculated. Of course, these days a calculator, laptop or PC takes away the need to do the arithmetic.

The “middle” is not the only parameter that characterises a set of numbers. The two sets of numbers 1, 2, 3, 4, 5, 6 and 2, 3, 3, 4, 4, 5 both have 6 members and a mean of 3.5, but they are not the same. They differ on how the numbers are spread about the mean, and this is a second important characteristic of data. The usual measure of spread is variance or its square root, standard deviation. There are other more subtle measures. In a textbook on marine applications such as this it is not possible to go into much detail in the way of statistical theory, nor would it be desirable. The many specialist texts on statistics that start as we have by introducing measures of central tendency, go on to discuss topics such as standard deviation, distributions, probability, and then to applied topics which include sampling, regression, hypothesis testing and experimental design. All of these have a role to play in marine science and coastal engineering, and we will give them a brief airing in later paragraphs, but it would be over-ambitious to try to cover them in any depth in this book. Perhaps the most important point to make is that the central purpose of statistics is *inference*. The reason why data are analysed is to enable scientists and engineers to es-

tablish hypotheses (in a statistical sense) from the data. It is however also necessary to give some ideas about probability as these form a central part in the understanding of wave spectra that will be met in Chapter 6. Before doing this, having defined and calculated measures of central tendency as statisticians call them, lets discuss measures of spread in more detail.

As mentioned above, the most common of these is called *variance* together with its square root *standard deviation*. The variance of a set of numbers measures how spread out they are from their mean. It is defined by the formula:

$$\sigma^2 = \frac{\sum(X_i - \bar{X})^2}{N},$$

where the symbols have the following meanings:  $X_i$  denotes the data (i.e. the numbers themselves),  $\bar{X}$  is the arithmetic mean,  $\sum$  is the summation sign which means that each number has the mean subtracted from it before it is squared, then the whole is divided by  $N$ , the number of numbers in the data set. The reason behind squaring each difference is that this makes all entries under the summation sign positive, hence making sure that the result of this sum is indeed a true representation of the spread of the data from the mean. Statisticians call this a “measure of dispersion”, but this is not an appropriate expression to use in a book where dispersion has a physical significance of its own. In order to restore the dimensions, the variance is normally square rooted (hence the square on the left-hand side) and the symbol  $\sigma$  is called the standard deviation.

Again, computers and calculators take away the arithmetical drudgery, but let us go through a simple example:

**Example** Find the variance and standard deviation of the numbers:

$$5, 3, 6, 5, 4, 5, 2, 8, 6, 5, 4, 8, 3, 4, 5, 4, 8, 2, 5, 4.$$

**Solution** The answers are  $\sigma^2 = 3.116$  and  $\sigma = 1.765$ . If you “cheated” and used a calculator or a computer, this is no problem as long as you are sure of what you have calculated and know what standard deviation and variance actually indicate. The above answers only validate your arithmetic; they do not confirm your understanding. When a frequency table is involved, the definitions are of course the same but the method of calculation looks a little different. In fact, there is a very useful formula that can be derived from the definition of variance that proves useful in calculation. This states that the variance is given by the expression:

$$\sigma^2 = \overline{X^2} - \bar{X}^2,$$

which can be read as variance equals the ‘mean of the squares minus the square of the mean’. For grouped data, the following expression is the formula for variance:

$$\sigma^2 = \frac{\sum fn_i^2}{N} - \left( \frac{\sum fn_i}{N} \right)^2;$$

the standard deviation is of course the positive square root of the variance. The variance and standard deviation of the data presented in Table 1.1 are  $0.408 \text{ mm}^2$  and  $0.639 \text{ mm}$  respectively. The calculation of the mean and standard deviation of a set of data is one thing and is easy to do. In a practical context, the scientist (usually a marine biologist) should now go on to discuss the implications of these values. Before leaving means and standard deviation, a word needs to be said about how to compare two (or more) data sets. Everything said above has been about analysing a single data set. However, it is very common to have to compare two or more sets of data. In a marine context there are many examples: sea temperature and/or salinity and the biomass of some phytoplankton, wind speed and direction at different locations, etc. More will be said later on the detail on how such comparisons are done, but the direct measure of comparison of two sets of numbers is a simple extension of variance called *covariance*. Here is the definition:

$$\sigma_{xy}^2 = \frac{\sum (X_i - \bar{X}) \sum (Y_i - \bar{Y})}{N}$$

where  $Y_i, i = 1, 2, \dots, N$  denotes the second set of numbers. In some books the denominator in the definition of both variance and covariance will be  $N - 1$  and not  $N$ , this is because the definition is for a sampled set and not the entire set of data, and this difference pops out of the mathematics one goes through to define a variance for the sampled set that closely matches the real variance of the entire data set. Do not worry about it as for large  $N$  they are very close, and if  $N$  is not large any inference one makes will be unreliable anyway. We will return to covariance later when discussing regression and Principal Component Analysis.

Let us now discuss probability. The notion of probability is tied up with the outcomes of things called events. An example of an event is the tossing of a coin or the drawing of a card from a pack. It can also be the occurrence of a particularly large wave. Defining the *probability*  $p$  of an event occurring as

$$p = \frac{\text{number of ways it can occur}}{\text{total number of ways}}$$

is fine for coin tossing or card drawing, but is not very useful for practical problems such as wave forecasting. A more workable definition might be to use a large number of trials and define

$$p = \frac{\text{number of successes}}{\text{total number of trials}}.$$

The problem with this definition is that the probability calculated only approximates the actual likelihood of the event occurring. For example, no matter how many trials there are, tossing a fair coin will never result in exactly half being heads and half being tails. (One hesitates to say never, but if it occurred one would suspect tomfoolery.) In most textbooks, the theory of probability is housed in terms of set theory and Venn diagrams. A set  $A$  will contain events and the function  $p(A)$  the probability of event  $A$  occurring. The set  $S$  is the universal set that contains all possible outcomes, and of course  $p(S) = 1$  and  $A \subseteq S$  and so on. We shall not go this route here. The last chapter of James (2001) is a good introduction if you are keen. This is the third edition but there will be another out soon no doubt. However, there are some aspects of probability that are well worth covering. In this text we shall primarily be concerned with continuous rather than discrete variables. This is because the processes that interest us such as trying to determine the probability of particularly large waves stem from continuous processes (there is a continuum of waves in between the smallest and largest), however the following discrete example serves as a useful introduction.

