Copy B

Handbook No. 26

FISHERIES STATISTICS TRAINING COURSE Lecture Notes

Prepared by

P.W. Hodgkinson and M.J. Williams

South Pacific Commission
Noumea, New Caledonia
September 1985

13/86

Original text: English

Prepared for publication at
South Pacific Commission headquarters, Noumea, New Caledonia
and printed at Commission headquaters, 1985.

PREFACE

It has become increasingly clear that the collection and analysis of fisheries data must be an integral part of all fisheries development programmes. In response to this requirement, the Asian Development Bank and the South Pacific Commission organized a training course on fisheries statistics held at SPC headquarters in Noumea, New Caledonia, from 3 to 14 September 1984. The Asian Development Bank provided the major part of the funding for the course and financed the costs of the two instructors, travel and per diem costs of eleven trainees, and the publication of this manual. Travel and per diem costs of four participants were provided by the Fiji regional office of FAO/UNDP (South Pacific Regional Fisheries Development Programme), together with the donation of scientific pocket calculators for course participants. The Government of France funded the participation of three participants. The preparation and conduct of the course was assisted by Dr Paul J. Hooker, whose services as a consultant were kindly provided by the Indo-Pacific Tuna Development and Management Programme of FAO. Support facilities, secretarial services, computer resources and participation by staff of the Tuna and Billfish Assessment Programme as part-time intructors were provided by the SPC.

The course was attended by 21 trainees from the fisheries divisions of 16 different countries of the SPC. These people are all involved to some degree in handling fisheries statistics in their home countries. The course was uniformly judged by the participants to have been useful and of assistance in the conduct of their normal work load, and some sort of follow-up activity was requested. The publication of the lecture notes is one of the follow-up activities that is to be conducted.

As part of their activities, the two instructors, Mr Peter Hodgkinson, former SPC Statistician and Dr Meryl Williams, former Fisheries Statistician with the Tuna and Billfish Assessment Programme, prepared a set of lecture notes for the course. These notes form the basis for an elementary foundation in general statistics and are unique in their extensive use of fisheries data in the examples. It is partly due to this heavy use of fisheries examples which contributed to the success of the training course and which makes the lecture notes a valuable reference for workers in fisheries offices of the region. In preparing the notes, the authors have freely complemented their own thoughts and ideas with extracts from two previous SPC statistical publications authored by Mr. G.J. Eele (Eele 1982a, 1982b)[1], and a report published by the FAO/UNDP South China Sea Fisheries Development and Co-ordinating Programme on a similar training course on fishery statistics held in Thailand in 1981[2].

---

1 Eele, G.J. (1982a). Statistical operations and procedures (elementary level) lecture notes. South Pacific Commission, Noumea, New Caledonia.

Eele, G.J. (1982b). Statistical operations and procedures (intermediate level) lecture notes. South Pacific Commission, Noumea, New Caledonia.

2 Hooker, P.J. (1982). Report on the regional training course on fishery statistics, 1 September - 9 October 1981, Samutprakarn, Thailand. SCS/82/GEN/41a, Part II, Technical Report, Volume 2. South China Sea Fisheries Development and Coordinating Programme, Manila, Philippines.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

TOPIC 1 - INTRODUCTION : WHAT ARE FISHERIES STATISTICS

In this course we will be discussing some of the basic techniques used in collecting, presenting and analysing fisheries statistical data. But in this introductory topic, we will talk briefly about why we collect fisheries statistics, what types of statistics are collected by and for a fisheries statistical system, the types and levels of detail needed for some of these data and some possible types of statistics collections. Following this short introduction to the world of fisheries statistics, we begin the part of the course which is concerned with detailed studies of basic statistical methods which can be applied to fisheries data.

1.1 Components of a national fisheries statistical system

Fisheries statistics are essential to fisheries managers and scientists alike. A government fisheries department collects a wide range of types of fisheries data, from scientific to socio-economic data. We briefly examine the whole range of types of statistics which come under the control of a government fisheries department. Not every department, of course, will deal with all types. In each category of statistics, we mention whether we would normally expect to collect a census of the statistics or just a sample.

Domestic large-scale fishing

Census statistics on large-scale fishing are usually gathered by detailed catch reports and by landing data returns. Examples are pole-and-line, purse-seine and longline operations for tuna and trawling for prawns and fish. In the case of logsheet data, we may have to process the data by computer since the amount of detail is large.

Foreign large-scale fishing

Census statistics are collected under government-to-government access agreements. Detailed processing of the logsheets is carried out on computer at the South Pacific Commission (SPC), but some preliminary statistics should be recorded in-country.

Domestic small-scale commercial fishing

Sample or census logbook statistics are collected from fishing units and/or from landings. Examples are local deep-bottom fishing, trolling for pelagic species, and shell collecting for export.

Local artisanal and subsistence fishing

Sample surveys are conducted, usually by fisheries officers, to estimate catch and effort. Careful planning, recording and analysis of the samples is required. The estimates of non-commercial catch are often quite imprecise and some of the difficulties involved in making catch estimates are discussed in this course.

Market statistics

Economic data on market sales, costs, profits and number of operators for the commercial fish catch should be collected where possible.

Export statistics

Values and quantities of fish and fish products exported should be monitored.

Import statistics

Quantities of fresh, frozen and processed fish products imported should be known, especially with a view to supplanting some imports with local products.

Fish consumption statistics

The quantities and types of fish and fishery products consumed in a country are of considerable interest in planning for development of fisheries and fishery facilities, e.g. processing plants, and in defining the role of fish products in economic and nutritional terms. Fish consumption statistics may be obtained from household surveys, catch, import and export statistics.

Research and survey results

A government fisheries statistician should be familiar with the results of research and survey data on fisheries in the country and, where possible, should be given copies of the results to supplement routine data collections. Detailed research and survey programmes can help explain some basic observations on changes in the status of stocks which may have been noted in routine catch and effort statistics.

Aquaculture

Pond sizes, equipment, capital outlay, employment and production data are basic to monitoring the progress of aquaculture projects. A census of aquaculture data should be attempted.

Fishing vessels

Fisheries departments frequently are responsible for licensing commercial fishing vessels and keeping estimates of the numbers of non-commercial vessels involved in fishing. Even if the fisheries department is not responsible for licensing of vessels, the fisheries statistician should be familiar with current information on fishing vessels. In some cases, the sizes, numbers and types of boats used for fishing may have to be determined by household and village sample surveys.

Data on fishermen

Often the number of commercial fishermen is easy to determine because such fishermen usually have to be licensed. Alternately, we may get estimates of numbers from market sales. Determining the numbers of non-commercial fishermen, however, is more difficult, but fisheries departments frequently are responsible for sample estimates of numbers of people involved in fishing. In conjunction with other government departments, fisheries departments often carry out socio-economic studies of people involved in fishing so that the value of fishing in economic, employment and nutritional terms may be understood. Socio-economic studies usually are carried out by household surveys, e.g. the 1981 survey of artisanal fishing in South Tarawa (Anon 1982), the 1981/82 study of the status of women in fisheries activities in Fiji (Lal and Slatter 1982).

1.2 Why collect fisheries statistics

Before examining the problems of which statistics to collect and how to go about collecting them, we might ask the question - why do we collect fisheries statistics?

The uses of fisheries statistics may be classified as for (a) assessment, (b) monitoring, or (c) planning and management. Within each category we have scientific, economic and socio-economic fisheries statistics. In note form, we may summarise the uses of various types of fisheries statistics in the following way. Many of the ways in which fisheries statistics may be used will be seen in greater detail in following sections of the course.

1.2.1 Assessment

Scientific:

Data types: catch and effort, length frequency, species composition and catch rates, biological samples

Uses:
- assessment of current status of stocks
- potentials of new fishable stocks
- determination of biological parameters of populations, e.g. recruitment, reproduction, age structure

Economic:

Data types: landing data, market, import, export, consumption, vessel, gear, employment, foreign and local catch and effort statistics

Uses:
- to calculate values of catch, catch per unit effort, vessels and gear, import, export and domestic revenue
- to determine employment and occupational involvement in fisheries
- to calculate revenue and fees for foreign fisheries operations

Socio-economic:

Data types: national census, sample and household surveys

Uses:
- to determine employment and occupational importance of fishing to members of the population
- to assess the nutritional importance of fish and fishery products

1.2.2 Monitoring

Scientific:

Data types: time series of data types given in section 1.2.1 (Scientific: Data types)

Uses:
- changes in status of stocks
- effects on catch per unit effort due to interaction of different fisheries
- changes in biological parameters
- changes in fishing techniques and consequent effects on catches

Economic:

Data types: time series of data types given in section 1.2.1 (Economic: Data types)

Uses:
- to monitor changes in values of catch, catch per unit effort, vessel values, etc.
- to monitor employment and occupational involvement in fisheries
- to monitor returns in revenue from foreign fishing interests

Socio-economic:

Data types: time series or periodic updates of data types given in section 1.2.1 (Socio-economic: Data types)

Uses:
- to monitor changes in occupational importance of fishing
- to monitor changes in nutritional importance of fishery products

1.2.3 Planning and management

The end result of conclusions drawn from assessment and monitoring information is the input to planning and management of fisheries and fishing related enterprises.

Scientific:

Uses:
- regulation of fishing operations for conservation of stocks
- development of new fisheries and techniques

Economic:

Uses:
- planning of capital expenditure on vessels, ports, landing facilities, processing plants, etc.
- regulation of fishing operations for economic reasons
- calculation of access fee levels for foreign fishing

Socio-economic:

Uses:
- development of employment and occupational fishing projects
- regulation of fisheries based on social considerations

1.3 What fisheries data should be collected

In this section, we will discuss the form of data related specifically to fishing operations. Other forms of fisheries statistics listed in section 1.1, e.g. market, import, export, socio-economic and aquaculture statistics, are similar in their form and method of collection to many other types of non-fisheries related statistics, e.g. agricultural, income, educational. The collection and analysis of fisheries statistics not related to fishing operations are dealt with more generally in Topic 6 on sampling. The form and collection of data related to fishing operations, however, present features peculiar to fishing and to no other activity and an introduction to some of these features is now given.

In all fisheries a basic set of fisheries catch statistics is needed to satisfy the routine requirements of government fisheries departments in regard to research, stock assessment and economic management. The 1981 ICLARM/CSIRO Workshop on the Theory and Management of Tropical Multispecies Stocks (Pauly and Murphy 1982) identified the following requirements:

- Reliable catch by species and associated effort data.

- Length composition by species or, if appropriate, by groups of species. Where discarding of part of the catch at sea is known to occur, it will be necessary to sample discards for length composition as well as by species to enable conversion of length composition of landings to length composition of catch.

- Indices of abundance calculated from records of catch and effort and expressed in units of catch per unit of standardised fishing effort. Research vessels or selected commercial vessels may be used for this purpose.

- Related to these data requirements is the problem of obtaining satisfactory species identification. With the large number of species, special efforts are needed to provide field workers with easily used taxonomic aids.

- Age composition of selected species as a basis for using standard techniques of assessment and for calibrating length-structured models.

Except perhaps in the cases of large-scale foreign and domestic fisheries, even basic data requirements for many fisheries may be difficult to meet in the Pacific since:

- A large number of different species are caught.

- A wide variety of fishing techniques are used, often including several different techniques for the same species.

- The subsistence and artisanal catching sectors are extremely important and in some cases commercial fishing is almost negligible.

- Fishing is usually done by a large number of small fishing units.

- The requirements for trained manpower to collect statistics from remote atolls, islands and villages are often prohibitive.

- Small-scale fishing methods may vary considerably with time of day, phase of moon, season, food and money needs of the people involved, etc.

1.3.1 Catch data

Catch data, by species of fish, may be collected directly from the fishing units, e.g. individual vessels or fishermen, and/or in aggregated form as landings of fish.

Data collected directly from fishing units, e.g. on catch reports or logsheets, will be much more detailed than aggregated landing data. However, the accuracy of data reported directly from fishermen needs to be checked. Catch report forms must be well designed and accompanied by clear instructions on how to fill them in. In addition, fisheries officers must rely on fishermen being able to identify fish species accurately and to record weights, numbers, positions of catch, etc. carefully.

Fish landings are useful for estimating total catch and for cross-checking detailed logsheet data. In addition, species composition of the catch may be more accurately determined from landing data. For example, small yellowfin and bigeye tuna caught by purse-seiners are very difficult to tell apart and catch reports filled out on board fishing vessels often lump the two species as "yellowfin". On landing, however, greater care is taken to distinguish the species because they are sold for different prices.

The problem of which species to record as separate species and which to record as groups of species or as "other species" must be considered carefully. Species which are to be recorded separately must be able to be identified accurately and must make up a measurable fraction of the catch. In tropical fisheries, the diversity of species is often large and no small number of species dominates the catch. For example, Munro (1982) reported that in a trap fishery in Jamaica, 35 species each comprised at least 1.5 per cent of the catch by weight.

One approach to the species problem is to combine all species of one biological group, e.g. deep water snappers (*Etelis* spp.), parrotfish (*Scarus* spp.), groupers (*Epinephalus* spp.), when collecting routine catch data and then to estimate the species composition of the group from a subsample of the catch only. Alternately, we may collect catch data for one or a few indicator species in each group and total catch for the group as a whole. In general, catch totals for the group only will not be sufficient since any changes in species composition within the group will be missed. An example of a typical breakdown of species and species groups is shown in Table 1.1. The species are all caught in a deep-bottom dropline fishery and a trolling fishery operating in the same general area. We see that the group of miscellaneous and minor species constitutes only three per cent of the catch. Three species are recorded as individual species and all others are put into family or sub-family groups.

Special care must be taken to record the occurrence of species which may be only a minor component of one fishery but which are major or target species of other fisheries, e.g. yellowfin tuna in local troll fisheries and in large-scale tuna fisheries. Catch records of such species are very

important if we are trying to see whether the catches in the big fisheries are having any impact on catches in the smaller fisheries, i.e. if any interaction is occurring between the fisheries.

TABLE 1.1 : AN EXAMPLE OF FISH CATCH COMPOSITION RECORDS - DEEP BOTTOM DROPLINE AND SURFACE TROLL FISHERIES

Species/Group	Approximate Percentage of catch by weight
Etelis carbunculus (Deep-water red snapper)	12
Pristipomoides auricilla (Gold tailed jobfish)	4
Pristipomoides zonatus (Banded flower snapper)	14
F. Lutjanidae, s-F. Etelinae (Other deep-water snappers)	10
F. Lutjanidae, s-F. Lutjaninae : F. Lethrinidae (Shallow water snappers and emperors)	8
F. Serranidae (Groupers)	12
F. Carangidae : F. Scombridae : F. Thunnidae (Coastal pelagic species)	15
Ruvettus pretiosus (Oilfish)	7
F. Sphyraenidae : F. Scorpaenidae : F. Labridae : Unident. (Miscellaneous bony fish)	3
F. Carcharhinidae : F. Hexanchidae (Sharks)	15

Length-frequency distributions of the major species or of indicator species are of considerable value in monitoring the state of fish stocks. A sampling programme carried out at markets or landing sites or by research surveys is recommended. Provided sufficient samples are taken regularly, the size composition of stocks can be monitored for recruitment, change in sizes of fish taken and, in some cases, growth. Information on where the samples of fish were taken and what fishing gear was used must also be carefully recorded.

1.3.2 Time and area details

The amount of detail required in specifying time and area of catch is important when designing a data collection system for fisheries statistics.

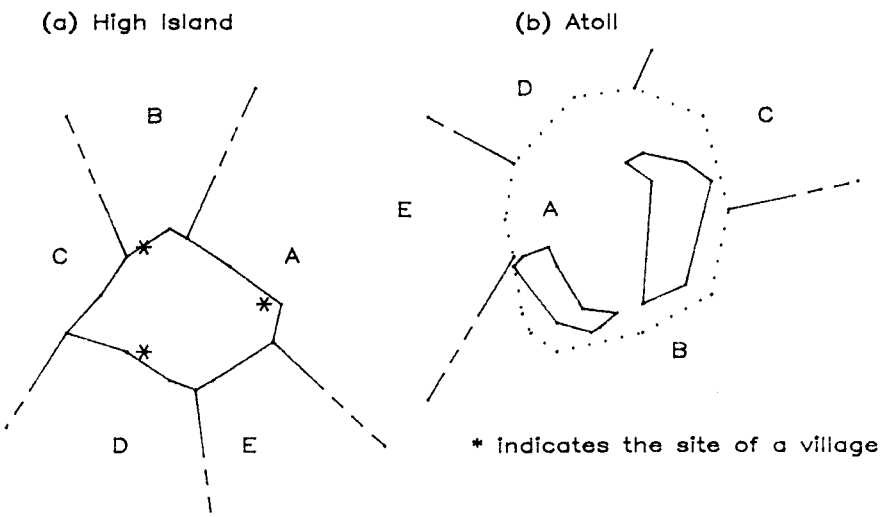
Time resolution is often straightforward because catch or landings are usually recorded by day or 1-2 day trip date. Greater precision on time of catch, e.g. hour of day or night, is more difficult to collect accurately.