In many practical examples of discrete probability there are only two possible outcomes from an event. Most events can be considered either a success or a failure. This gives the two possible outcomes and two associated probabilities,  $p$  and  $1 - p$  linked with each and the trial is called a Bernoulli trial. Certainty has the probability 1 and impossibility the probability 0. This follows from either of the definitions given above. For the toss of a fair coin, if  $p$  is the probability of getting a head then  $p = 1/2$ , disallowing the coin landing on its edge. There is a reasonably simple formula that gives the probability of  $r$  successes in  $n$  trials. If we put  $q = 1 - p$  then expand  $(p + q)^n$  in a binomial expansion then the  $r$ th term gives the probability of  $r$  successes in  $n$  trials. Mathematically,

$$(p + q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{2!}p^{n-2}q^2 + \dots + \binom{n}{r}p^r q^{n-r} + \dots + q^n.$$

The  $r$ th term is:

$$\binom{n}{r}p^r q^{n-r}.$$

It is the probability of  $r$  successes multiplied by the probability of  $n - r$  failures multiplied by the number of ways that this particular arrangement can occur in  $n$  trials. This arrangement is written using the “ $n$  choose  $r$ ” notation which is defined by:

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

As each arrangement of success or failure is independent, if we add up all possibilities we regenerate the binomial expansion of  $(p + q)^n$  which is of course 1 since  $q = 1 - p$ . It is not difficult to see direct applications of Bernoulli trials in marine biology. Detecting the presence of diseases in phytoplankton or zooplankton for example (success – no disease, failure – disease) is one example that belongs in Chapter 7. Whether or not a toxic level of a contaminant is present in a river, estuary or coastal sea is another example that belongs in Chapter 5. The tossing of coins will be what are called independent events, that is the probability of getting a head or a tail is one half no matter what the previous results are. Drawing cards from a pack without replacement is a different matter. The chances of selecting the ace of spades for example will be  $1/52$  from a full pack but after not drawing it will reduce to  $1/51$  for the next selection. Once the ace of spades is in your hand, the chances of getting another one drops to zero of course (assuming a standard deck; and no card sharps). In many practical examples the independence of events is assured due to the scientific stringency of how the observations or experiments are carried out. However, sometimes events are not independent. For example, suppose the events are  $A$ : “the occurrence of waves of height over  $10m$ ” and  $B$ : “the occurrence of storms where the winds exceed  $50ms^{-1}$ ” then these are obviously not independent. The probability of  $A$  occurring will be smaller than the probability of  $A$  occurring given that  $B$  has occurred. Sometimes it will be larger; consider if we replaced  $B$  by “wind is less than  $20ms^{-1}$ ”. In order to distinguish between the two cases, the notation for the former is the straightforward  $p(A)$  whereas for the probability of  $A$  given  $B$  it is  $p(A|B)$  and there are rules linking the two, in fact:

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

where  $p(A \cap B)$  is the probability of both  $A$  and  $B$  occurring. We have no room to explore such subtleties, but hopefully this has made you think about the use of statistics a little more deeply.

For the discrete Binomial distribution, the mean is  $np$  and the variance  $np(1 - p)$ , but to go any deeper into pure probability theory would not

be appropriate here; we do return to the binomial distribution when we discuss maximum likelihood a little later. Meanwhile, note that if  $n$  gets very large, it becomes too unwieldy to deal with and instead the Poisson distribution is used whereby:

$$y = \frac{\lambda^n e^{-\lambda}}{n!}$$

with both mean and variance equal to  $\lambda$ . In other circumstances continuous distributions are required at the outset. In Chapter 6 various distributions will be met that are very useful when considering the prediction of large waves. These distributions are special to the subject. Most of the rest of the world either meet the (discrete or continuous) Normal distribution or the Poisson distribution. The formula for the Normal distribution is:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty,$$

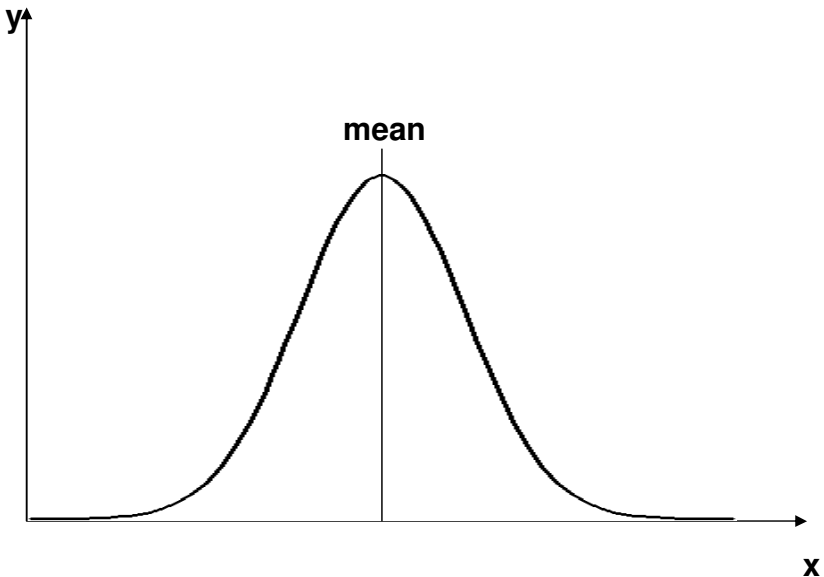


Fig. 1.5 Gaussian profile

with mean  $\mu$  and variance  $\sigma^2$ . The shape is shown in Figure 1.5. Here  $y$  is in what statisticians call the sample space which means it is wave height, or some other quantity that we require to estimate. The random variable  $x$  can be time or space, but most commonly in our applications it will be

the frequency of the wave. In Chapter 6 the spectrum is indeed defined by a functional form giving its distribution with frequency. Figure 6.3 in Chapter 6 is a good example. Once the mean and standard deviation are fixed, in many diverse applications away from wave prediction it is often acceptable to assume that data presented in the form of a frequency distribution approximate closely to normal. Believe it or not, this assumption tends to be universal. In fact the normal or Gaussian distribution tends to be assumed even when it is not appropriate; users of statistical routines need to be aware of this. The first thing to remember is that the normal distribution assumes that the variation of frequency  $y = f(x)$  with random variable  $x$  is the above function. Although the normal distribution is very widely used, the particular mean and standard deviation of each problem is going to be different. It is obviously desirable to have only a single standard normal distribution to cater for predictions, and this is easily accomplished by adjusting the variables to a mean of zero and a standard deviation of one. Remember therefore to transform your data  $X$  into the normal variable  $z$ , sometimes called the  $z$ -statistic, through the simple transformation:

$$z = \frac{X - \mu}{\sigma},$$

before doing any statistical testing, and testing is what it's all about. The reason for proposing a normal distribution is to test whether your particular data fits this distribution. The commonest of tests to use is the  $\chi^2$  test, which can be used to test whether or not a particular set of data fits a given hypothesis. In order to get the idea of testing a hypothesis, let us return to tossing a coin. If a given coin is tossed 1000 times, say, and the outcomes are recorded, then this test can be used to decide whether the coin is biased or fair. Similarly, the  $\chi^2$  test can be used to decide whether or not data fit the conclusions drawn from a particular model. As hinted at above, however, one never gets *the* answer, and the criterion for acceptance or rejection of a hypothesis, to the applied marine scientist or coastal engineer, is not God-given but is in fact dependent on assumptions involving the normal distribution.