With respect to area of catch, the amount of precision we need or can obtain is determined both by the type of fishery and by the precision with which the fishing boats are able to report catch. Thus, in inshore reef fisheries we may be interested in knowing the area fished to within a few kilometres accuracy whereas in pelagic fisheries we may be satisfied with several tens of kilometres accuracy.

In large-scale fisheries where vessels have good navigation instruments, fishing positions can readily be given in precise degrees and minutes. In such cases, computers can be used to produce graphic presentations of large amounts of detailed catch and effort data and to summarise the data in numeric form.

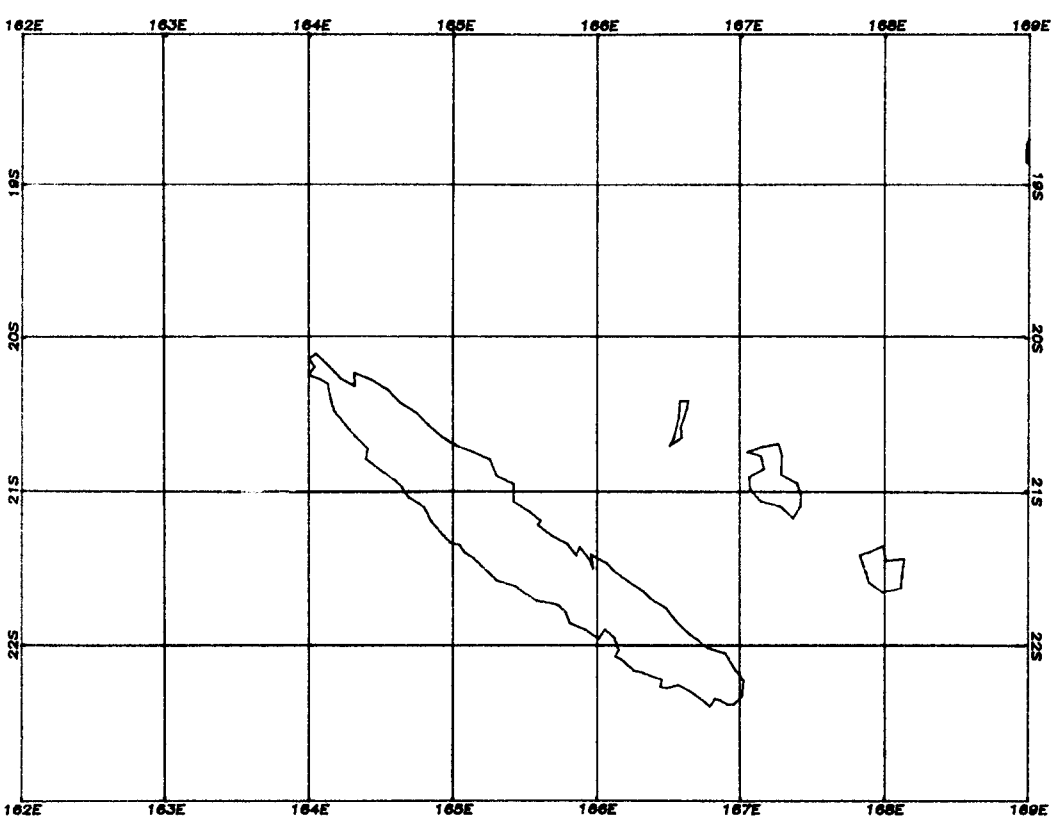
In practice, in small-scale fisheries, exact fishing positions usually cannot be obtained. However, a local system for describing approxiate areas should be set up. For example, the waters surrounding a high island may be divided into discrete areas which represent different possible fishing areas (Figure 1.1(a)); an atoll with lagoon may be divided into lagoon plus several offshore areas (Figure 1.1(b)). Particular note should be taken of recording fishing around fish aggregation devices (FADs).

FIGURE 1.1 : EXAMPLES OF ZONE AREAS AROUND ISLANDS



If some navigation equipment is available to local vessels, a grid system may be set up, e.g. a 1/4x1/4 degree, or 1x1 degree grid system. Figure 1.2 shows a 1x1 degree grid system.

FIGURE 1.2 : EXAMPLE OF 1x1 DEGREE GRID FOR NEW CALEDONIA



In general, the smaller the areas defined, the better the catch and effort data may be monitored, but the greater the difficulty in collecting and processing the data.

1.3.3 Fishing units

The fishing unit is defined as the smallest discrete, complete unit necessary for a fishing activity. The fishing unit varies from fishery to fishery but always consists of the fishing gear, persons (crew or fishermen and fisherwomen) and often fishing vessel or vessels. In some fisheries, the unit is obvious, e.g. in tuna longline fishing the unit is the longline vessel, crew and fishing equipment. In smaller fisheries the unit is not so clear, e.g. in shellfish gathering the unit may be one or more people and their collecting equipment. A table of common Pacific fishing methods and the fishing unit for each method is given in Table 1.2.

Identification of fishing units is important when designing the collection of fisheries statistics and also in choosing the way in which to measure fishing effort.

TABLE 1.2 : FISHING UNITS FOR SOME COMMON PACIFIC FISHERIES

Type of Fishing	Fishing Unit
Tuna longline	Longline vessel, crew and gear
Tuna pole-and-line	Pole-and-line vessel, crew and gear
Tuna single purse-seine	Purse-seine vessel, crew, gear and helicopter for locating fish
Tuna group purse-seine	Net vessel, skiff, carrier vessel(s), crew and gear
Trolling	Canoe/motor vessel, crew and gear
Deep-bottom fishing	Canoe/motor vessel, crew and gear
Trap fisheries	Canoe/motor vessel, crew and traps
or	Crew and traps if shore-based
Spearfishing	Fishermen, spears and boat(s)
Gillnets, set nets, lift nets, beach seines	Fishermen, nets and boat(s)
Cast netting	Single fisherman and net
Shellfish collecting	Collectors, equipment and boat(s)

1.3.4 Fishing effort

Catch data alone tells us little of the state of a fishery. For example, if the total catch of all reef fish is 50 tonnes in one month but only 10 tonnes in the next month, we have no way of knowing whether the drop in catch was due to reduction in available fish or to a drop in the amount of fishing, for whatever reason, carried out in the second month.

If we are to monitor the changes occurring in fisheries, therefore, we must measure or estimate not only the catch but the amount of fishing carried out. We quantify the amount of fishing by choosing a measure called a unit of fishing effort, depending on the fishing gear, skill and time required to catch fish. We require, therefore, some method for measuring the amount of fishing effort used by each fishing unit to catch fish. Ideally, the measure chosen should be such that catch is proportional to effort expended under given conditions. For example, if two people fish with lines in the same area on the same day, and one uses one line and the other uses two lines, we expect the person with two lines to catch about twice as much fish as the other since he/she is using twice the fishing effort.

Fishing effort is measured in different ways for different types of fisheries. Table 1.3 gives the usual measures of effort for different fisheries. The recommended measures for each fishing method are marked with a (1). Often it is difficult to collect data for the best measure of effort in a fishery and instead a less ideal but more easily measured unit must be used. For example, in set net fisheries on a reef, we may easily find out the total number of sets but may less easily obtain data on the size of nets and on the actual catching time for each set.

In practice, a unit of fishing effort can vary in effectiveness from fishing unit to unit and over time and area for the same unit. For example, 2000 longline hooks set for a standard period of time will vary in their success from time to time and from vessel to vessel. Several factors can cause variation in the effectiveness of a unit of fishing effort.

(i) Learning and technological changes:

In a new fishery, fishing skills and knowledge change rapidly over the first few years so that catches may improve with little apparent change in effort. The effective fishing effort is constantly changing as the fishermen become more skilful, but the measures of fishing effort, e.g. number of hours fished, number of sets made, will not show the changes.

The effectiveness of a unit of fishing effort may also change when new fishing gear or fish-finding equipment is introduced, e.g. a new type of net, a motor added to a canoe, better navigation and depth-sounding equipment, or when changes occur in the method of fishing, e.g. fishing at different depths from the usual, using a different type of bait, setting a purse-seine on different types of tuna schools.

(ii) Competition between units of gear:

Physical competition exists when the setting of additional units of gear directly interferes with the gears already fishing, e.g. heavy fishing of a school of fish may disperse the school. If many boats fish in an area, each boat may catch less per unit effort than if only a few boats fished.

(iii) Saturation of gears:

Some types of fishing gear cease to fish effectively once a certain amount of fish have been caught, e.g. set longlines, fish and crab traps.

TABLE 1.3 : FISHING EFFORT MEASURES

Type of Fishing	Gear Size/ Number	Bait	Vessel Size	Actual Catching Time	Searching Time	Changing Target Species	Skill	Within-fleet Communication
Deep sea handlines	** (1)	*		*** (1) (or no. of trips)		*		
Deep sea longlines	*** (1)	*		*** (1) (no. of sets)		*	*	
Reef, shore, gillnets, set nets	***			*** (1) (no. of sets)		*	*	
Traps, pots	*** (1)	*		*** (1) (no. of sets)		*		
Coastal pelagic - troll	*** (1)	*		*** (1)	*		*	
Oceanic - tuna purse-seine	**		*	*** (1) (no. of sets)	**	*	**	**
Oceanic - tuna pole-and-line	**	*	*	*** (1) (no. of days fished)	*	*	**	**
Oceanic - tuna and billfish longline	*** (1)	*	*	** (1) (no. of days fished)	*			*

Note: The preferred measure of fishing effort is marked by (1).

The degree of relative importance for each factor is shown by the number of asterisks (*). Three asterisks indicates that the factor is very important, two asterisks that the factor is moderately important and one asterisk that the factor is of low importance.

(iv) Co-operation between fishing units:

A single fishing unit may be much more successful when fishing in co-operation with other units or when receiving information from other fishermen. In some cases, we may have to redefine our fishing unit to account for co-operation, e.g. group purse-seiners.

(v) Expansion of fishing areas:

As a fishery expands to use new fishing grounds, the effectiveness of a fishing unit may be increased as previously untouched stock are fished.

(vi) Differences in skill between fishing units:

Fishing skill is hard to measure, but differences between fishing units cause some of the greatest differences in effectiveness of units of fishing effort. Adjustments may be made by comparing the long-term catches of different fishing units to the catches of a standard research or survey unit or to a particular unit of the local fishery. For example, if a particular fishing team or group of teams habitually catch one and a half times as much fish as other teams, each unit of effort from the successful teams is effectively equal to one and a half times similar units of effort from the other teams. Such adjustments to units of fishing effort are difficult to carry out, however, and a large amount of detailed data analysis is required.

Despite the care which must be taken in collecting fishing effort data and in using these data, fishing effort is still a very useful measure to have, both from the biological and socio-economic point of view.

1.3.5 Using catch and effort data

Catch and effort data are used by scientists, economists and planners as simple indicators of what is caught and how much effort is expended in fishing, and in calculations for stock assessment.

At the most basic level, if we know any two of the three statistics, catch, effort and catch per unit of effort, we may estimate the third statistic. If we know total catch and have a sample of catch with effort data, we may estimate total effort, provided the sample is representative of the whole. Conversely, if we know total effort and have a sample of catch with effort data, we may estimate the total catch. The values for total catch and effort derived from samples will only be estimates of the actual (unknown) values. If neither total catch nor total effort is known, however, both may still be estimated by sampling, but the estimates will be approximations only.

For the purposes of stock assessment, catch per unit effort (CPUE) is commonly used as an index of abundance of fish stocks. CPUE is calculated by dividing catch by effort, perhaps after first standardising effort. Unfortunately, CPUE will not always be proportional to fish abundance. Some of the factors that influence the relationship between catch and effort are:

(i) Multispecies fisheries:

When more than one species is caught in a fishery, CPUE of any one species may not be a reliable measure of the abundance of that species. Where certain species are highly sought after and fishing practices target on the preferred species, CPUE may be a misleading indicator of abundance of all species.

(ii) Standardisation of effort:

Individual units of effort may need to be standardised. If the measure of effort is not a true measure of effective fishing effort due to varying skill among fishermen, changes in technology, etc., the measures of effort may have to be adjusted to make CPUE values comparable between times and areas.

(iii) Discards:

Undesirable species and undesirable specimens of the preferred species, e.g. small, very large or damaged fish, may be discarded after catching. Discards are rarely recorded in routine fisheries statistics. The regional longline and purse-seine tuna catch reports have columns for discards of tuna and other species but not all vessels fill in the columns. Observers aboard vessels have provided useful information on the extent of discards but such information is only a very small sample of the whole catch. The Deep Sea Fisheries Development Project at the SPC keeps careful records of fish discarded, especially where the species of fish discarded, e.g. oilfish (Ruvettus pretiosus) and some sharks, have potential as food species. In general, research and survey programmes are required to provide reliable information on discards.

(iv) Changes in catchability of fish:

Changes in catchability of fish due to behavioural and physiological factors, e.g. schooling, reproduction, moulting in crustaceans, or to environmental factors, e.g. temperature, winds, moon phase, tides, may cause large fluctuations in CPUE. Such fluctuations, however, are not indicative of changes in abundance of the fish stock.

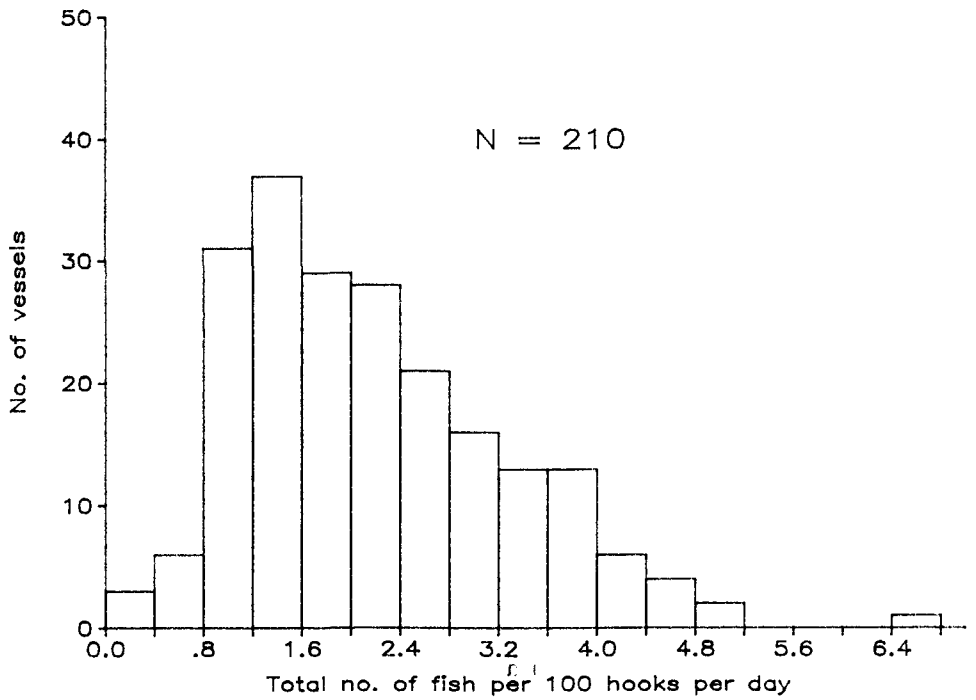
Take one further general observation at this point, and this is that the frequency distribution of CPUE in many fisheries is skewed to the right. A few fishing units have high CPUEs, but the majority have relatively low CPUEs. A typical example is the set of 210 longline boats which fished in one country over a three-month period (Figure 1.3). The CPUE (total number of fish per day fished) for the boats is skewed to the right. We will look at the presentation of such data as these CPUE data in more detail in Topic 3.

1.3.6 Length frequency

The uses of length frequency data are beyond the scope of the present course, but it is important to know that good length frequency data are being increasingly used for stock assessment purposes in tropical

fisheries. Workshop papers in the book on tropical multispecies fisheries edited by Pauly and Murphy (1982) may be referred to for further details.