Before we can do examples, we need to introduce the subject of hypothesis testing a bit more formally. This is the traditional first step on the road to *inference*, the main purpose behind most of statistics. Suppose we have some data, perhaps from observations taken on a field trip. These data form what statisticians call a population. It is a collection of numbers arranged in a table or represented graphically. There will be certain statistics associated with the data - we have calculated the mean and standard

deviation, but there are others. Now suppose further that we suspect that these data obey the form dictated by, say, the normal distribution. That is, we suspect that the mean and standard deviation conform to a certain normal, bell-shaped curve. We can use the  $\chi^2$  test to ascertain the truth of this hypothesis. This hypothesis is called the *null hypothesis* and is given the symbol  $H_0$ . If  $H_0$  is rejected when in fact it is true, we say that a type I error has occurred. If we accept  $H_0$  when it is actually false, we say that a type II error has occurred. Unfortunately, it is all too easy to make both sorts of errors, and it is always best to take a cynical look at the data, looking for oddities (*outliers* as statisticians call them) which may be due to human error in observing, or instrumental failure, and which could distort the data and be the underlying cause of the type I or type II error. Finally, statisticians give the symbol  $H_1$  to an alternative to the null hypothesis. Hopefully some of this will come alive through the next two examples.

The first of these examples is an introductory one involving as said above that old standby, the tossing of coins; the second is a more practical example involving real marine data.

**Example** Suppose a coin is tossed 1000 times, and the outcome is 530 heads and 470 tails. We might expect the outcome to be 500 heads and 500 tails, but then again it is the nature of chance that most of us would actually be surprised at such a precise obedience of the laws of probability. The pertinent question to ask is: is the coin fair? In other words, can the deviation from the ideal answer be attributed to chance, or is there a bias in the coin? In this case, the null hypothesis might be:

$H_0$ : heads and tails occur with equal frequency.

**Solution** We shall use the  $\chi^2$  test. In order to do this, we need an appropriate distribution. The  $\chi^2$  distribution can be found on the web (I found one at <http://rvgs.k12.va.us/statman/Table-C.jpg> for example, but type “chi-square table” into Google). In the  $\chi^2$  table, the top row, labelled  $\chi^2$  which denotes the *levels of significance*, gives a choice of thirteen numbers. These numbers represent significance levels so that, respectively, the columns that they head are appropriate to testing at the 99.5%, 99%, 97.5%, 95%, 90%, 75%, 50%, 25%, 10%, 5%, 2.5%, 1% and 0.5% levels. Let us choose the value 0.01, so that we are testing at the 99% significance level. Coin tossing is a process that has two possible outcomes (heads or tails); therefore the first row is chosen. The number in this row is  $\chi^2 = 6.635$ . Now we calculate the value of  $\chi^2$  according to the formula:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Remember, the summation sign is not a sum over 1000 trials, but a sum over all possible outcomes. The calculated value is:

$$\chi^2 = \frac{(530 - 500)^2}{500} + \frac{(470 - 500)^2}{500},$$

so that  $\chi^2 = 3.6$ . This value is less than the value in the table, so we accept the null hypothesis  $H_0$  and conclude that, at 99% significance level, the coin is not biased. This is probably the correct conclusion, but if on examining the data we found 200 consecutive heads, we would want to research further into how the coin was tossed, etc. This latter point may seem a little silly here, but if we were dealing with real data, it is analogous to “eye-balling” the figures and spotting if anything suspicious is present in the data. Mind you, one is much more likely to look if the hypothesis is rejected.

To the relief of most of you, the next part comprises a marine related example.

Here is an example involving fish. Table 1.2 gives the actual and expected values for catches of five species of fish.

Table 1.2 Actual and Expected catches of five species of fish

	Species A	Species B	Species C	Species D	Species E
Expected catch	25	5	7	31	35
Actual catch	20	4	17	26	30

First, the null hypothesis for this problem  $H_0$  states that: “the expected catch and the actual catch are the same”. Using a  $\chi^2$  test with parameter 0.01, do we reject  $H_0$ ? To answer this we calculate  $\chi^2$  from the formula and get the appropriate value of  $\chi^2$  from a  $\chi^2$  table; these are

$$\chi^2(\text{calculated}) = 17$$

$$\chi^2(\text{table}) = 13.3.$$

On the face of it, these results indicate that we should reject the null hypothesis. However, if we glance at the table of data, there is a very large discrepancy between expected and actual catch for species C. Without species C data the calculated value of  $\chi^2$  would have been well below 13.3 and  $H_0$  would have been accepted. The correct conclusion to draw therefore is that the figures for species C need to be re-examined and the reason for the glut of fish or the serious under-estimation of the catch ascertained. In passing, note that for an  $n$ -variable problem ( $n = 2$  for the coin, and  $n = 5$  for the fish) we look at the line  $n - 1$  rather than line  $n$  in the  $\chi^2$  table. The

reasons for this are rather technical and have to do with the (statistical) degrees of freedom of the system.

When using hypothesis testing, it is always necessary to put data in the form of frequency. The  $\chi^2$  test simply does not work for dimensional data in the form of lengths or masses. If your data is in such a form, classify them in some way; batch them up to rid the numbers of dimension. Once the data is put into these  $m$  classes, then the degrees of freedom, the row along which to look up the value in the  $\chi^2$  table is  $\nu = m - k - 1$  where  $k$  is the number of parameters to be estimated (usually zero for us). Let us go no further here; interested readers are directed to statistics books - there are plenty to choose from. (The statistics chapters James (2001) are particularly accessible.)

### 1.4.1 Maximum Likelihood Estimation

At the beginning of this section about statistics, certain names were mentioned as pioneers of the subject. Maximum Likelihood Estimation (MLE) belongs securely in the Ronald Fisher (1890 - 1962) realm whereby sampling methods are used. He first used it in the context of agriculture around the time of the first World War, but it remains a useful technique. The idea is to be able to estimate the value of some parameter given the outcome of a number of trials and assuming that these trials fit a particular distribution. In keeping with this section, general theory is avoided (it looks very daunting) and instead we do some examples to get the general idea.

**Example** Suppose a die is thrown 12 times with the result: 5, 6, 6, 4, 2, 6, 6, 6, 3, 1, 6, 6. It is suspected that there is a bias. What is the maximum value of the probability of getting this result?