FIGURE 1.3 : FREQUENCY DISTRIBUTION OF CPUE FOR LONGLINE VESSELS



1.4 Collecting fisheries catch and effort statistics

Collecting accurate and reliable fisheries catch and effort data is often difficult due to the problems mentioned in section 1.3. We just briefly mention here the nature of data collection in large- and small-scale fisheries. The collection of other fisheries statistics, such as market statistics and socio-economic data, is treated in Topic 6.

1.4.1 Large-scale domestic and foreign fisheries data

A census or total collection of large-scale domestic and foreign fisheries data should be attempted. A census rather than a sample is possible since large-scale fishing is conducted by large, well-organised fishing units, capable of catching large quantities of fish and of keeping accurate records of such catches.

Large-scale domestic operations are commercial concerns which are usually licensed under government regulations and should be required to provide complete catch and effort returns. Good examples are seen in the region, e.g. the Papua New Guinea prawn fisheries, the Fiji pole-and-line tuna fishery and the Tonga longline tuna fishery.

With respect to foreign fishing operations, the foreign affairs department of each country is responsible for agreements ensuring that fishing vessels provide catch and effort data. Fisheries access agreements usually include regulations on the type of catch reports to be used and how these are to be returned to the country fished. In practice, some vessels may fail to comply with the regulations and we must rely on surveillance and law enforcement authorities to provide us with information on the extent of the problem.

1.4.2 Small-scale fisheries

By their very nature, small-scale fisheries present great difficulties with respect to data collection. In many Pacific countries, the subsistence and artisanal fisheries catches are of greater magnitude than commercial catches. A census of fish catch and effort in non-commercial fisheries requires a prohibitive amount of work. Estimates of total catch and effort, however, may be made with a regular sampling programme and/or by short-term, intensive surveys.

The methods for designing sampling programmes are discussed in Topic 6. Such methods apply equally well to catch and effort estimates as to socio-economic statistics estimates, except that great care must be taken in considering all the possible sources of bias and variability affecting the sampling scheme. For example, how may we best sample sporadic or irregular types of artisanal fishing activity, or night fishing, or fisheries directed to spawning runs on certain lunar periods?

Apart from sampling, another useful method of collecting highly specific data on fishing catch and effort is by the use of fisheries surveys. Fisheries surveys are systematic fishing activities designed to collect data on what types of fish are available and on what quantities of fish may be caught by certain gear types. Fisheries surveys may be carried out once only or they may be repeated at regular or irregular intervals as the need arises. In particular, exploratory surveys are often used to test catches in new fishing areas and/or using new types of fishing gear. The Deep Sea Fisheries Development Project of the South Pacific Commission is an example of a very successful programme of exploratory surveys in the Pacific region. In this programme, the results from surveys in each country are published with full details of the methods used and the results obtained. A typical example is a report by Taumaia and Preston (1984). Intensive fisheries surveys may be of considerable importance to stock assessment and economic studies since the sampling methods usually are well documented and the results are available for future comparisons. By contrast, units of effort and identification of catch from routine data collection are often non-standardised and make comparison from year to year and area to area very difficult.

1.5 How do we present and analyse the data we have collected

Most of the remaining topics in this course will be concerned with simple but useful methods for processing, summarising and analysing the data collected. In Topic 2 are introduced some basic statistical concepts and terms which will be necessary for future work. Topic 3 covers frequency distributions, and Topic 4 deals with statistics which may be used to summarise and describe data. In Topic 5 we will look at some methods for analysing the relationship between different statistics such as catch against time. Topic 6 is devoted to the common methods of sampling as a means of obtaining estimates of statistics when we cannot get a complete collection of data.

References

ANON (1982). South Tarawa artisanal fishery 1981: Report on a fisheries survey carried out on south Tarawa (19 January to 13 February 1981). Fisheries Statistics Unit, Fisheries Division, Tanaea, Tarawa, Republic of Kiribati, 47 pp.

LAL, P.M. & C. SLATTER (1982). The integration of women in fisheries development in Fiji: Report of an ESCAP/FAO initiated project on improving the socio-economic condition of women in fisherfolk communities. Fisheries Division, Ministry of Agriculture and Fisheries, Fiji, Centre for Applied Studies in Development, University of the South Pacific, Fiji, 177 pp.

MUNRO, J.L. (1982). Estimation of biological and fishery parameters in coral reef fisheries, p. 71-82. In Pauly, D. and G.I. Murphy (eds.). Theory and management of tropical fisheries. ICLARM Conference Proceedings 9, 360 p. International Center for Living Aquatic Resources Management, Manila, Philippines and Division of Fisheries Research, Commonwealth Scientific and Industrial Research Organisation, Cronulla, Australia.

PAULY, D. & G.I. MURPHY (1982). Theory and Management of Tropical Fisheries: Proceedings of the ICLARM/CSIRO Workshop on the Theory and Management of Tropical Multispecies Stocks, 12-21 January 1981, Cronulla, Australia.

TAUMAIA, P. & G.L. PRESTON (1984). Report of visit to Cook Islands, 10 September 1981-29 March 1982. Deep Sea Fisheries Development Project. South Pacific Commission, Noumea, New Caledonia.

TOPIC 2 - STATISTICAL METHODS : SOME IMPORTANT CONCEPTS

2.1 Some Basic Definitions

In statistics, as in any other subject, we need to define some words and phrases so that we can use them to have a specific meaning in a particular situation. We shall try to avoid using much "jargon", but we shall need some technical terms. One possible source of confusion here is that many of these words are used in everyday English, but in statistics their meaning is a little different from normal usage. It is worthwhile, therefore, taking a little time to make sure that these terms are understood since we shall use them a lot throughout this course.

When we collect data, for whatever reason, we need to know exactly what kind of information we want, who or what this refers to, how we are going to obtain the data, and for what group of people or items. We use special terms to refer to each of these things; we talk about observing characteristics for statistical units in some population. The terms observing, characteristic, statistical unit and population all have special meanings and we shall look at all of these to see what their definition is and how the term is used.

Statistical unit - We use this term to mean any person, group of people, item or thing about which we wish to obtain some numerical information. Examples of statistical units are: a person, a family, a household, a village, a building, a province, an island, a fish, a boat, a port, a period of time such as a week or a year, a church, a business establishment, and so on. We could think of many more examples.

Population - When we collect statistical data we are interested in obtaining information about a group of statistical units - we use the word "population" to refer to this whole group. In English we usually use the word "population" to mean a group of people; we talk about the population of Suva or the population of Tonga, for example. In statistics we can use "population" to mean a group of any type of statistical unit; thus we can refer to the population of all households in Apia, the population of fishing boats on Rarotonga, the population of all fish caught in Tuvalu in the year 1983, and so on. When we want to collect statistical data we have to be very careful to define exactly what population it is that we are interested in.

Observation - We use this word to stand for the method we use to collect any particular item of information. Usually an observation will be carried out by a person, sometimes with the help of instruments, but there are examples of some machines which will make observations and record the data automatically. It is important to realise that, in the statistical sense, observation can mean any method of collecting data, not just the physical act of seeing and noting something down. Common methods of statistical observation are: measurement, counting, personal judgement, conducting an interview, copying from existing records, a person completing a questionnaire, using self-recording instruments, and so on.

Characteristic - This word is used to stand for some feature or property of the unit that we are interested in. We could, for example, record or observe the weight of a fish, the area of a farm, the value of all goods imported in a port, the annual income of a household, the number of people living in a village, and so on. In most situations, of course, one unit may have many different characteristics, all or some of which we may observe.

For example, when data are obtained about a pole-and-line fishing trip in a country's waters some of the characteristics of the trip which may be collected are:

port of departure;
country of registration;
gross tonnage of vessel;
number of crew;
days spent fishing in territorial waters;
species of fish taken;
quantity of fish taken;
average weight of fish.

A characteristic can be one of two types. In the first case it may be expressed only in numerical values, and we call this type of characteristic a variable. Other characteristics do not take numeric values, and have to be described in words. This second type is referred to as an attribute.

From the list of characteristics of the pole-and-line trip given above, we can identify the following as variables:

gross tonnage of vessel;
number of crew;
days spent fishing;
quantity of fish taken;
average weight of fish.

The attributes in the list are:

port of departure;
country of registration;
species of fish.

It is often more convenient, especially when using a computer, for statisticians to work in numbers rather than in words. Therefore, we sometimes allocate numerical codes to attributes. For example, we might allocate code 001 to skipjack, 002 to yellowfin, 003 to bigeye, 004 to albacore, and so on. Then we would key into the computer this code number, rather than the name of the species. However, it is important to recognise that, even though this characteristic has now been recorded in numeric form, it is still an attribute, not a variable.

2.2 Notation

During this course we will make use of some special statistical notation. This is the statistician's shorthand way of expressing a concept which would otherwise be cumbersome and long-winded to express. We will try to keep the notation as simple as possible.

This special notation will be introduced progressively during the course, but a few basic symbols should be described immediately.

n, N : The number of observations under consideration is denoted by "n" in the case of a sample, and "N" in relation to the whole population. Thus, if we collect data from a sample of 17 fishing boats for a survey, we say that $n = 17$. If the fleet consists of 80 boats, we say that $N = 80$.

x — When one variable is being considered, "x" is used to denote the values of the observation. This symbol is often followed by a subscript to describe exactly which observation is referred to. That is, x_1 refers to the value of the first observation, x_2 to the value of the second observation, and so on up to the final (i.e. nth) observation, which is denoted by x_n.

Thus, if we were measuring fork length of fish in centimetres, and the length of the first fish in our sample was 62 cm, we would say that $x_1 = 62$. (The whole set of n observations can be described by reference to x_1, x_2, x_3, ... x_n.)

y — When we are considering two variables, the second variable will be denoted by y with subscripts as required. So if we were conducting length-weight comparisons, and the first fish weighed 3.8 kg, we would say that $x_1 = 62$ and $y_1 = 3.8$.

i — For convenience any particular observation is referred to as the ith observation. So in order to refer to the value of the first observation we say i = 1 and so on. We will find that "i" is most often written as a subscript. So that for instance x_i means the ith observation of our variable x.

$\sum$ — This is the Greek letter, capital sigma, and means simply "the sum of". It must not be confused with σ, which is the ordinary Greek letter sigma, and which will be introduced later in the course.

2.3 Diagrams

In the following topics in this manual, we will be representing a number of statistical concepts in diagrammatic form. We will use three types of diagram - a scatter diagram, a graph and a bar chart. At the same time we will make brief mention of pie charts, which, although not specifically required in the later topics, are a very useful way to portray information diagrammatically.

The topic of diagrams is a very important one, and the way diagrams are used to portray results of statistical surveys can greatly affect the understanding of the results. However, for the present we are not going to explore this topic in detail; we will simply touch on the basic principles of constructing these types of diagram, as a lead-in to the following topics.

A diagram is used to demonstrate the relationship between characteristics. The main components are:

(a) Heading: essentially a diagram number and a title, describing what the diagram represents;

(b) Two axes: a vertical or "y" axis and a horizontal or "x" axis. These meet at the origin ("0"). Each axis must be clearly labelled and values of the variable or attribute are plotted along it according to some scale;

(c) The data: plotted on the diagram, depending on the type of diagram being used.

2.3.1 Scatter diagram

When we draw a scatter diagram, we are looking at the relationship between two characteristics. In general, we shall have a number of observations of both of these for a series of statistical units. We shall be concerned here almost exclusively with variables. We shall assume that we have a sample of n units, and for each unit we shall observe two variables which we can denote by x and y. Thus, for the first unit, the observations can be written as x_1y_1, for the second unit x_2y_2, and so on. In general, for the ith unit, our observations will be written as x_iy_i and there will be n such pairs. If we have a graph with two axes to represent the two variables, then each pair of observations can be plotted as a point; the co-ordinates of the pairs will be the values (x_i, y_i). This kind of graph with all n observations plotted as a number of points is called a "scatter diagram". It is a very useful first step in looking at the relationship between x and y.

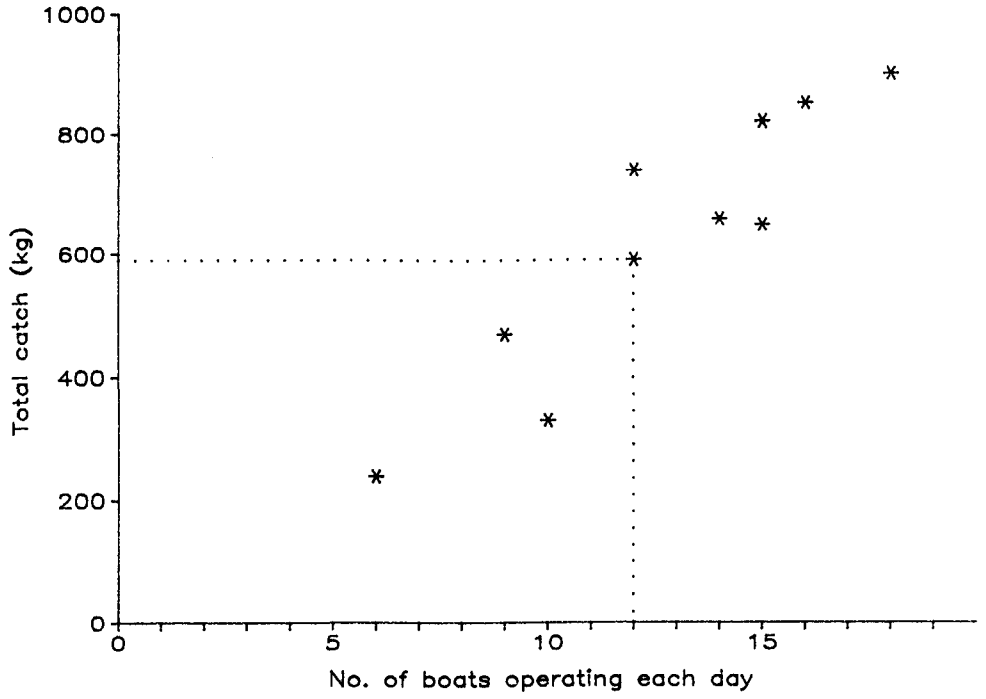
For example, suppose that on 10 selected days we recorded details of the number of boats fishing, and the total daily catch from an artisanal fishery (Table 2.1):

TABLE 2.1 : BOATS OPERATING AND DAILY CATCH AT AN ARTISANAL FISHERY

Day	Boats operating (x)	Total catch (kg) (y)
1	12	590
2	15	820
3	10	330
4	12	740
5	18	900
6	14	660
7	6	240
8	15	650
9	16	850
10	9	470

In a scatter diagram we will show number of boats on the x-axis and catch on the y-axis. The result for the first day is plotted as a point, where the vertical distance above the x-axis is equal to 590 kg and the horizontal distance from the y-axis is equalivalent to 12 boats. This point represents the pair of observations x_1y_1, as shown in the diagram. The dotted lines are also drawn in to show exactly how this point was located. In practice, however, in constructing a scatter diagram we are interested in showing just the distribution of points. Figure 2.1 is the scatter diagram showing all 10 pairs of observations.

FIGURE 2.1 : NUMBER OF BOATS OPERATING AND TOTAL CATCH PER DAY

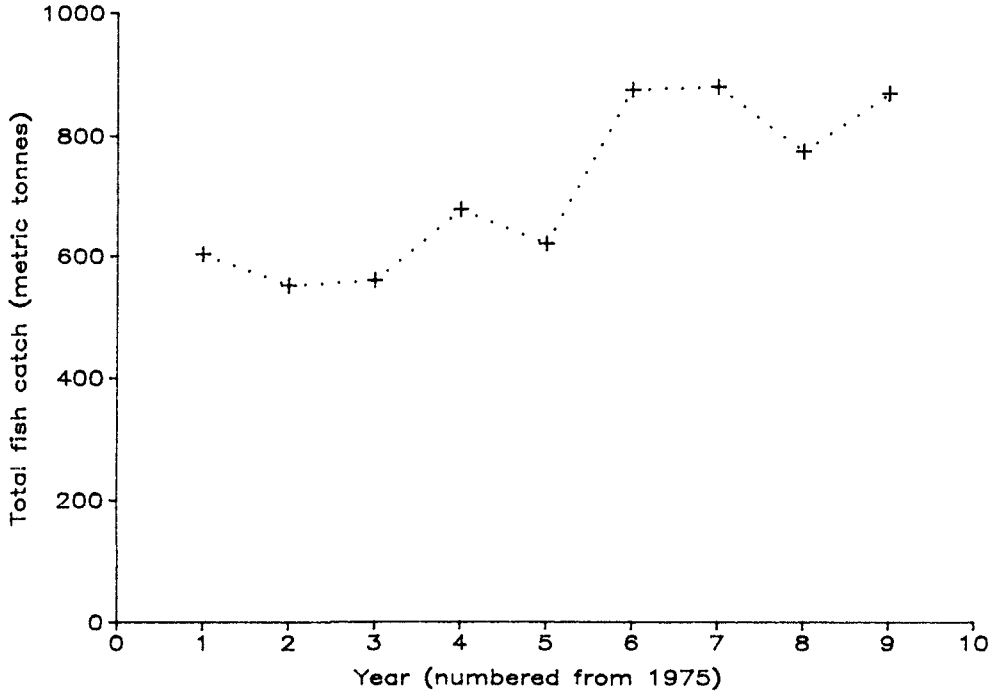


2.3.2 Graphs

A graph is very similar to a scatter diagram, showing the relationship between two variables, but with the points linked up by lines, to show the trend in the relationship.

Graphs are very useful when we show how some variables change over time, as in Figure 2.2.

FIGURE 2.2 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC



We may note that drawing lines to link up points has a real meaning in this situation, because the slope of each line shows whether catch is going up or down from one year to the next. However, linking up points in the previous diagram would not make sense. It would not be showing a trend.

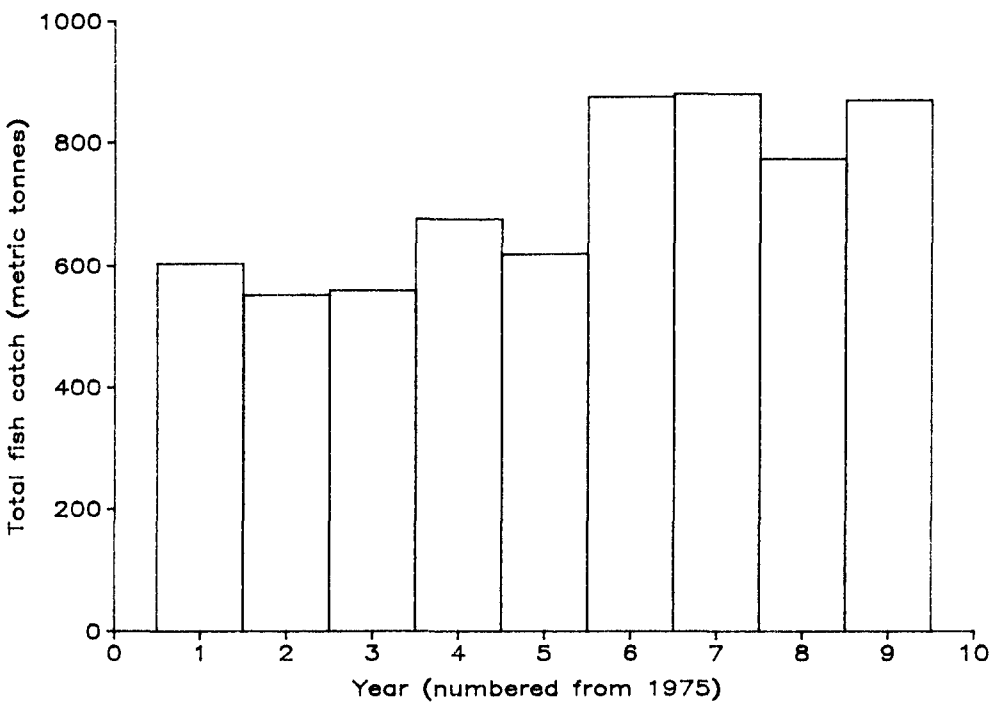
Graphs are not necessarily constructed by straight lines joining up a series of points, as in that illustration. Often we have curves, to represent the shape of different distributions, and we will study that in the next topic.

2.3.3 Bar charts

If we wish to prepare a diagram of data classified by some attribute, rather than by a variable, then a line graph is not suitable. For example, if we have statistics on production of fish by district, we cannot place districts along the x-axis and join up a series of points by a line. Such a line would be meaningless. In this situation the best form of presentation is a bar chart.

However, the use of this type of diagram is not restricted to attributes. We can also depict relationships between variables on a bar chart. Figure 2.3 shows the same data of fish catch over several years, which we previously depicted by a line graph, in the form a bar chart.

FIGURE 2.3 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC



The bars can be drawn fairly close together, or further apart, and may be shaded or cross-hatched to improve the appearance. A little later, we will look at a particular type of bar chart, called a histogram, in which data is represented in a series of bars which are contiguous.

2.3.4 Dependent and independent variables

Having chosen the type of diagram we require in order to best illustrate the data, we next need to decide which characteristic will be plotted on the x-axis, and which on the y-axis. To determine which way round to draw the diagram, we need to see if there is likely to be any form

of relationship between the characteristics. In many cases we can say that we are interested in seeing how one variable changes as another variable or attribute changes. In our example of total fish catch in Figures 2.2 and 2.3, we are trying to show how the level of catch changes as time changes. In Figure 2.1, we are interested in how the catch varies according to the number of boats engaged.

In Figures 2.2 and 2.3, we may say that the total catch depends on time; in Figure 2.1, that catch depends on the number of boats engaged. More formally we say that we have a dependent variable and an independent variable (or attribute). In these situations we cannot reverse the relationship. It would be silly to say we were looking at how time varied depending on the level of fish catch, for example.

Given this kind of independent-dependent relationship we always plot the dependent variable on the y-axis and the independent variable or attribute on the x-axis. This is a mathematical convention, it is used for convenience, and it makes diagrams easier to understand.

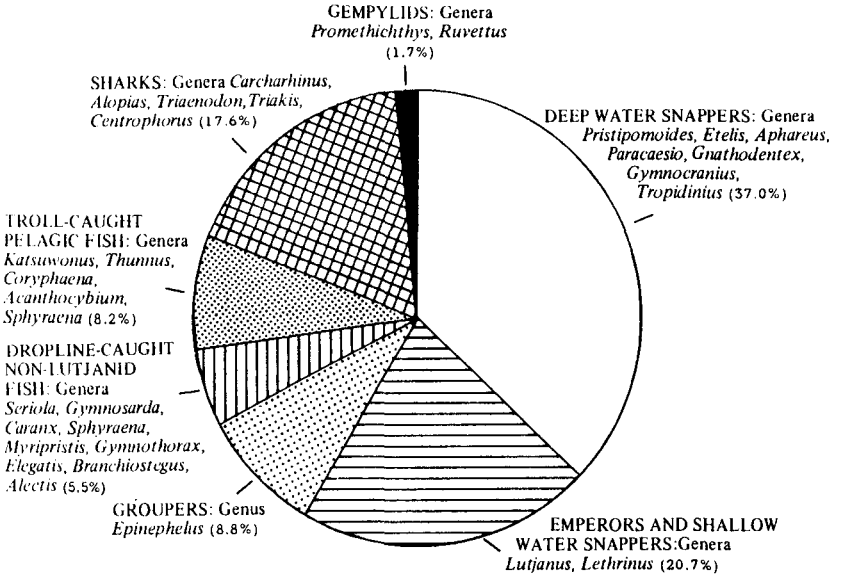
Where there is no clear direction of dependence between the characteristics then it does not matter much which one is plotted on which axis. This situation will not occur very often in practice, as we will usually find that one variable can be considered to depend on the other.

An example, which we will study later, is the relationship between the length and weight of fish. It may be argued that each one depends on the other, and that there is no clear dependent/independent relationship. Even here though, there is a convention, and it will be found that weight is always plotted on the y-axis, and length on the x-axis.

2.3.5 Pie charts

A basic pie chart consists of a circle divided into a number of sectors. Each sector is used to represent a particular value of a characteristic, the area of each sector being proportional to the share of that characteristic to the total. Either variables or attributes can be portrayed in this manner, but a pie chart is especially useful for attributes, as in Figure 2.4.

FIGURE 2.4 : CATCH COMPOSITION BY WEIGHT (GENERA WITHIN GROUP LISTED IN DECREASING ORDER OF IMPORTANCE)



The normal practice, as shown in this figure, is to commence at the top of the chart (the "12 o'clock" position) and to work clockwise from there, with the largest or most important sector being shown first.

A pie chart such as this is quite easy to prepare. The area of any sector of a circle is proportional to the angle at the centre between the two radii. Since the value of the characteristic has to be proportional to the area, all we have to do is to draw a number of sectors with angles proportional to the value of the characteristic. The sum of the angles will of course be 360 degrees. The calculation for any category is then quite simple.

Notice also, that since we are dealing with proportions we can prepare the pie chart either from the actual data or from a percentage distribution. With too many categories a pie chart becomes confused and difficult to read; as a general rule eight is about the maximum number that should be included.

2.4 Rounding of numbers

During later topics we will encounter situations where we need to round numbers to a certain number of significant digits, or to the nearest one decimal place. It may also be necessary in publishing survey data to present results rounded to the nearest tonne, or to the nearest thousand, etc. To ensure that this is done in a consistent way, we need a standard rounding procedure.

The basic principle is to round to the nearest significant digit. Thus, if we wish to round 428,548 to the nearest thousand, we would record this as 429,000. When we need to round a number which is exactly halfway between two significant digits, we adopt a convention of rounding so that the last significant digit is even. So we would round the number 428,500 to 428,000, in preference to 429,000.

We should note here that, when a series of numbers and the total of those numbers are rounded, it may happen that, after rounding, the sum of the components is not equal to the total. For example, let us consider the following, where a set of numbers is to be rounded to the nearest thousand.

128,613	rounded to	129,000
428,548		429,000
37,924		38,000
595,085		?

Clearly the total should be rounded to 595,000 according to our rules, and yet the sum of the three rounded numbers is 596,000.

This gives us a problem in how to present the data in rounded form, and as similar situations will often arise in practice, we need a convention to deal with it. It is recommended that each number be rounded correctly according to the rules (the total in the example being rounded to 595,000), so that the sum of the components may not be exactly equal to the total. This disadvantage of this is that users may notice that the total does not correspond exactly with the sum of the components, and may conclude that an error has been made. To counteract this it is normal practice to include in publications a note, such as "Any discrepancy between totals and the sum of components is due to rounding".

TOPIC 3 - FREQUENCY DISTRIBUTIONS : REDUCING A LOT OF DATA INTO A MANAGEABLE FORM

3.1 Background and Introduction

When we undertake a statistical investigation we end up with a series of observations of some characteristic for a number of units. Usually, of course, we will have a number of observations of different variables and attributes for each unit, but to keep the situation fairly simple at present we will only look at one variable. Given, then, the series of observations, we want to find out some way we can summarise this information, so that we may begin to make sense out of it. We may want to make some kind of decision, or inference, about the whole group of units, perhaps to compare them with some other group; we may want to make some kind of estimate for the whole population of units of which our group may only be a small part; or we may just want to have some convenient way to summarise the basic data, to reduce the amount of information to a manageable size.

We shall start off by looking at a frequency distribution as a method of summarising the basic data; later on in this topic we shall see how we can develop a theoretical basis. As an example, consider the information given in Table 3.1. This represents a summary of 1,870 observations of the weights of mangrove crabs packed for export. We prepare a frequency distribution by dividing the range of weights we observe into a number of classes and then counting the number of observations in each class.

TABLE 3.1 : EXAMPLE OF A FREQUENCY DISTRIBUTION (NUMBER OF MANGROVE CRABS PACKED FOR EXPORT ACCORDING TO WEIGHT)

Weights of Crabs (g)	Number of Crabs
200 to less than 300	55
300 to less than 400	302
400 to less than 500	540
500 to less than 600	357
600 to less than 800	290
800 to less than 1,000	176
1,000 to less than 1,200	59
1,200 to less than 1,600	52
1,600 and over	39
Total	1,870

This frequency distribution has summarised 1,870 observations into 9 groups or classes, together with a total figure. Obviously this is much easier to comprehend than a list of 1,870 individual values would be, even if those values were sorted into size order. At the same time we have lost some of the original detail; we do not know the actual value of any of the observations.

A typical frequency distribution then, divides the range of values of the characteristic we are considering into different classes and counts the number, or the frequency, of observations within each class. We can construct frequency distributions for both variables and attributes, but procedures for attributes are quite straightforward so in this topic we will concentrate only on variables. Later on we shall distinguish between two basic types of variables.

A frequency distribution is particularly useful if we wish to find out how the values of a variable are distributed. It shows at a glance the range of values, how many high values and how many low values have been observed, what the most frequently occurring values are, and whether the values are symmetrically distributed along the range, or whether they are mostly at one end.

3.2 Construction of a frequency distribution

Preparing a frequency distribution from a list of the basic data consists of three steps:

(a) specifying the classes into which the data are to be grouped;

(b) sorting the data into these classes; and

(c) counting the number of observations in each class.

The last two of these steps are quite straightforward, but it can be quite difficult to decide on the number of classes we need, and on the range of values for each class. In Table 3.1, on crab weights, we chose nine classes, and the class interval (or range) was 100 g for the first four classes, 200 g for the next three, 400 g for the next one, and unlimited for the last class (i.e. it was open-ended, and any observation of 1,600 g or more would have been included in it).

Although in principle we can decide on any set of classes we like for a distribution, and the definition of each class will depend on the purpose of the distribution, there are some guidelines which it is useful to follow:

(a) The classes chosen must span a range sufficient to encompass every observation, from the lowest to the highest.

(b) There should be no gap or overlap in the classes; each should be separate and distinct. It is particularly important that the range of each class should be defined so that each observation can only go into one class. If in our previous example we had carelessly described the classes as 200-300, 300-400, 400-500 and so on, we would not know how to classify a crab weighing exactly 400 gms, as it could go into either of two classes. We must be sure that there is no ambiguity, and that observations which are on the border between two classes will fit into only one of them.

(c) There should not be too many classes (as this will lose the advantage of a frequency distribution over the raw data), nor too few (as too much information will be lost). In general, it is suggested that more than 5, and not more than 16, separate classes are desirable, but these are no firm limits. The number of classes formed will depend on the nature of the data, on the number of observations, and on the type of distribution.

(d) It is advantageous, whenever practicable, for the class intervals, that is, the range (or length) of each class, to be the same. Class intervals of equal length make it much easier to comprehend the distribution and to draw suitable diagrams. If unequal intervals are used it is often difficult to compare one class frequency with another. Sometimes, however, it is impossible to avoid unequal intervals; the variability of the data requires their use.

(e) Whenever possible avoid the use of open-ended intervals, that is classes at the ends of the distribution in which one end value of the class is not stated.

As we have already observed, our frequency distribution of crab weights does not have equal class intervals, as we have suggested in (d) above. We could try to combine data into classes with equal intervals of, say 400 g, and the frequency distribution would then become as shown in Table 3.2.

TABLE 3.2 : NUMBER OF MANGROVE CRABS, ARRANGED IN CLASSES WITH EQUAL INTERVALS

Weights of Crabs (g)	Number of Crabs
Less than 400	357
400 and less than 800	1,187
800 and less than 1,200	237
1,200 and less than 1,600	52
1,600 and over	39
Total	1,870

It will be seen that this distribution does not contain as much useful information as the first table, as one class now contains nearly two-thirds of all cases (1,187 out of 1,870). The breakdown of this class into smaller classes, as we had originally shown, is desirable to give additional information on the size of crabs in this large class. Generally it will be found that the more evenly the observations are spread, the easier it will be to construct a frequency distribution with equal class intervals.

It will also be noted that the final class of our distribution is open-ended, despite the recommendations of principle (e) above. However, it is often quite difficult, or even virtually impossible to avoid their use. For example, there may have been one or two very large crabs, of say 3,000 g, and we would have had to break down the final class into several classes, with different intervals, in order to avoid the open-ended class we have used. Since there are only two per cent of all observations in this class, a further breakdown would not have been desirable.

We will now use a different set of data to illustrate how a frequency distribution is constructed. The data represent the weight in kilograms of a sample of 63 yellowfin tuna caught by pole-and-line method (Table 3.3). For this exercise it is assumed that all weights are rounded to the nearest 1/10 kg.