**Solution** Assuming  $p$  is the probability of throwing a 6. Then the probability of getting 7 sixes and 5 other numbers in 12 throws will be

$$\binom{12}{7} p^7 (1-p)^5.$$

In this example, the order that the 7 successes come is not relevant. To maximise this for  $p$  we need to differentiate and put the derivative equal to zero. We could differentiate  $p^7(1-p)^5$  as it stands, but the trick is to take logarithms and differentiate these as follows:

$$y = p^7(1-p)^5$$

$$\text{so } \ln y = 7 \ln p + 5 \ln(1-p)$$

$$\text{differentiating, we get } \frac{1}{y} \frac{dy}{dp} = \frac{7}{p} - \frac{5}{1-p}$$

which has to equal zero in order for  $p$  to be a maximum. Hence

$$p = \frac{7}{12} = 0.5833.$$

Now if the dice was a fair one, the probability of throwing a 6 or any other number would be  $1/6$ . The value 0.5833 seems to indicate that with this die, there is a better than even chance of throwing a 6. Note the phraseology here; in particular the use of “seems” and the presence of “likelihood” in MLE. Nothing is certain, and it just might be the case that the 12 throws were a lucky sequence as far as throwing sixes were concerned.

Here is a different kind of example. Suppose that samples are taken from a population that has a normal distribution. Normal distributions have two free parameters the mean usually denoted by  $\mu$  and a variance usually denoted by  $\sigma^2$ . (The square root of variance,  $\sigma$ , is the standard deviation.) For this example, suppose that the variance is known but that  $N$  samples are taken. Let the sizes of all these samples be given by the numbers  $X_1, X_2, \dots, X_N$ . By definition, the probability density function of each of these samples will be:

$$f(X_i|\mu) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

where each sample has the same mean and variance as the population from which the samples are taken. However we do not know the value of  $\mu$  and we can use the MLE to estimate this. If all the samples are independently taken, then the likelihood function is simply the product of the density functions as follows:

$$\begin{aligned} L(X_1, X_2, \dots, X_N) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_1-\mu)^2}{2\sigma^2}} \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_2-\mu)^2}{2\sigma^2}} \times \\ &\dots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X_N-\mu)^2}{2\sigma^2}}. \end{aligned}$$

The right hand side can be tidied up to give

$$L(X_1, X_2, \dots, X_N) = \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^N e^{-\sum_{i=1}^N \frac{(X_i-\mu)^2}{2\sigma^2}}.$$

As before, logarithms are taken before the right hand side is differentiated with respect to  $\mu$  to estimate its likelihood. So:

$$\ln(L) = N \ln \left( \frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{\sum_{i=1}^N (X_i - \mu)^2}{2\sigma^2}$$

and differentiating gives:

$$\frac{d}{d\mu}(\ln(L)) = \frac{\sum_{i=1}^N (X_i - \mu)}{\sigma^2}$$

noting that the first term does not contain  $\mu$  so differentiates to zero, and differentiating the second term gives rise to a series of terms in which the 2's cancel and the minus sign disappears. Setting this quantity to zero gives

$$\sum_{i=1}^N (X_i - \mu) = 0$$

or

$$\sum_{i=1}^N X_i - \sum_{i=1}^N \mu = 0.$$

The second of these terms is simply  $N\mu$  thus we have

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

as the Maximum Likelihood Estimation for the mean. Thus for populations having a normal distribution, the MLE of the mean is simply the mean of the sample means of the population; no real surprise, but quite convenient.

### 1.4.2 Regression

The next topic to cover in this briefest of excursions into statistics is fitting lines to data. The most common example of this is the regression line, which is a line of best fit through a set of data points. Some time will be spent going through the principles of regression. After this, there will be a discussion of more sophisticated techniques for fitting data, these include EOF (Empirical Orthogonal Functions) which is really part of PCA (Principal Component Analysis) which itself is in fact a special case of Factor Analysis (which one hesitates to abbreviate to FA).

Let us start with a simple scatter plot as shown in Figure 1.6, for which there is a quite straightforward procedure for drawing a line of best fit through the data. An arbitrary line is drawn, then the square of the perpendicular distance of each point from this line is calculated. These are all added together, and the minimum value of this is found. The parameters of the line that correspond to this minimum value give the line of best fit. Difficulties arise only when the data are so scattered that there is virtually zero correlation, in which case the line of best fit has no meaning. In fact, there are always *two* regression lines. If  $x$  and  $y$  denote the standard axes, these regression lines are called “ $y$  on  $x$ ” and “ $x$  on  $y$ ”, and if there is no correlation then these two regression lines are at right angles to each other. The presence of these two lines was one reason for the more complex PCA

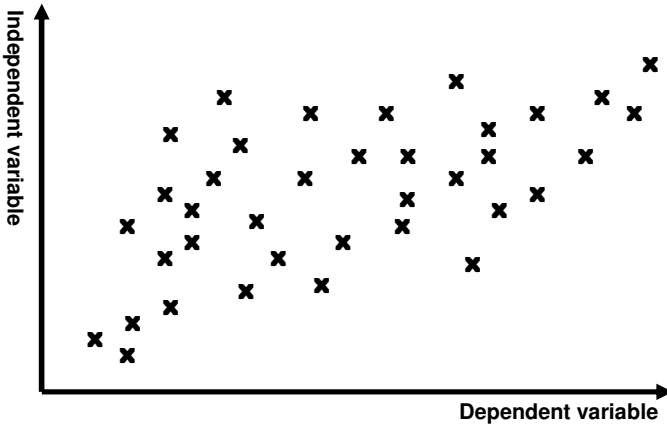


Fig. 1.6 A scatter diagram

and Factor Analysis. There will be more about PCA later. Recall that the term correlation refers to the measure of agreement between two sets of data. A correlation of 1 denotes perfect agreement, a correlation of  $-1$  denotes perfect disagreement (as an example of the latter, the rainfall at one point of an estuary, and the salinity of the water at the same point; as the rainfall increases, the salinity decreases and vice versa) and a correlation of zero denotes no relationship at all. Other complications occur when there is obviously a relationship between two variables, but this relationship is not a linear one. This takes us into log-linear and log-log plots. These days most schools seem to have abandoned logarithms because they are no longer of any practical use as a calculating tool. Teachers think that they have gone the way of the *ready reckoner* and the *slide rule*. Fortunately they have not gone completely as it is recognised that logarithms have another function that has not and is never likely to be superseded - they are used to represent data where some kind of exponential growth is taking place. Those students that need to know about such things, such as students of biology and marine science, may thus be faced with logarithms for the first time. The good news is that fortunately, there is no need to dwell at length on the many properties of logarithms, just a few; all that is necessary in fact is given below.