TABLE 3.3 : WEIGHTS OF 63 YELLOWFIN (in kilograms)

4.6	3.9	2.8	6.6	4.2	3.7	3.7	5.9
3.2	2.2	3.2	4.1	3.1	3.0	4.8	4.1
2.1	4.2	5.0	4.6	5.4	2.4	6.3	2.9
5.3	4.0	4.7	3.6	3.3	6.9	4.5	2.5
5.4	5.7	3.8	4.1	5.6	6.2	3.0	3.3
5.0	5.4	3.4	4.4	4.0	3.6	5.0	4.1
4.8	7.2	6.4	3.0	3.5	5.8	7.7	3.9
2.6	7.9	3.3	5.5	4.3	3.9	6.3	

To determine the different classes we first of all need to know the range, from the lowest to the highest value. In this case the lowest value is 2.1 kg, the highest 7.9 kg, the range therefore is 5.8 kg. We wish to split the range up into a number of classes, and the observations are so evenly spread that there seems no reason with this data why we should choose classes with unequal intervals. A suitable distribution, then, is given in Table 3.4.

TABLE 3.4 : FREQUENCY DISTRIBUTION OF YELLOWFIN WEIGHT DATA

Weights (kg)	Frequencies
2.0 - 2.9	7
3.0 - 3.9	19
4.0 - 4.9	16
5.0 - 5.9	12
6.0 - 6.9	6
7.0 - 7.9	3
Total	63

3.3 Definition of terms

We need to define some terms when talking about frequency distributions. Some of these we have used already; in this section we shall define the terms more closely.

(a) Class frequency

The class frequency in a distribution gives the number of observations falling within that particular class. When presenting a frequency distribution in tabular form, the classes always go in the left hand column, with the class frequencies on the right.

(b) Class limits

The smallest and largest values (rounded where necessary) that can go into any given class are termed its class limits. In the yellowfin weights table the class limits are 2.0, 2.9, 3.0, 3.9, and so on. We usually differentiate between the lower class limits (2.0, 3.0, 4.0, etc.) and the upper class limits (2.9, 3.9, 4.9, etc.).

(c) Class boundaries

These represent the actual, or true limits to a class. There is a fine distinction between class boundaries and class limits, and it is important to be clear on this distinction. In our example we may note that a fish weighing (say) 2.96 kg will be recorded in the survey as weighing 3.0 kg. The class boundaries in this example are actually 1.95, 2.95, 3.95, and so on.

(d) Class marks

The class mark is the mid-point of the class, and is obtained by taking the arithmetic mean of the upper and lower class limits. In the example, the class marks are 2.45, 3.45, 4.45, etc. These are often also referred to as mid-marks, mid-points, mid-values, etc.

(e) Class interval, or range

The class interval is the length of any class, the range of values it contains. The class interval of a class is the difference between the lower class limit of that class and the lower class limit of the next class. If all the intervals are equal then it is also equal to the difference between successive class marks. For example, the class interval for the yellowfin weights is 1.0 kg and is equal for all classes. Note that the class interval is not necessarily the difference between the upper and lower limits of the class. (In our table this is equal to 0.9 kg.)

3.4 Continuous and discrete data

At this point in the study of frequency distributions, we need to distinguish between two different types of variable, because the problems of constructing a frequency distribution and drawing diagrams of the distributions are somewhat different for each type. The first is where the variable is allowed to take any value within a specified range, and the second where the variable can only take certain values. The first type we call a continuous variable and we also refer to continuous data; the second type we call discrete. Examples of continuous variables are:

(a) fork length of fish;
(b) weight of fish;
(c) water temperature of sea surface.

Examples of discrete variables are:

(a) number of canoes in a village;
(b) number of longline sets;
(c) crew number.

In most cases discrete variables take whole number (or integer) values, although this is not essential.

In practice, the dividing line between continuous and discrete data is often very difficult to discern. Continuous data will not normally be recorded in an absolutely continuous way, as there are limitations to the accuracy with which we can measure or record a variable. For example, we may be able to measure the weight of a fish to the nearest 1/10 kg - or even to the nearest gram if we had accurate enough equipment - but that is as fine a breakdown as we could hope to achieve.

Likewise some discrete data can be dissected so finely that it looks like continuous data. Revenue from fish sales could be recorded to the nearest dollar, or even to the nearest cent, and we would have such a spread of recordings that we could treat this as though it were continuous. There is a fine distinction, however. Fish weights could take any value whatever in a range, and it would still be meaningful; revenue on the other hand cannot be expressed in fractions of a cent, because a cent is the smallest unit of currency which exists.

While quite often the division between continuous and discrete data is blurred, there are situations where the distinction is important. For instance, if we are discussing the number of fishing units in a village, or the number of landing sites on an island, we will have a discrete distribution which does not look at all like a continuous distribution.

We can illustrate this by looking at two examples of constructing frequency distributions from discrete data. The situation is simple as long as we have one value of the variable only in one class. For example, we can quite simply present data summarising the number of villages with 0, 1, 2, 3, 4, etc. powered fishing boats as in Table 3.5.

TABLE 3.5 : NUMBER OF VILLAGES, BY NUMBER OF POWERED FISHING BOATS PER VILLAGE

Number of powered boats	Number of villages
0	20
1	7
2	12
3	28
4	17
5	10
6 or more	4
Total	98

Often, however, the range of values is so great that we have to combine values in each class. For example, a country may be interested in the distribution of visits by longline vessels to its territorial waters, classified by number of longline sets made during a visit. The distribution might be as shown in Table 3.6.

TABLE 3.6 : NUMBER OF VISITS BY LONGLINE VESSELS, CLASSIFIED BY NUMBER OF SETS PER VISIT

Number of longline sets	Number of visits
1 - 5	17
6 - 10	25
11 - 15	23
16 - 20	39
21 - 25	49
26 - 30	33
31 - 35	15
36 - 40	4
Over 40	5
Total	210

Although this table looks like the table for the distribution of weights of yellowfin, we must be careful to remember that the data are discrete. We cannot talk about a vessel making 18.3 longline sets.

3.5 Cumulative frequency distributions

A frequency distribution provides information on the number of observations in the different classes; we can tell at a glance from the table how many small-sized and large-sized observations there are. Often, however, we have a situation where slightly different information is required. What is of interest in this case is to find out how many observations are larger than some specified value, or how many observations

are less than a certain amount. For example, in our earlier exercise on crab weights, we may well be interested in how many crabs weighed less than 500 g, how many weighed 800 g or more, and so on. This type of information can be readily obtained from a cumulative frequency distribution.

Table 3.7 shows how we can construct a cumulative frequency distribution for the data on longline sets per visit.

TABLE 3.7 : EXAMPLE OF A CUMULATIVE FREQUENCY DISTRIBUTION

No. of longline sets per visit	No. of visits by longline vessels	Cumulative frequency (less than)	Cumulative frequency (greater than)
1 - 5	17	17	210
6 - 10	25	42	193
11 - 15	23	65	168
16 - 20	39	104	145
21 - 25	49	153	106
26 - 30	33	186	57
31 - 35	15	201	24
36 - 40	4	205	9
Over 40	5	210	5
Total	210		

The cumulative frequency distribution is obtained by calculating the progressive totals of the frequencies in each class. This can be done in one of two ways as illustrated in Table 3.7. The cumulative frequency distribution (less than) is calculated by starting with the first class, and then adding the cumulative frequency in each class until the last. The cumulative frequency distribution (greater than) is calculated by starting with the last class and working upwards.

As their name suggests, the two cumulative distributions are used to answer the questions: how many observations are greater than a certain value? Or, how many observations are less than a certain value? From the table we can see at a glance that 65 of the vessels made 15 sets or less during a visit to the country's territorial waters, and 106 made over 20 sets in a visit.

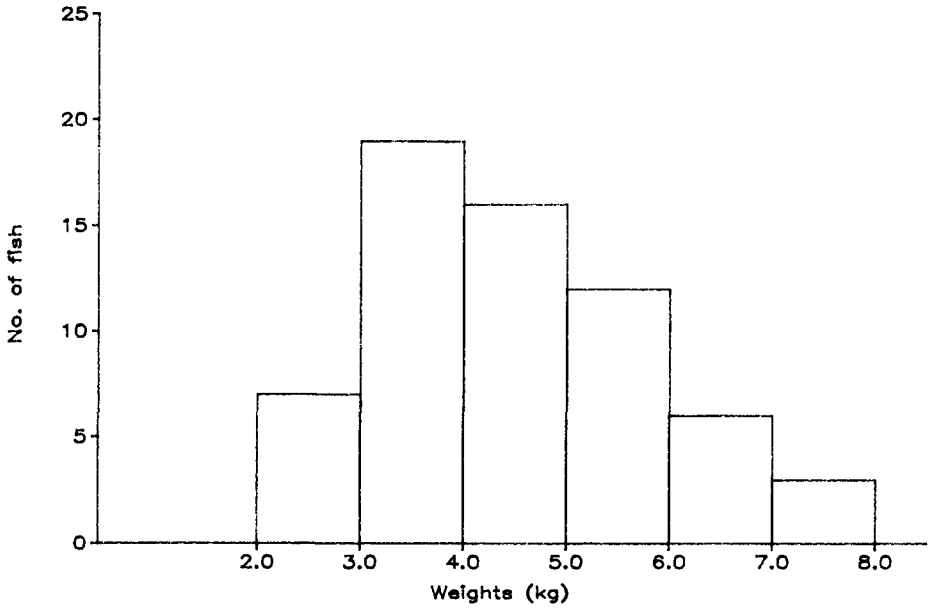
3.6 Diagrams of frequency distributions

A frequency distribution gives information on the way that a number of observations of a particular characteristic are "distributed" or spread out over a range of values. As well as preparing tables of these distributions, it is also important to have methods of representating them graphically, since in this way the different patterns in the data can be seen at a glance.

(a) Frequency histogram

The most common method of representing a frequency distribution is by drawing a frequency histogram. A histogram for our earlier data on yellowfin weights would be drawn as shown in Figure 3.1.

FIGURE 3.1 : FREQUENCY HISTOGRAM OF WEIGHTS OF 63 YELLOWFIN



To draw a frequency histogram we observe the following general principles:

(a) Magnitudes are drawn along the horizontal axis.

(b) Frequencies are plotted along the vertical axis.

(c) The general practice of all diagrams should be followed. Where applicable, headings, footnotes and full details of both axes should be provided.

(d) Frequencies should only be represented by a rectangle covering the whole of an interval if the data is continuous; discrete histograms are drawn somewhat differently, as will be shown a little later in this section.

(e) When plotting a continuous distribution it is the <u>area</u> of each rectangle and not the height that is proportional to the frequency. It is only in the case of equal intervals, as in the previous diagram, that the frequency is proportional to the height.

It is worth noting here that we have shown the weights of the classes of yellowfin in the previous diagram as 2.0, 3.0, 4.0, etc. These values represent the lower limit of each class. At the same time, we must remember that all the weights have been rounded to the nearest 0.1 of a kilogram and that the <u>true</u> class limits are 1.95, 2.95, etc.

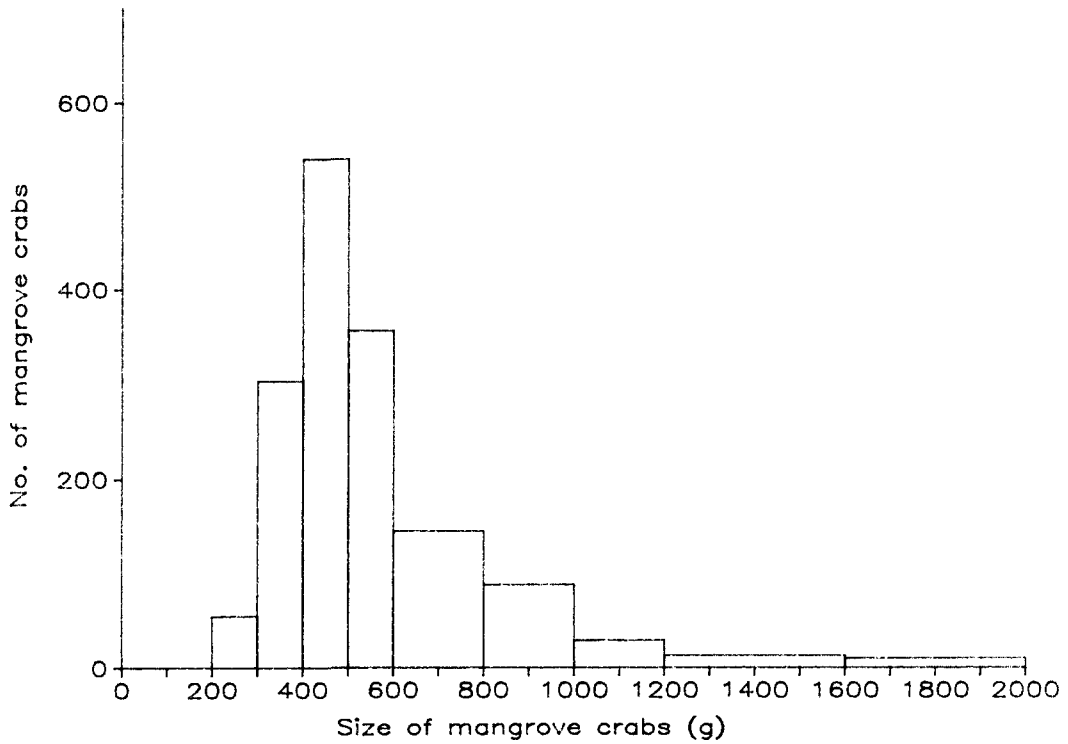
To illustrate how we should deal with a distribution with unequal intervals, we will return to our earlier example of mangrove crabs. In this case we have several different intervals for the crab weights. We must draw each rectangle so that its width is proportional to the class interval, but so that its area is proportional to the observed frequency. Perhaps the best way to achieve this is to calculate for each class a frequency which is an equivalent for the smallest class interval in the table (i.e. 100 g). Thus we can say the the 290 observations for the class 600 to less than 800 g is equivalent to 290 x 100/200 = 145 observations per 100 g interval. In this way all figures are brought to a common basis, and the heights in the diagram will be proportional to these. The calculations are as shown in Table 3.8.

TABLE 3.8 : CALCULATIONS FOR PLOTTING A HISTOGRAM WITH UNEQUAL CLASS INTERVALS

Class	Frequency	Equivalent frequency per 100 gram interval
200 and less than 300	55	55
300 and less than 400	302	302
400 and less than 500	540	540
500 and less than 600	357	357
600 and less than 800	290	290 x 100/200 = 145
800 and less than 1,000	176	176 x 100/200 = 88
1,000 and less than 1,200	59	59 x 100/200 = 30
1,200 and less than 1,600	52	52 x 100/400 = 13
1,600 and over	39	- -

The histogram of this then can be drawn as in Figure 3.2.

FIGURE 3.2 : NUMBER OF MANGROVE CRABS PACKED FOR EXPORT ACCORDING TO WEIGHT

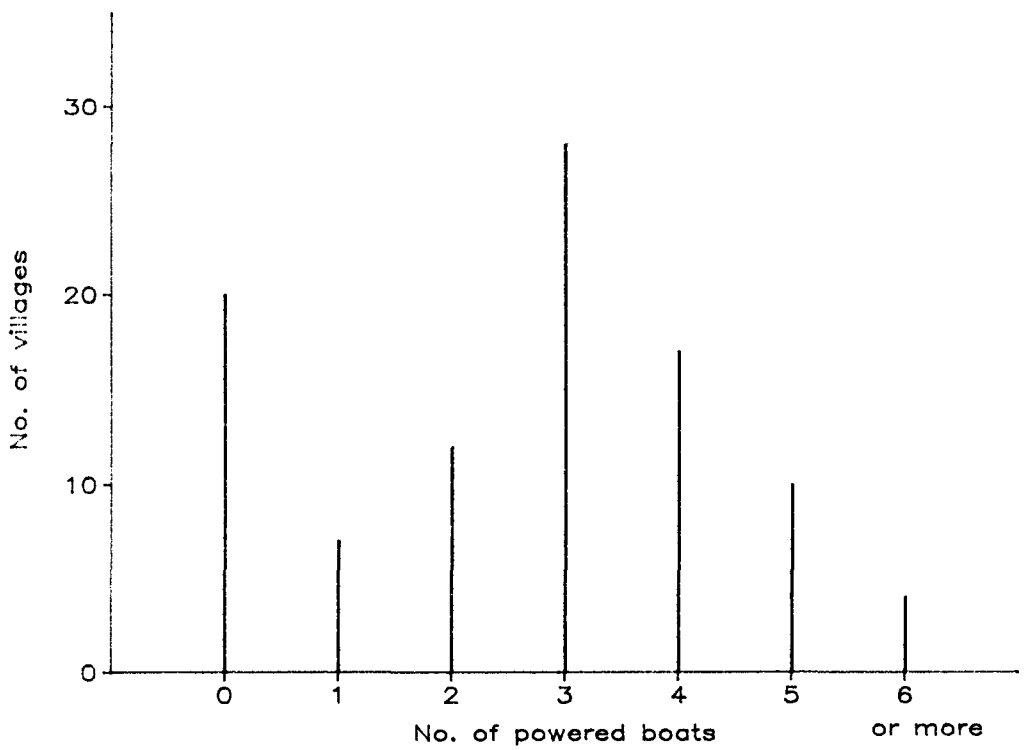


Apart from the fact that this distribution has unequal intervals, there is one other point that is interesting. There is a problem in deciding how to deal with open-ended classes, in this case the class "1,600 and over". Since we do not know the class width we cannot calculate the height of the rectangle to represent this part of the total frequency. It would be wrong to leave it out altogether, so we have to decide what to do. Basically the problem can be dealt with in one of two ways. Firstly, we can assume an upper limit for the distribution, and draw the rectangle accordingly. The second alternative is to draw the rectangle with a nominal height but leave it open, as in Figure 3.2. This indicates an open-ended interval. This second method only works when the frequency in

the open-ended interval is small (as is often the case), and in these circumstances, it is probably the better presentation.