If an animal is growing exponentially, then its weight  $w$  might be related to time  $t$  through a relationship such as:

$$w = a + b \exp(ct)$$

Table 1.3 Some common relationships

Equation	Straight line	Description
$Y = \frac{1}{a + bX}$	$\frac{1}{Y} = a + bX$	A hyperbola: use ordinary linear spreadsheet
$Y = ab^X$	$\ln Y = \ln a + X \ln b$	An exponential curve: use a log-linear relation
$Y = aX^b$	$\ln Y = \ln a + b \ln X$	Geometric curve: use a log-log relationship
$Y = \frac{1}{ab^X + c}$	$\frac{1}{Y} = ab^X + c$	Logistic curve: use a log-linear relationship (with care)

where  $a$ ,  $b$  and  $c$  are known constants that are fixed once the species and its environment are also fixed. In reality of course this growth will stop, and these more sophisticated models are introduced in Chapter 7, but this is a simple illustration only. If we wanted to make  $t$  the subject of this formula, then we would subtract  $a$  from both sides before taking logs to obtain:

$$t = \frac{1}{c} \ln \left( \frac{w - a}{b} \right),$$

where the symbol ‘ln’ denotes the natural or Napierian logarithm. This particular logarithm function is the inverse of the exponential function, and is the ‘log’ referred to in the phrase ‘log-linear’ as in graph paper. We have still not given the reasons for needing to know about such graph paper. To do so, consider the expression just derived,

$$t = \frac{1}{c} \ln \left( \frac{w - a}{b} \right).$$

If data corresponding to  $(w - a)/b$  were to be plotted on one axis of log-linear paper, and data corresponding to  $t$  be plotted on the other, then provided  $w$  and  $t$  were related in the way dictated by the above equation, the plot would be a straight line (with slope  $c$ ). Once a scatter plot can be assumed to contain within it an implied linear relationship, then all the regression methods developed for straight lines can be brought to bear on the data. Table 1.3 gives some examples of relationships and the correct page of a spreadsheet (for example the graphics facility of Excel) that should be used to display them as a straight line. In what follows,  $X$  and  $Y$  are the independent and dependent variables, respectively.

There is specialist software that renders variables that are related logarithmically (the last entry in Table 1.3) as a straight line. However a log-linear

relationship can be used provided the equation is transformed into exponential type by treating  $(1/Y) - c$  as a variable.

One important question we have not yet addressed is how to assess whether or not a particular law is suitable for a given set of data; we cannot always rely on simply “eye-balling” it.

If we wish to compare two sets of figures in a quantitative manner, then we calculate a correlation coefficient. There are several such coefficients to choose from, but the one most commonly used is the Pearson correlation coefficient, which is 1 for perfect agreement,  $-1$  for perfect disagreement, and 0 for no relationship at all. To calculate the Pearson correlation coefficient,  $r_{XY}$ , the formula:

$$r_{XY} = \frac{(1/N) \sum [(X_i - \bar{X})(Y_i - \bar{Y})]}{S_X S_Y},$$

where

$$S_X^2 = \frac{1}{N-1} \sum (X_i - \bar{X})^2 \text{ and } S_Y^2 = \frac{1}{N-1} \sum (Y_i - \bar{Y})^2,$$

is used. All summations are over all the data points. As mentioned on page 34, the presence of  $N - 1$  rather than  $N$  in some of these expressions may perplex some readers, but as stated there when  $N$  is large this difference is unimportant, and when  $N$  is small any correlations will have large uncertainty anyway. Although the above formula gives the definition of  $r_{XY}$ , we give below the most widely used practical formulae for calculating not only  $r_{XY}$  but also the regression line of  $Y$  on  $X$  in the form  $Y = AX + B$ . Purists will also notice a missing factor of  $(N - 1)^2/N^2$  in the formula for  $r_{XY}$ , but again this quantity is very close to one in most practical examples. In fact, if it is not, then any straight line drawn through such sparse data has only scant value.

$$r_{XY}^2 = \frac{(N \sum XY - \sum X \sum Y)^2}{|N \sum X^2 - (\sum X)^2| |N \sum Y^2 - (\sum Y)^2|},$$

$$B = \frac{N \sum XY - \sum X \sum Y}{N \sum X^2 - (\sum X)^2}, \quad A = \frac{\sum Y - B \sum X}{N}.$$

Let us now do an example.

**Example** Table 1.4 gives the discharges of nitrogen (N) and total phosphorus (P) through the River Göta in tonnes per year, as measured in the years 1972-82 (inclusive). the quantity  $Q$  denotes the river discharge in  $m^3 s^{-1}$ .

**Solution** First, we need to plot the two scatter diagrams of the discharges of nitrogen and phosphorus. These are shown in Figures 1.7 and

Table 1.4 A Table of river discharge data

Discharge	1972	1973	1974	1975	1976	
N(t/yr)	13600	8900	13500	12700	7000	
P(t/yr)	310	210	250	220	120	
$Q(m^3s^{-1})$	150	365	505	515	240	
Discharge	1977	1978	1979	1980	1981	1982
N(t/yr)	16700	14900	13600	18700	18000	16400
P(t/yr)	350	290	310	390	330	270
$Q(m^3s^{-1})$	535	535	435	645	620	535

1.8, respectively. The variable  $Q$  is the independent variable, and it is seen that the data are suitable for a linear regression line to be appropriate. Calculate the two correlation coefficients  $r_N$  and  $r_P$  using the formula to obtain  $r_N = 0.73$ ,  $r_P = 0.53$ .

Although both correlations are positive, they are not particularly high, so it is not obvious that linear regression is the best way to obtain reliable predictions. One may find a better non-linear relationship, but looking at scatter plots does not immediately suggest any obvious alternative candidates. We therefore still press ahead and calculate the linear regression lines, but bearing in mind that predictions need to be treated with some caution. It is possible in fact to place error bars on the values of  $r_{XY}$ , but such refinements are considered outside the scope of this introductory text. The regression lines for Nitrogen and Phosphorus are

$$N = 17.06Q + 6121.06, \quad \text{and} \quad P = 0.258Q + 158.22.$$

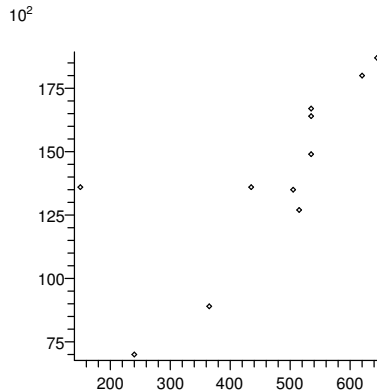


Fig. 1.7 Total nitrogen against river flow: the  $x$ -axis is  $Q m^3 s^{-1}$ , the  $y$ -axis is nitrogen  $N$  in tonnes per year

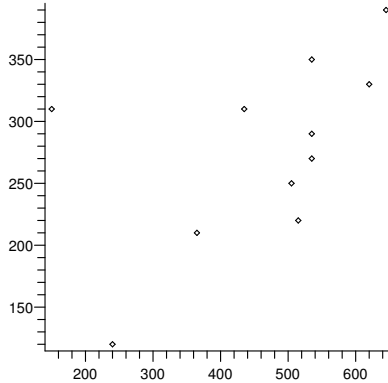


Fig. 1.8 Total phosphorus against river flow: the  $x$  axis is as before and the  $y$  axis is phosphorus,  $P$ , also in tonnes per year

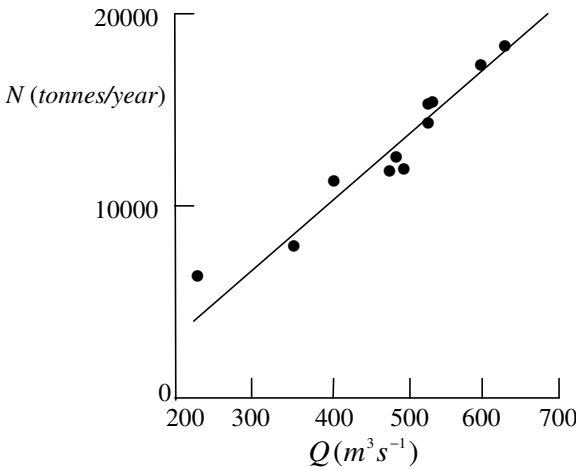


Fig. 1.9 Regression line for nitrogen

Note that no attention has been paid to the units here. The data are given in a mixture of units, as is quite typical (one might even say prevalent) in marine science with its long nautical traditions, and no conversions to, say, standard SI units have been made. This may annoy the purists, but in the calculation of lines of best fit, the geometric distance between the data points and the regression line has been minimised and this process is independent of units. Only if we wish for sensible units for the constants  $A_N$ ,  $A_P$ ,  $B_N$  and  $B_P$  does it become necessary to standardise. Finally in

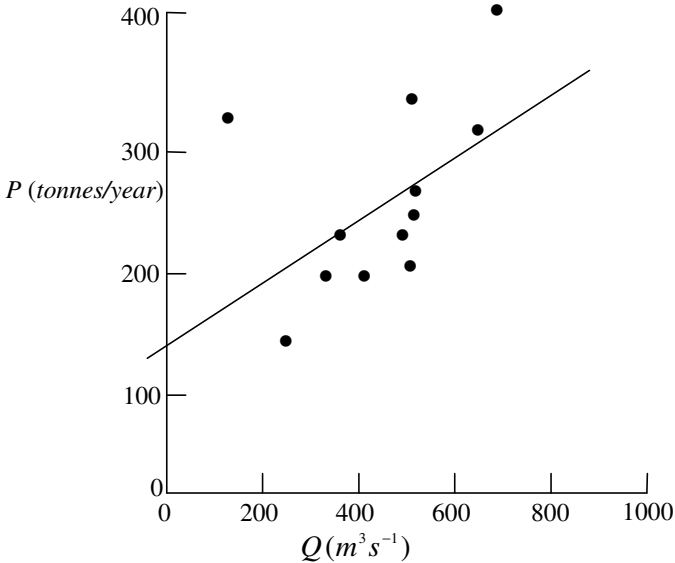


Fig. 1.10 Regression line for phosphorus

this example, let us do some predicting. We use the regression lines to predict the values of nitrogen and phosphorus in the River Göta when the river discharge is  $800 \text{ m}^3 \text{ s}^{-1}$ . Either from drawing these lines on the graphs (shown in Figures 1.9 and 1.10) or, more accurately, from inserting  $Q = 800$  into the formula for each line in turn we get

$$N = 19769 \quad \text{and} \quad P = 364.$$

The first albeit less accurate method is acceptable, particularly because it keeps you in touch with the data, reminding you how scattered the points are, and hence how low the correlation is. Most importantly, it indicates how much (how little?) reliance can be put on these predictions. A correlation of 0.9 would be considered a reasonable figure, and the data falls well below this.

### 1.4.3 Principal Component Analysis

There are several more advanced techniques that are now readily used by not just marine scientists and coastal engineers but all kinds of researchers. Many of them are not particularly new, but the advance in software has meant that they are much more available and easy to use than used to be the

case. Let's start with Factor Analysis. The development of Factor Analysis occurred well away from marine science and coastal engineering; indeed well away from science. It finds its greatest use in marketing, psychology and social science. In these subjects, there are a plethora of (ill defined) variables and to make sense of any large data sets is bewilderingly difficult. Factor analysis is a technique for reducing the number of variables to a salient few. In these less exact fields, the variables are usually assumed to be linear combinations of factors. When it comes to more precise sciences, we home in on a method called Principal Component Analysis (PCA for short) which is a very useful method for analysing the variability, spatial or temporal, of physical fields. Not only scalars such as temperature or salinity but vectors such as currents too. We have already discussed regression in the previous section, and in some ways PCA is a generalisation of regression. In other ways it is far more sophisticated. Briefly, PCA is a method of identifying patterns in data. The identification of the line of best fit or regression line is one. The line that is perpendicular to this, also a regression line of course, but hardly a line of best fit is another. The point about PCA is that the data can be  $n$  dimensional and it can pick out the main trend, second best, third best etc. For those that know about Fourier series or harmonic analysis (of tides) PCA is very like retaining the first two or three terms as a picture of the main ingredients. Neglecting all but the first few Principal Components does mean the loss of information, but in practise it is these first two or three that contain most of the information and with PCA it is in a far more digestible form. So having sold it, how do we do it?

Many of the building blocks are already here, and as with useful methods, to do the general theory will look far too daunting. Let us go through it in two dimensions. This is useful in two respects. First of all being two dimensional means data can be displayed on paper or screen, and secondly the relationships with regression are easier to see. The first thing that is needed is a reasonably full data set. Suppose that this is in the form of pairs  $(x_i, y_i)$  with  $i = 1, 2, \dots, N$  displayed on a scatter diagram. It is necessary for these to be standardised in some way before any analysis takes place in much the same way as data is "normalised" for use with the normal distribution. This is important and needs to be remembered when doing real data analysis as in Chapter 6 when an example involving coastal erosion is done. Here, suppose that this has been done and the scatter diagram is ready. Figure 1.11 shows the data. One can see that the data in some sense follows a linear trend, and two lines are drawn that are the first and second principal axes. For those that have a mechanics background these