When we are dealing with discrete data we can sometimes consider it to be approximately continuous if the unit of measurement is small compared with the size of the observation. Thus, for example, we may have a distribution of the number of villages according to the number of people in each village. Strictly speaking this is a discrete distribution since we cannot have fractional parts of a person. In practice, however, we can probably treat the data as continuous since the unit of measurement, one person, is small compared with the range of the data, which might be 1,000 people for instance. When we have a discrete distribution where the unit of measurement is large compared with the range of values, such as in our earlier example of the number of powered boats in a village, then we cannot draw a continuous histogram. Instead, we plot the frequencies by means of a simple bar chart of the type we studied in the previous chapter. Sometimes a single line instead of a bar is used as in the example in Figure 3.3.

FIGURE 3.3 : NUMBER OF VILLAGES BY NUMBER OF POWERED FISHING BOATS PER VILLAGE

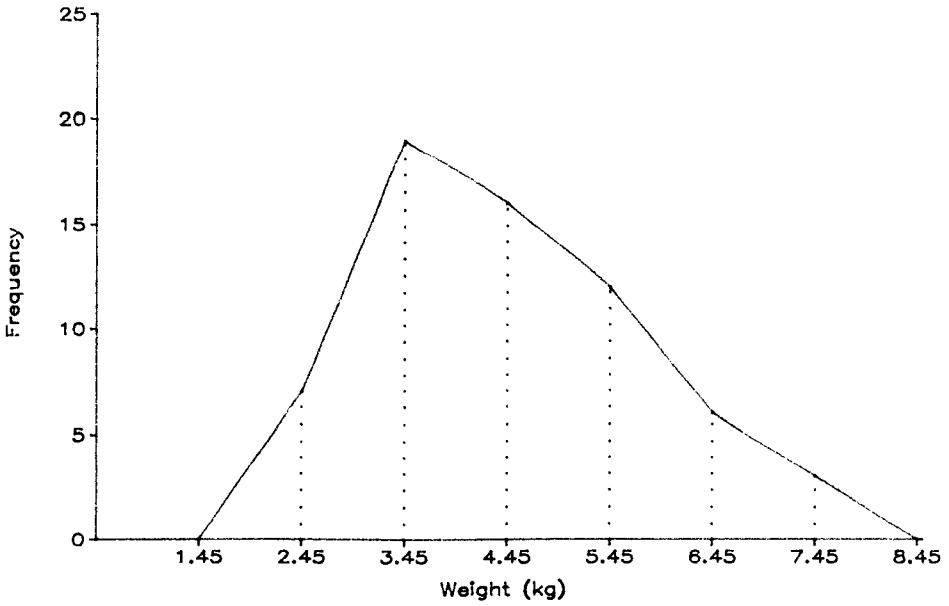


Notice the way the problem of class "6 or more" has been dealt with.

(b) Frequency polygon

An alternative type of diagram, suitable for continuous data, or for discrete data which can be considered to be approximately continuous, is a frequency polygon. In this type the frequencies of each class are plotted at the class mark, and successive points joined up by straight lines. An example of such a polygon for the yellowfin tuna data is given in Figure 3.4.

FIGURE 3.4 : FREQUENCY POLYGON FOR YELLOWFIN WEIGHT DATA

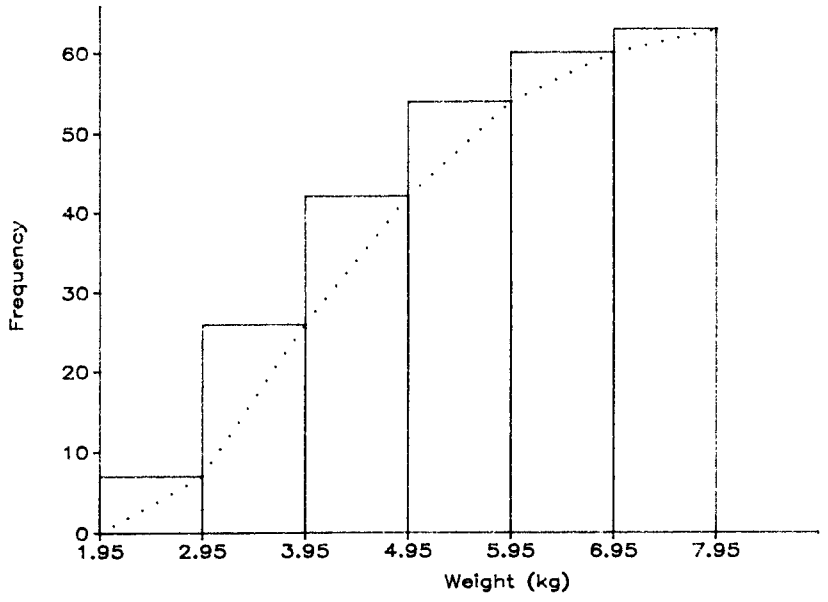


The beginning and end of the polygon should be extended to the x axis, at the mid-points of the classes below and above those covered by the distribution. The area under the polygon then is equal to the area of the rectangles in the earlier diagram. This area, as in the histogram, represents the total frequency.

(c) The ogive

Just as for some purposes it is better to construct a cumulative frequency distribution, so we also find it useful to represent cumulative distributions graphically. If we constructed a cumulative frequency distribution for the yellowfin data, and then draw a histogram, this would appear as a series of rectangles as shown in Figure 3.5.

FIGURE 3.5 : CUMULATIVE FREQUENCY HISTOGRAM OF YELLOWFIN WEIGHTS



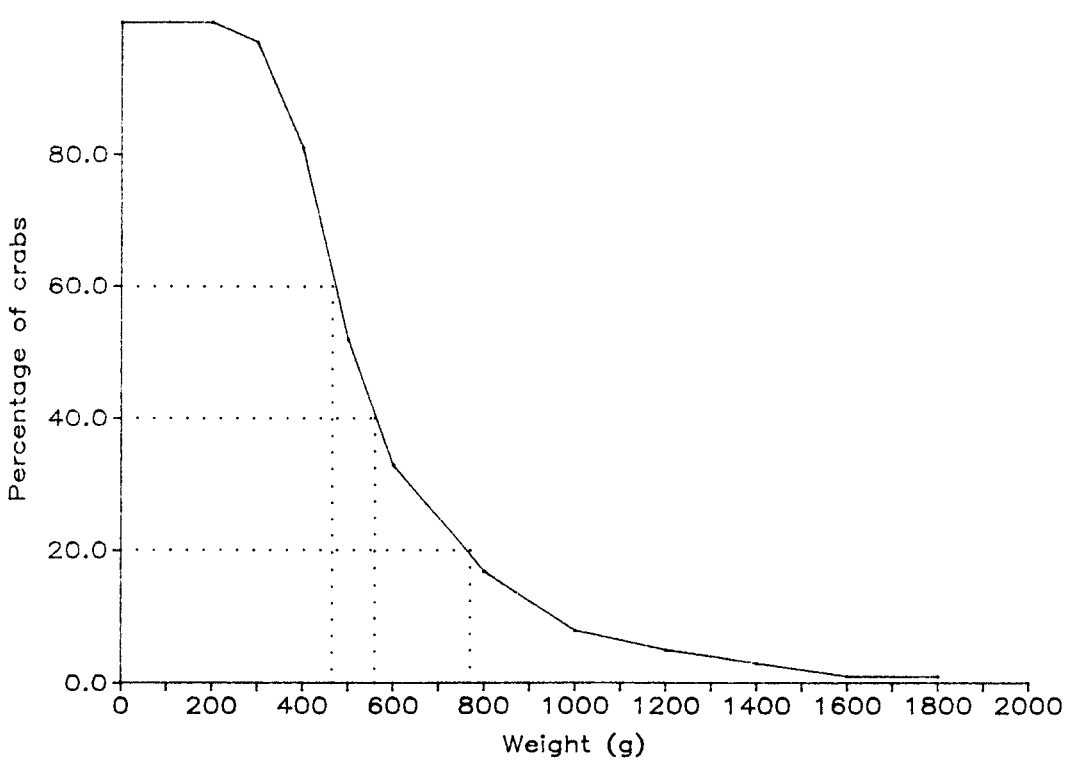
The dotted line in Figure 3.5 represents the cumulative frequency polygon; this is also called the ogive. This joins up the top right hand corner of each rectangle. This diagram represents the cumulative frequency distribution (less than). We can also draw a similar ogive for the cumulative frequency distribution (greater than).

Note that the ogive joins up points on the Figure which represent all the observations up to that point. The first class must therefore terminate at 2.95 kg, and not 3.0 kg; a yellowfin weighing (say) 2.97 kg would not be included in this first class.

This shows the need for care and precision, and highlights the importance of defining exactly what we mean by all our terms. In our crab weights example, we used a different approach and defined classes as "400 and less than 500 g", etc., so in that case the true class limits are 400 g etc. We could have adopted the same approach for the yellowfin, and defined classes as "2.0 kg and less than 3.0 kg", etc. In that case weights would not be rounded, and a fish weighing 2.97 kg would be included in that class. If we had done so, the ogive would be drawn through points at 3.0, 4.0 kg etc. (Our class marks would also be different - at 2.5 kg, etc.) We can record our data either way, whichever is more convenient for us, but we must then take care to make all our calculations accordingly.

Instead of plotting the actual frequencies on the y axis, as shown in Figure 3.5, we could convert this to percentages, and show the relative frequencies instead. The shape of the ogive would be exactly the same, but the y axis would be marked in percentages, from 0 to 100, instead of in numbers. The ogive (greater than) for our previous example of mangrove crab weights, expressed in percentages, is given in Figure 3.6. We can see from the lines drawn on this graph that it is very easy to derive estimates that (for example) 60 per cent of crabs weigh more than 485 g, 40 per cent weigh more than 560 g and 20 per cent weigh more than 755 g.

FIGURE 3.6 : PERCENTAGE OGIVE (greater than) FOR DATA ON WEIGHT OF MANGROVE CRABS



It is worth noting here that all figures of frequency distribution have exactly the same shape regardless of whether they are expressed in absolute values or in percentages. It is often preferable, particularly when attempting to draw general conclusions from data, to express these diagrams in percentages.

Ogives are useful for many purposes. We will see in Topic 4 that they can be used to calculate certain measures, particularly the median, very easily.

3.7 The distribution of the population

Sometimes when we construct a frequency distribution the data we use comprises the whole population that we are considering. Very often, however, the data we have is only part of the population, and we wish to make assumptions about the population from our sample. For example, we are not only interested in the frequency distribution of the 63 yellowfin we have been discussing, but we would hope to be able to say something about the whole population of yellowfin which are caught by pole-and-line method.

In this section we shall see in a very simple way how we can extend the idea of a frequency distribution to a population which has no limit on its size. We call such a population infinite, which indicates that we cannot count all the items in it. The same ideas will also apply to finite populations with a very large number of units; in many cases they can be assumed to be infinite.

Consider the frequency polygon for the yellowfin data which we drew earlier (Figure 3.4). Imagine that instead of six classes we have twelve, but that we double the size of the sample. If we draw a frequency polygon of this data we would expect it to have roughly the same shape as before, but because there are more lines, the shape would appear smoother.

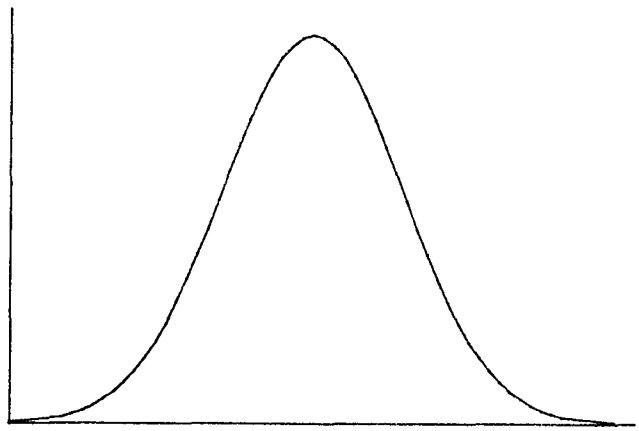
Let us suppose that we repeat the process again, taking a larger number of classes, but also a larger number of observations. The frequency polygon would be made smoother still. Eventually, as we take more and more classes, but increase the number of observations as well, we could expect to end up with a smooth curve. This curve represents the distribution of the whole population.

The idea of a population distribution is very important in statistics; it forms the basis for a great deal of more advanced statistical theory. We have no time to go into this theory in detail; all we can do is to introduce some of the concepts.

When we collect data from many different sources and then construct frequency distributions and draw histograms and polygons, in many cases we find that the shape of the distribution is quite regular. The distributions we have looked at so far all have quite a simple shape. This leads statisticians to seek a simple mathematical function that will describe or "fit" this shape.

Such a mathematical function, in the case of continuous data, will be smooth and will describe the distribution of the population. If we find that many different populations are distributed in more or less the same way, then we can use these mathematical functions to answer important questions about the population. For example, are the fish in one part of the sea generally larger than those from another part? Does one group of people have a larger income than another? What is the likely range of catch per unit effort from a certain fishing technique? and so on.

One of the most important shapes of a population distribution is known as a "normal distribution". A typical normal distribution is shown below:



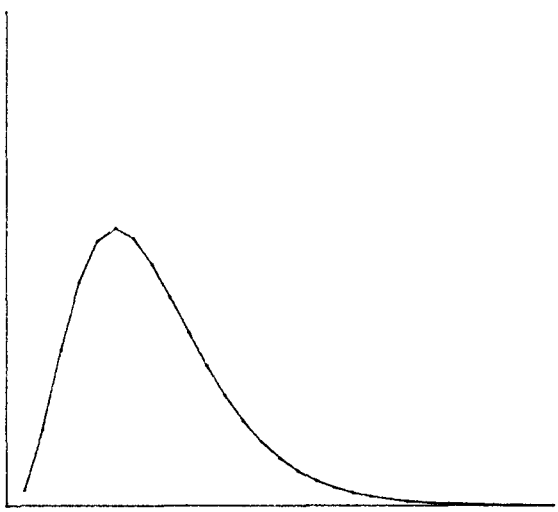
Normal

The normal distribution is symmetrical, that is, both sides are of the same shape. Some examples of distributions which have this kind of shape are:

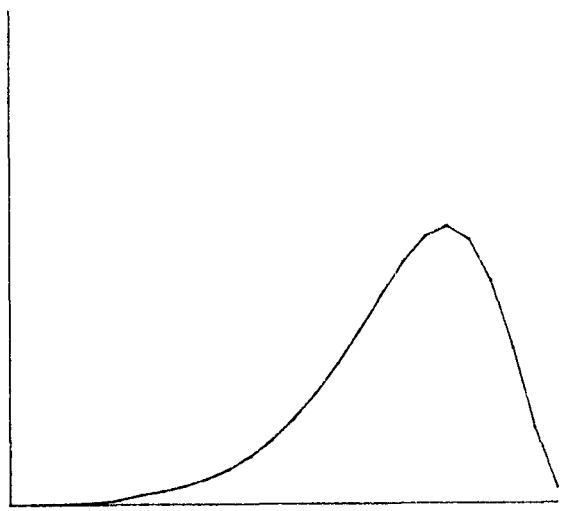
(a) numbers of adult men of different heights;
(b) the number of rainy days in a year;
(c) length frequency of a species of fish for a certain age class.