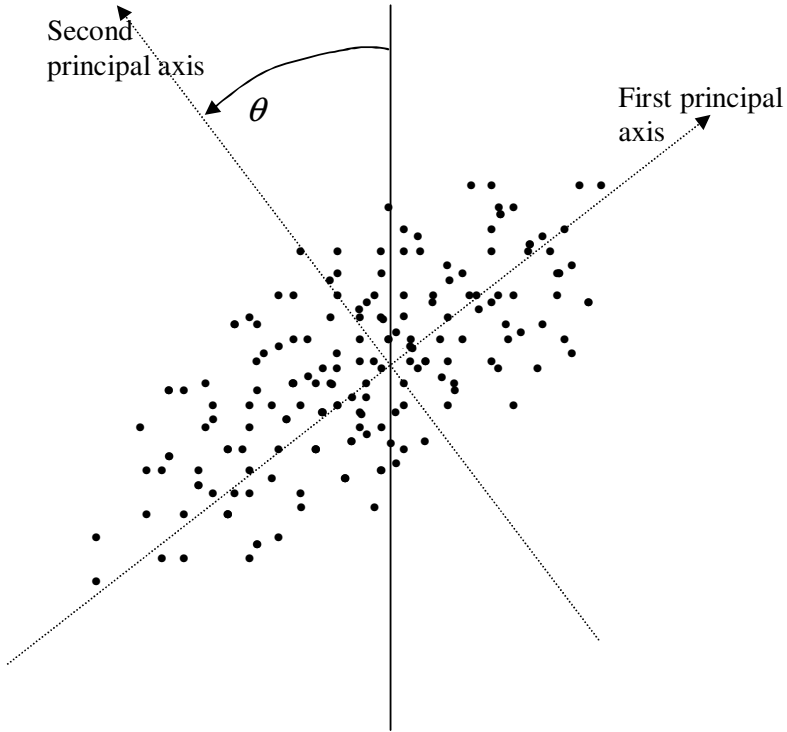


Fig. 1.11 A bivariate scatter plot showing the first and second principal axes and the angle  $\theta$  through which the graph needs to rotate

are analogous to the principal axes of a rigid body. In order to calculate them here recall the definition of covariance given earlier in this chapter: if the points are  $(x_i, y_i)$  with  $i = 1, 2, \dots, n$  then the covariance is

$$\sigma_{xy}^2 = \frac{\sum(x_i - \bar{x}) \sum(y_i - \bar{y})}{n}$$

however, it is best if the mean is subtracted from the data first, in effect translating the origin of the scatter plot to the middle of the points. The covariance is then:

$$\sigma_{xy}^2 = \frac{\sum(X_i) \sum(Y_i)}{n}$$

where  $X_i = x_i - \bar{x}$  and  $Y_i = y_i - \bar{y}$ . Of course  $\sigma_{yx}$  will be the same as  $\sigma_{xy}$  by symmetry. The variances of the data are

$$\sigma_x^2 = \frac{\sum(X_i) \sum(X_i)}{n}$$

and

$$\sigma_y^2 = \frac{\sum(Y_i) \sum(Y_i)}{n}.$$

In order to carry out PCA, the *covariance matrix* needs to be found. For the bivariate case it is the following  $2 \times 2$  matrix:

$$\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}.$$

The most natural way to proceed would be to rotate the axes such that one axis went through the “axis” of the data and resembled the regression line, then the other axis would be perpendicular to this. The way to do this is to define new axes  $\xi, \eta$  that are rotated through an angle  $\theta$  (shown in Figure 1.11). These new coordinates are related to  $(X, Y)$  through:

$$\xi = X \cos \theta + Y \sin \theta, \quad \text{and} \quad \eta = -X \sin \theta + Y \cos \theta,$$

where there should be no confusion using  $\eta$  here as its normal meaning as surface elevation does not feature until Chapter 3. The angle  $\theta$  is for the moment arbitrary. The variance of the data along the direction dictated by  $\theta$  is easily calculated as  $\sigma^2(\theta)$  where

$$\sigma^2 = \sum_{i=1}^N [X_i \cos \theta + Y_i \sin \theta]^2 = \sigma_{xx}^2 \cos^2 \theta + 2\sigma_{xy}^2 \sin \theta \cos \theta + \sigma_{yy}^2 \sin^2 \theta.$$

Using double angle formulae, this can be written:

$$\sigma^2(\theta) = \frac{1}{2}(\sigma_{xx}^2 + \sigma_{yy}^2) + \sigma_{xy}^2 \sin 2\theta + \frac{1}{2}(\sigma_{xx}^2 - \sigma_{yy}^2) \cos 2\theta.$$

The angle  $\theta$  should be chosen so that the line goes through the data in such a way that  $\sigma^2(\theta)$  is an extremum. This is found by choosing the variance to be an extremum (usually maximum or minimum) with respect to  $\theta$  so that using calculus (for a maximum or minimum the derivative is set to zero) we get:

$$\frac{d}{d\theta}(\sigma^2(\theta)) = (\sigma_{yy}^2 - \sigma_{xx}^2) \sin 2\theta + 2\sigma_{xy}^2 \cos 2\theta = 0$$

so that the values of  $\theta$  are given by

$$\tan 2\theta = \frac{2\sigma_{xy}^2}{\sigma_{xx}^2 - \sigma_{yy}^2}.$$

This formula gives two angles, 90 degrees apart. Using the second derivative as a check reveals that the two principal variances are given by

$$\frac{1}{2} \left[ (\sigma_{xx}^2 + \sigma_{yy}^2) \pm \sqrt{(\sigma_{xx}^2 - \sigma_{yy}^2)^2 + 4\sigma_{xy}^2} \right]$$

where the plus sign gives the maximum variance (the principal axis) and the minus sign the axis perpendicular to this. It is found that for these special values of  $\theta$  the covariance is zero, so the covariance matrix is diagonal. Anyone who knows about diagonalising matrices will perhaps have heard of eigenvalues and eigenvectors, of which more in a moment. The two directions in the plane of the data are sometimes called EOFs (Empirical Orthogonal Functions), “use of EOFs” is often preferred to PCA in oceanography but they are the same thing. EOFs are often used in association with the use of data assimilation and what amounts to a smoothing routine called Kalman filtering, but discussion of this is postponed until the next chapter.

In two dimensions this looks reasonably straightforward, but if the data had a larger number of dimensions (instead of  $(x_i, y_i)$  we would have to write  $(x_{1i}, x_{2i}, \dots, x_{Ki})$  where perhaps  $K = 20$ ) then the covariance matrix would be  $K \times K$  and using calculus and geometry as we did above is not an option. Instead one needs some matrix algebra. Some of you will know enough about eigenvalues and eigenvectors of a matrix to follow what comes next, but others will not. For those in the latter category, here is a quick diversion.