In these examples, observations will have some high and some low values but a predominance of "average" values.

There are a number of other shapes of population distributions which occur quite frequently, probably the most common being the skewed distribution. This looks like a normal distribution which has been pushed out of shape sideways so that it is no longer symmetrical, as depicted below:



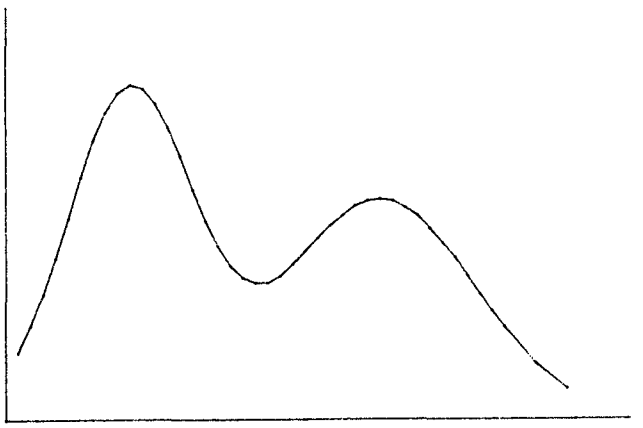
Skewed to right



Skewed to left

An example of a distribution which would almost certainly be skewed to the right is income distribution of the population; in many countries it will be found that there is a heavy concentration of incomes at fairly low levels, and then a long "tail" in the graph stretching out to the right, representing a fairly small number of people with very high incomes. Distributions of catch per unit effort are also usually strongly skewed to the right.

Another distribution quite often encountered is a "bi-modal" distribution. This has two distinct peaks, viz:



Bi-modal

Perhaps the best-known bi-modal distribution is of deaths by age. There is a first sharp peak at age 0, representing infant death, and then (as would be expected) a second, broader peak of deaths in higher age-groups. The distribution of fork length of yellowfin tuna taken by purse-seiners in the Pacific is also bi-modal.

If we examine the diagrams of distributions presented earlier in this topic, we find that they have shapes similar to the types described above. The distribution of crab weights is strongly skewed to the right; that for yellowfin taken by pole-and-line is much closer to a normal distribution, but still skewed a little to the right; and the distribution of villages by number of powered boats is bi-modal.

TOPIC 4 - DESCRIPTIVE STATISTICS : SUMMARISING THE OBSERVATIONS

4.1 Introduction

In the previous topic we saw how a number of observations of some variable could be summarised by forming a frequency distribution. This distribution will contain a lot of information about the variable, it will show how many high values there are, how many low ones, and by looking at the frequency histogram we can often get some idea of the distribution of this variable in the population. In many situations this is sufficient, but we often find that we need to reduce the amount of information in the frequency distribution even further. If we want to compare two distributions, for example, it can be difficult and confusing to have to look at all the information. In this topic we shall see how we can calculate one or two values that can be considered to represent some feature or property of the distribution. We can then use these values to make comparisons and to form the basis of more complex decisions.

As an example, consider Table 4.1 which shows two frequency distributions of fork length of yellowfin taken by purse-seine vessels.

TABLE 4.1 : COMPARISON OF TWO FREQUENCY DISTRIBUTIONS

Country A		Country B	
Fork length (cm)	Number	Fork length (cm)	Number
30-39	132	Less than 40	167
40-49	219	40-49	345
50-59	253	50-54	369
60-69	126	55-59	492
70-89	61	60-64	318
90-109	124	65-69	160
110-129	182	70-79	114
130+over	135	80-99	281
		100-119	294
Total	1232	120+over	203
		Total	2743

Using the data presented in Table 4.1, comparison is difficult. We have the same variable in each case, but different numbers of observations and different classes. What we need to do is to look at the distributions for the two countries and to find some way of describing certain characteristics of each one, which we can then compare quite easily. There are several different characteristics that we could choose, but in practice we tend to concentrate on just two: the <u>average</u> size and <u>dispersion</u>. We choose these because they have an obvious meaning and most people can understand them, and because in practice we find that they describe the whole distribution effectively. These two measures form the basis of almost all statistical inference, but we shall only be dealing with averages and dispersion as ways of describing or summarising a set of observations.

4.2 Some special notation and concepts

In this topic we shall be concerned with a number of observations of some variable and, as before, we shall only be dealing with one variable at a time. The observations may be grouped into a frequency distribution or they may be in their original state, but the principles in each case will be the same. In order to be able to make general statements that will be true about any set of data, however, we shall need to use some special statistical notation. We can use certain letters and symbols to stand for some items, and these usually will be the same as those introduced in the preliminary session to this course. There are, however, one or two new ideas that we must mention before we can go on to look at average and dispersion in detail.

We will need to distinguish between populations and samples because there will be some important differences. When we are dealing with the whole population we generally use letters from the Greek alphabet to denote values we calculate; in particular, we shall be using the letters μ (mu) and σ (sigma). For a sample, on the other hand, we use ordinary letters to represent values.

In practice we are usually interested in the population values of averages and measures of dispersion, rather than just the sample values. The population values are referred to as parameters of the population to distinguish them from values derived from samples. Very often we do not have information about a population, rather we have a series of values from a sample. What we do is to estimate the population parameters by calculating sample statistics.

4.3 Measures of average values

An average is a measure of the size of a set of variables and it forms the basis of a lot of more advanced statistical work. There are several different types of average that we can calculate, or find, and they have different properties; which one we use in any particular situation will depend upon what we want to do. We shall look at three types of average: the arithmetic mean, the median and the mode; these are the ones most commonly used, although there are other types for more specialised uses.

4.3.1 The arithmetic mean

The most common and widely understood type of average is the arithmetic mean. If we have a sample of values, x_1, $x_2 \ldots x_n$ of some variable x, then the arithmetic mean of this sample, which is denoted by $\bar{x}$ (pronounced 'x bar'), is given by:

$$\bar{x} = (x_1 + x_2 + \ldots + x_n)/n$$

This may be written as $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

As we noted earlier, the use of the subscript 'i' is a convention to indicate the observation under consideration. Thus, $\sum_{i=1}^{n}$ means the sum of all the observations, from the first to the nth; similarly $\sum_{i=2}^{4}$ means the sum of the second, third and fourth observations, and so on. For the rest of this course we will be referring to the sum of n observations. Usually, for the sake of simplicity, we will abbreviate this and just use the symbol Σ on its own, and will omit the subscript i. So we will write the formula for

the mean as simply $\bar{x} = \Sigma x/n$, and we understand that this is really a shorthand way of writing $\frac{1}{n}\sum_{i=1}^{n} x_i$.

We can express the same idea in words by saying that the arithmetic mean of a set of values is given by the sum of the values divided by the number of values in the set. The arithmetic mean is easy to calculate and will always exist for any set of values.

We use the symbol $\bar{x}$ to stand for the arithmetic mean of a sample, and we use the Greek letter μ (mu) to stand for the mean of a population. For a finite population we will have: $\mu = \frac{1}{N}\Sigma x$. Often we calculate $\bar{x}$ and use it to try to estimate μ. Obviously, if we have an infinite population it is impossible to calculate μ, and then there is no alternative but to use $\bar{x}$ to estimate μ.

Calculating the arithmetic mean of a frequency distribution

The mean which we discussed above is sometimes called a simple mean, and each value in the set is given the same weight, or importance. In the case of a frequency distribution, the mean must be obtained as a weighted mean.

We will first consider the simpler case of a discrete frequency distribution, and will use as an illustration the data on powered fishing boats per village, which we used in the previous topic. To calculate the mean number of powered boats per village, it is not correct to calculate a simple average of the different numbers of boats (0, 1, 2, etc.) shown in that distribution. There are many more villages with three boats than with one boat, for instance, and we have to take this into account.

In addition, we have to deal with the last group, '6 or more'. Since the frequency of this group is small, not much error will be introduced by the way we treat it; in this example we shall assume an average size of 7 boats per village for all units in this last class.

The arithmetic mean in this example is obtained as the sum of the products of the two columns of data, divided by the total number of observations, as follows.

No. of powered boats (x)	No. of villages (f)	Product (fx)
0	20	0
1	7	7
2	12	24
3	28	84
4	17	68
5	10	50
6 or more (Est. = 7)	4	28
Total	98	261

The arithmetic mean number of boats per village is then 261/98 = 2.66. We may note that, if we had gone back to the raw data and calculated the mean of the 98 individual observations we would have obtained almost exactly the same result. The only reason we have to say 'almost' exactly

is that we do not know the precise values for the 4 villages in the open-ended '6 or more' class.

Just as we had a mathematical formula for the arithmetic mean of a set of numbers, so we can derive a similar formula for use with a frequency distribution. In this case we call the number of classes 'k', the value for the ith class will be denoted by x_i and the frequency of each class by f_i. The formula for the mean $\bar{x}$ is then given by:

$$\bar{x} = \sum_{i=1}^{k} f_i x_i \; / \; \sum_{i=1}^{k} f_i$$

In words this says that the mean of the distribution is given by the sum of the frequency of each class multiplied by the value of the variable for that class, all divided by the total frequency.

We will abbreviate this formula by omitting reference to i and k, as

$$\bar{x} = \Sigma fx \; / \; \Sigma f$$

In the example above, $\Sigma fx = 261$ and Σf (which is equal to n) = 98. We may note that there are 7 classes in the frequency distribution, so k = 7.

When dealing with a continuous distribution we use the class mark as our values of x, as in the following example using the yellowfin data from the previous topic:

Class (kg)	Class Mark (x)	Frequency (f)	Frequency x Class Mark (fx)
2.0 - 2.9	2.45	7	17.15
3.0 - 3.9	3.45	19	65.55
4.0 - 4.9	4.45	16	71.20
5.0 - 5.9	5.45	12	65.40
6.0 - 6.9	6.45	6	38.70
7.0 - 7.9	7.45	3	22.35
Total		$\Sigma f = 63$	$\Sigma fx = 280.35$

$$\bar{x} = \Sigma fx / \Sigma f = 280.35/63 = 4.45$$

The arithmetic mean in this case is obtained as the weighted mean of the class marks or midpoints of the class intervals, the weights being the frequencies, f_i, or relative frequencies, $f_i / \Sigma f_i$. What we have done is to assume that all the units in a class interval have the corresponding midpoint as their value.

It follows then that we cannot expect the arithmetic mean we have calculated to be exactly the same as the mean we would obtain by going back to the individual raw data. In fact if we make the calculation of the arithmetic mean from the 63 individual observations of yellowfin weights, we find that $\bar{x} = 278.9/63 = 4.43$.

The arithmetic mean has a lot of advantages as an average: it is easy to calculate, most people understand it, and it is easy to use in more

advanced statistical work. It does, however, also have some disadvantages which can produce difficulties in some situations. The value of the arithmetic mean can be quite affected by one or two large observations, especially in a small sample; this can happen when we have a non-normal distribution. In this kind of situation, using the arithmetic mean may be misleading.

For instance, in distributions of income, which are strongly skewed to the right, it is not unusual for the incomes of a few very rich people to be so large that they pull the arithmetic mean to a higher level than is earned by the great majority of people.

In similar fashion, the arithmetic mean of a bi-modal distribution quite often falls between the two peaks, and is therefore not a good representation of the distribution.

The other difficulty that can arise with the arithmetic mean is that we can obtain a value that obviously does not exist. There is no problem with continuous data: it is easy enough to envisage a mean weight of 4.45 kg, for example. However, for discrete data the situation is different. We calculated the mean number of boats per village as 2.6. Obviously we cannot have 0.6 of a boat, and many non-statisticians find this kind of answer difficult to understand. We could round the answer to the nearest whole number, in this case 3, but we lose a lot of information if we do. What we have to realise is that the arithmetic mean is an artificial concept; we use it because it is convenient, not because it has any natural meaning. If we found that the mean number of boats had been 2.2 in 1980 and is 2.6 in 1984, we could draw some conclusions about trends. We can use the mean to make useful comparisons, but we must not assume that the mean value must actually exist.

4.3.2 The median

The median is a very simple concept which can be quite useful in practice, although it is difficult to deal with mathematically. It is the value of the middle observation of a set of numbers; half the numbers will be larger than this value and half will be less. For data in the raw form all we have to do is to rank the observations in order of size, and the median will be the value of the middle one. If we have n observations, the median will be the value of the $\frac{n+1}{2}$ th observation.

If n is odd there is no problem, but if n is even there are two middle observations; in this case we take the median to be the arithmetic mean of the two values. As an example, consider the following three sets of observations, which have been sorted into size order.

(a) 19, 22, 26, 31, 34, 37, 42, 44, 49, 55, 63

(b) 12, 19, 23, 27, 30, 30, 47, 49, 60, 93

(c) 128, 186, 193, 207, 218, 222, 286, 346

In set (a) there are 11 observations; the median is given by the 6th one, so is equal to 37. In set (b) there are 10 observations; the median is given by the mean of the 5th and 6th ones, and as each of these is equal to 30, there is no problem in obtaining 30 as the median value. In set (c) there are 8 observations so the median value is the arithmetic mean of the 4th and 5th observations, i.e. the median = (207+218)/2 = 212.5.

To calculate the median from a frequency distribution we use the same principle, but we start by determining the class within which the median value lies. If this class contains a single value, then there is no problem. If, however, it contains a range of values, then we have to estimate the median value, using simple interpolation. For the second case, because we are dealing with a range of values, we use the formula n/2, not (n+1)/2 to determine the median observation. This is illustrated in Table 4.2.

TABLE 4.2 : CALCULATION OF THE MEDIAN FOR A FREQUENCY DISTRIBUTION

(I) Using our powered fishing boats example:

No. of boats	Frequency	Cumulative frequency
0	20	20
1	7	27
2	12	39
3	28	67
4	17	84
5	10	94
6 or more	4	98

The median value is given by the (98+1)/2th observation, i.e. by the mean of the 49th and 50th. These lie in the class "3 boats per village", so the median is 3.

(II) Using our yellowfin weights example:

Class (kg)	Frequency	Cumulative frequency
2.0 - 2.9	7	7
3.0 - 3.9	19	26
4.0 - 4.9	16	42
5.0 - 5.9	12	54
6.0 - 6.9	6	60
7.0 - 7.9	3	63

In a frequency distribution, the median position is n/2, not (n+1)/2 as in the case of an array. With 63 observations, the median position is the 63/2th or 31.5th. From the cumulative frequency column we see that the 31.5th position falls in the class 4.0 - 4.9 kg with actual class limits 3.95 - 4.95. There are 26 observations prior to the interval beginning 3.95, and 16 in this interval, so we calculate the median weight as 3.95+(31.5-26)/16x1.0. Thus the median is equal to 4.3 kg.

Once again we must recognise that (as for the arithmetic mean) the calculation of the median which we obtain from a frequency distribution is only approximately the same as we obtain from a list of all the individual observations. Indeed if we refer back to our original 63 yellowfin weights and sort them into order, we find that the (63+1)/2th (or 32nd) value is 4.1 kg.

Another way to determine the median is directly from the ogive of the frequency distribution. In this case, though, the accuracy of the median is determined by the accuracy with which the graph is plotted. We draw a horizontal line from the y-axis at the position corresponding to the median observation, and from the point where this line cuts the ogive, we draw a vertical line. The point where this vertical line meets the x-axis gives us our reading of the median value.

If the ogive of our example is accurately plotted, as in Figure 4.1 below, we should be able to observe that the median (i.e. the 31.5th) value is 4.3 kg, which is the figure we have already calculated from the frequency distribution.

4.3.3 Quartiles

The median is the value of that observation which divides the total frequency into two equal parts. In the same way we can determine other values which divide the frequency into other fractions. The most important of these are called the quartiles. Quartiles, as their name suggests, divide the total frequency into four equal parts.

The first, or lower, quartile will then have one quarter of the observations less than this point and three quarters greater. The middle quartile is equivalent to the median. The upper quartile has three quarters of the observations less than this value and one quarter more.

In an array, the lower quartile is the (n+1)/4th value, the middle quartile (which corresponds with the median) is the (n+1)/2th value, and the upper quartile is the 3(n+1)/4th value. For the 63 observations of yellowfin weights we can see that the lower quartile is the 16th value and the upper quartile the 48th. A study of the individual weights, sorted into order, will show that the quartile values are 3.3 kg and 5.4 kg respectively.