If  $A$  is a square matrix, then a number  $\lambda$  such that

$$|A - \lambda I| = 0$$

is called an *eigenvalue* of the matrix  $A$ . Here  $I$  is the unit matrix that has ones along the diagonal but zeros everywhere else. The vertical bars denote the determinate of the matrix which when multiplied out leads to an equation of the same order in  $\lambda$  as the size of the matrix  $A$ . That is a  $2 \times 2$  matrix leads to a quadratic in  $\lambda$ , a  $3 \times 3$  matrix a cubic and so on. Solving this equation (called the characteristic equation) for  $\lambda$  will therefore give as many different values for  $\lambda$  as there are rows or columns of the matrix  $A$ . This is not always so (some quadratics can have only one root,  $\lambda^2 - 2\lambda + 1 = 0$  for example) but we will gloss over these special cases. Here is an example; consider the matrix:

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}.$$

This has eigenvalues given by the roots of the equation

$$\begin{vmatrix} 2 - \lambda & 3 \\ 1 & 4 - \lambda \end{vmatrix} = 0.$$

This is expanded to:

$$(2 - \lambda)(4 - \lambda) - 3 = 0$$

or

$$\lambda^2 - 6\lambda + 5 = 0.$$

This has the two solutions:

$$\lambda = 1 \quad \text{and} \quad \lambda = 5$$

and these are the two eigenvalues of the matrix

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix}.$$

Having found the eigenvalues, there are eigenvectors associated with each of them. The theory says that if there is a scalar  $\lambda$  such that for the matrix  $A$  we have

$$|A - \lambda I| = 0$$

then associated with each of these  $\lambda$  there is at least one vector  $\underline{x}$  such that

$$A\underline{x} = \lambda\underline{x},$$

and these vectors  $\underline{x}$  are the eigenvectors of the matrix  $A$ . In order to find these for our  $2 \times 2$  matrix we solve the two sets of simultaneous equations:

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1 \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 5 \begin{pmatrix} x \\ y \end{pmatrix}.$$

These two sets of equations can be written:

$$\begin{pmatrix} 1 & 3 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0 \quad \text{and} \quad \begin{pmatrix} -3 & 3 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 0.$$

You may have noticed that neither of these are true simultaneous equations. The first is the one equation  $x + 3y = 0$  written twice and the second is the one equation  $-x + y = 0$  twice over. This is correct because the eigenvalues  $\lambda$  are precisely those values that render the determinant of the matrix  $(A - \lambda I)$  singular (having zero determinant), and a zero determinant means redundancy which means for two equations that they are the same. To solve  $x + 3y = 0$  we can choose  $x = 3$  and  $y = -1$ , and to solve  $-x + y = 0$  choose  $x = y = 1$ . Therefore we have eigenvalues 1 and 5 with associated eigenvectors  $\begin{pmatrix} 3 \\ -1 \end{pmatrix}$  and  $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$  respectively. These vectors can be multiplied by any number (scalar) and they would still remain eigenvectors, that is because it is the *direction* that is important not the length of the vector. In Principal Component Analysis, it turns out that it is the *normalised* eigenvectors that are required, that is the eigenvectors of unit length. The length of a vector is the square root of the sum of the squares

of the components, so for the example here, the normalised eigenvectors are, uniquely:

$$\begin{pmatrix} 3/\sqrt{10} \\ -1/\sqrt{10} \end{pmatrix} \text{ and } \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

The method generalises to any square matrix, though of course the arithmetic gets more involved as do the number of special cases to consider. Never mind about the technical stuff, what does this all mean. We will come to the context of Principal Component Analysis in a bit, but if the matrix  $A$  represented a geometrical description of a quadric surface (an ellipsoid perhaps; the shape of the earth is an ellipsoid called an oblate spheroid) then the eigenvectors would represent the main axis (there is only one) and the axes of symmetry perpendicular to this. If the cross-section of the ellipsoid is circular then there would be an infinite choice of the two perpendicular axes. It is precisely this case that arises when eigenvalues are a multiple root of the characteristic equation. Turning to mechanics, if  $A$  represented the matrix of moments of inertia, then the eigenvectors would represent what are the principal moments of inertia if you know about such things. In Principal Component Analysis, the eigenvectors represent the direction through the data where there is greatest variation. A regression line will be one such, but for multivariate data these will be more than one. In Figure 1.11, one eigenvector or principal axis is obvious and drawn, a second is less so but is also displayed. If there were 20 variables, then the matrix of covariances would be  $20 \times 20$ ; there would be 20 eigenvalues and, associated with each of these an eigenvector. There will always be enough eigenvectors, but sometimes the number of eigenvalues will be less; the coincident roots case of the characteristic equation gives rise to a whole plane of eigenvectors from which one can choose any two at right angles, this is the case alluded to above. The role of finding eigenvalues and eigenvectors is to pick out directions through data where there is most variation. Once the direction of maximum variation is found, we then seek the second largest, then the third largest etc. We have not said anything about how to order these yet, but it turns out that the first two or three eigenvectors or principal components carry most of the variation and so analysing the data is much simplified. A 20 dimensional problem has been reduced to 2 or 3 dimensions. Also it is the largest eigenvalue that is associated with the greatest variation, so although there will be as many eigenvectors as dimensions, as said above only a few will be required to capture most of the variation. In Chapter 6 there will be an example, but this is as far as we will go into the theory of

using statistics for modeling. The next step would be to look at non-linear regression and to include placing confidence intervals on predictions. Those interested in these topics need to consult more specialist statistics texts.

## 1.5 Exercises

- (1) For the following phenomena, list the effects that need to be included in a *first* model:
  - (a) Gulf stream,
  - (b) A pollution accident in an estuary,
  - (c) A surface oil slick 5km. offshore of Aberdeen, Scotland.
- (2) Explain how you think modeling can help in the clean up of eastern European rivers.
- (3) Explain why calculus is important to know for the modeler of coastal and offshore processes.
- (4) Say which of the following variables are scalars and which are vectors: pressure; wind; salinity; temperature; force; current; stress.
- (5) Use Green's theorem in the plane to show that if the current is in the form  $\mathbf{u} = \nabla\phi$ , then the circulation around any closed curve in the fluid must be zero.
- (6) The following table gives two data sets for measurements of inputs into the North Sea; units are  $10^3 \text{ km}^3$ :

Location	A	B	C	D	E	F
Expt. 1	0	9.5	34	3.4	0.4	0.4
Expt. 2	10	11	40	5.2	0.2	0.3

$A$  is the Pentland Firth,  $B$  between Orkney and Shetland,  $C$  between Shetland and Norway,  $D$  through the straits of Dover,  $E$  river discharges and  $F$  represents input from precipitation. Use the  $\chi^2$  test to decide whether or not the two sets of data are in agreement, and comment accordingly.

- (7) Suppose a die is thrown six times and the results are: 3, 6, 6, 1, 6, 2. The probability of throwing a six should be  $1/6$  but these results indicate a bias. Use Maximum Likelihood Estimation to estimate the maximum value of the probability of throwing a six.
- (8) It is suspected that the relationship between the probability of an extreme wave occurring  $p$  and time  $t$  is of the form:

$$p = \exp\{-\exp\{(t - a)/k\}\}$$

where  $a$  and  $k$  are constants. Find a function of  $p$  that renders this relationship linear.