Whenever the quartile lies between two values, the value of the quartile is calculated by interpolation as in the case of the median.

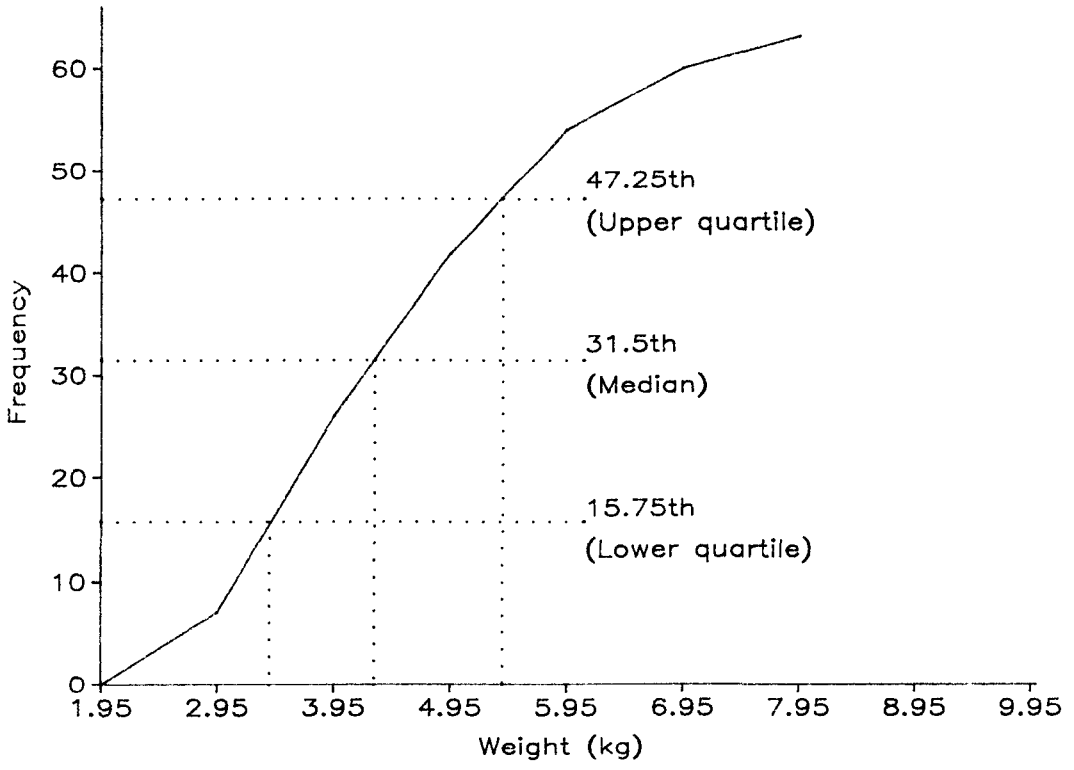
A similar method is used even for a frequency distribution, interpolating (as in the case of the median) within the quartile class. However, the quartile positions become n/4th, n/2th and 3n/4th. In our example Q_L is the 63/4th or 15.75th position. This falls in the 3.0 - 3.9 class, being the 15.75-7 or 8.75th item into this class. Therefore $Q_L = 2.95+(15.75-7)/19 = 3.41$, or 3.4. A similar calculation will give us a value for the upper quartile of 5.4 kg. These values can be compared with those obtained from the raw data, as shown above.

Later in this topic we will use the quartiles to determine a measure of the spread or dispersion among the observations.

Apart from splitting the total frequency into quarters we could use other fractions, or percentages, and we associate the word percentiles with these. There are also special names for certain commonly used fractions, e.g. deciles (which split the total frequency into 10 groups) and quintiles (into 5 groups). We should note here that these are not averages, since they do not give us a central value to represent the distribution, but it is convenient to discuss them at the same time as we discuss the median, which is an average. All of these measures can be read

directly from an ogive, and we illustrated that in Figure 3.6. The derivation of the median and quartiles for the yellowfin data is shown in Figure 4.1.

FIGURE 4.1 : CUMULATIVE FREQUENCY OF YELLOWFIN WEIGHTS SHOWING POSITION OF THE QUARTILES



The median and other percentiles are very useful when we want to describe what is happening in certain types of distribution. Quite often, for example, we would like to find out about part of a distribution, the poorest 20 per cent of fishermen, the largest 20 per cent of skipjack, and so on. By calculating percentiles we can do this, and we can compare values between distributions.

We also find that for skewed distributions (ones that are not almost symmetrical), the median is often a better measure of the average value than the mean. Figure 4.2 illustrates this; it shows the position of the mean and the median for a symmetrical and a skewed distribution. For a distribution that is almost symmetrical, the value of the median and the mean are very similar, and both are good measure of the average. For a distribution skewed to the right, however, the mean will be to the right of the median. In the sense of actually representing the data the median may be more useful; it is more stable, is not affected by the inclusion of a few very large values, and hence is probably better to use for the purposes of comparison.

4.3.4 The mode

If a population distribution has a peak in its distribution function at a certain point then there is said to be a mode at that point. Like the arithmetic mean and the median, the mode is a type of average.

When dealing with sample observations the concept of the mode is most useful in connection with frequency distributions. For a discrete distribution the mode is that value which occurs most often. For instance, in our earlier illustration of the number of powered boats, the modal value is 3 boats per village, because more villages had 3 boats than any other number. We may consider that this is a more useful summary of our information than is the arithmetic mean of 2.66 boats. It is interesting to note that in this case the median value is also 3.

For a continuous distribution the determination of the mode is rather complicated, and so for our purposes we shall be concerned only with the modal group or class. This is the class with the highest frequency (that is, 3.0 to 3.9 kg in the yellowfin data, for example), in a distribution which has equal class intervals.

However, determining the modal group for continuous frequency distributions, particularly where the distribution is a sample from a population, often produces problems. The modal group will very largely depend on how the classes are defined, and for data with a fairly even distribution between classes, a change in the definition of the classes can change the modal group. The smaller the sample, the more likely this is to happen. For instance, if we had grouped the yellowfin data into 3 classes, namely, 2.0-3.9 kg, 4.0-5.9 kg and 6.0-7.9 kg, we would find that the class with the highest frequency is 4.0-5.9 kg, so we would have quite a different modal class. Various different groupings, e.g. into a large number of smaller classes, would give us different results again. For this reason, the mode is of limited value, and should be used with care.

Diagrams illustrating the relationship between the arithmetic mean, the median and the mode for the most common distributions are given in Figure 4.2.

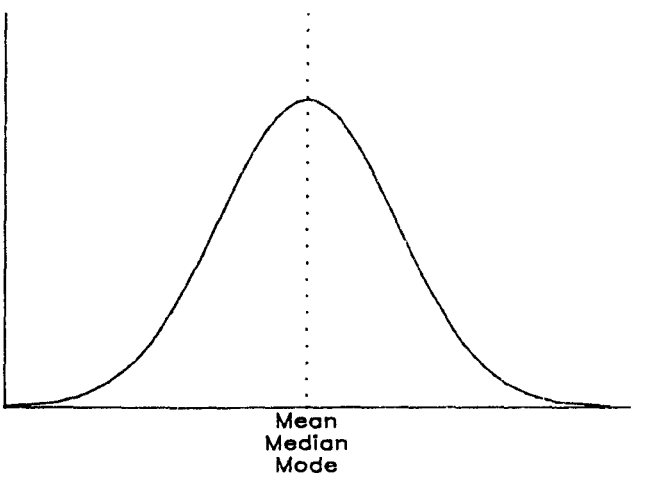
4.3.5 Summary of the different types of average

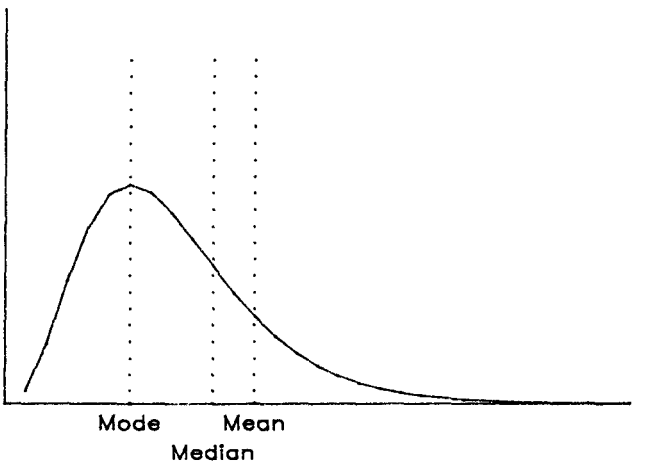
We have looked at three main types of average: the arithmetic mean, the median and the mode. All three of these have both advantages and disadvantages when used to describe or summarise a set of data. The mean is the most widely known and most widely used, but can be misleading when dealing with skewed distributions. In this situation the median, and various percentiles can be more useful, particularly when making comparisons between distributions. The mode has limited value, and should not be used with small samples.

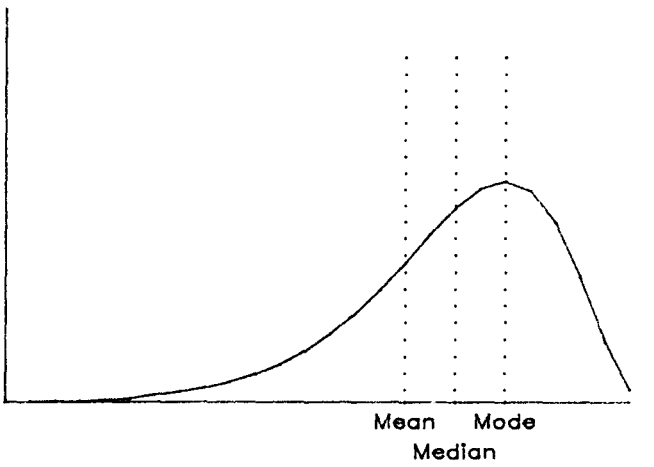
When we come to the problems of statistical inference, however, we almost always use the arithmetic mean. The reason for this is purely mathematical convenience. It is much easier to deal with the mean to derive more complex results; the ways the median and the mode are defined make these much more difficult to use. We tend, therefore, to concentrate on the mean just because this helps us when we want to study more complicated areas of statistics.

There are also other types of average, which we will not discuss in this course. The best known are the geometric mean (which is most useful for measuring rates of change) and the harmonic mean.

FIGURE 4.2 : THE RELATIONSHIP BETWEEN THE ARITHMETIC MEAN, THE MEDIAN AND THE MODE







4.4 Measures of dispersion

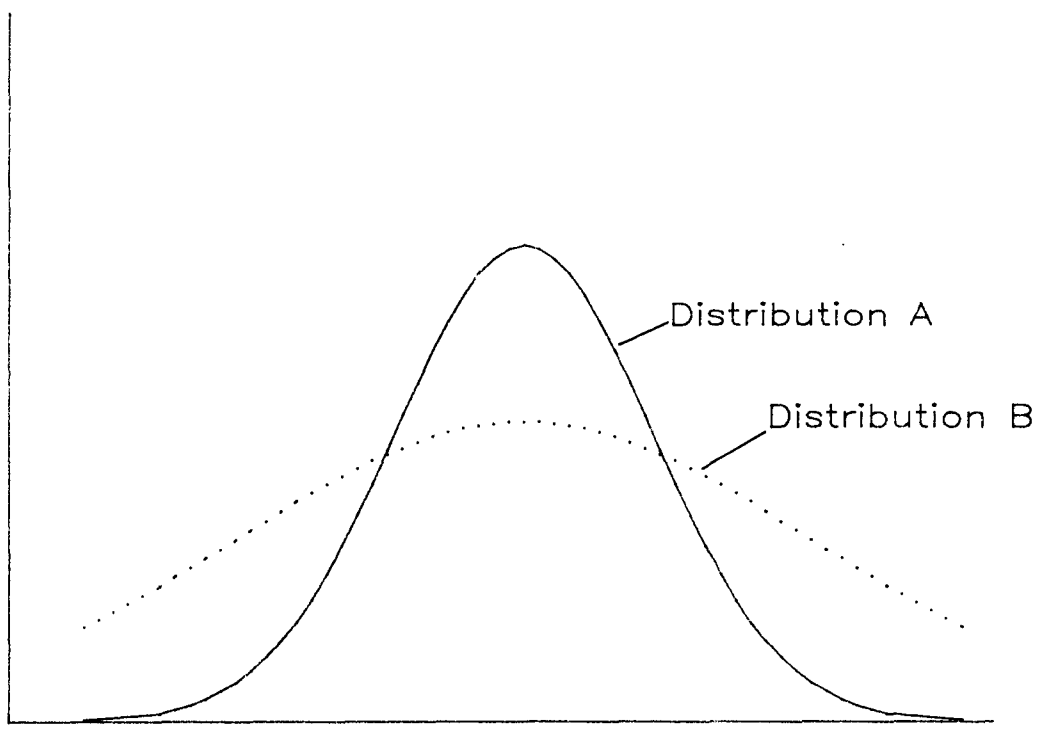
4.4.1 Basic principles

In the two previous sections we have discussed ways in which we can summarise statistical data, and can present it in a straightforward way which will be fairly readily understood. The frequency distribution is a method to summarise information in tabular or graphical form, while an average (such as the arithmetic mean) summarises this information into one single number.

We must recognise, however, that while in summarising we are attempting to distil from a mass of data the essential features which need to be highlighted, in so doing we are always losing some of the information. We have to be very careful that in the process we do not go too far, and leave out of our summary information which is necessary for a proper understanding of the situation. We will see in this section that an average, on its own, is often insufficient to describe a population adequately. In particular when we are endeavouring to compare the characteristics of different populations, some further measure in addition to the average is usually required.

Consider, for example, Figure 4.3 where there are two distributions shown. Both have the same average value, whether measured as a mean, a median or a mode, but we could not say that the distributions were the same. To describe and compare them we need additional information; we need alternative ways of describing the distributions. From the diagram we can see that distribution B is much more spread out than distribution A; in this section we shall look at different ways of measuring this spread, or dispersion.

FIGURE 4.3 : COMPARISON OF TWO DISTRIBUTIONS



We want to measure dispersion for two main reasons. In the first place we may well be interested in the actual level of dispersion and in comparing this with another distribution. The second reason for wanting to measure dispersion is that, even when we only want to compare average values, we still need to take variability into account. We want to be able to distinguish between differences that might have just happened by chance and those that indicate some real change.

In this topic we shall consider four different measures of dispersion, which are basically of two types:

(a) measures of the distance between certain representative values of the population; and

(b) measures of the deviation of every member of the population from some specified central value.

As examples of the first type of measure of dispersion we shall look at the range and the interquartile range, while under the second type we shall consider the mean deviation and the standard deviation (or the square of this, the variance).

4.4.2 Measure of the distance between selected points of the distribution

The most obvious way of measuring the dispersion in a set of observations is to calculate the range, which is just the difference between the smallest and the largest values. This is simple to understand and easy to calculate and so has an obvious appeal. It is used in practice, but is only really useful when the variable under consideration has a fairly even type of distribution over the range. It has some obvious drawbacks which tend to restrict its use in practice; some of the more important disadvantages are:

(a) Because the range is the difference between the smallest and the largest values, it is very sensitive to very large or very small observations; the inclusion of just one freak value will affect the range.

(b) The range depends on the number of observations. Increasing the number of observations can only increase the range; it can never make it less. This means that it is difficult to compare ranges for two distributions with different numbers of observations.

(c) The range ignores most of the observations; for example, the following sets of data all have the same range, even though we can see that the degree of dispersion is different.

(i) 3, 5, 7, 9, 11, 13, 15, 17
(ii) 3, 3, 3, 3, 17, 17, 17, 17
(iii) 3, 3, 3, 3, 3, 3, 3, 17

(d) It is difficult to calculate the range for data grouped in a frequency distribution. All we can really do is take the difference between the lower limit of the first class and the upper limit of the last class. This will obviously depend on our definitions of the classes, and is impossible if we have an open-ended class.

We can get round most of the disadvantages of the range as a measure of dispersion by using other points in the distribution rather than the two extremes. An obvious choice would be to measure the inter-quartile range - the difference between the upper and lower quartiles. Another alternative would be to use the difference between the 10th and the 90th percentile. As measures of dispersion, both these are quite useful. They are not affected by one or two wild observations, they are less dependent on the number of observations, and they will tend to differentiate between different sets of observations. In the case of frequency distributions, we can nearly always calculate these distances, the only problem being when one of the percentiles or quartiles falls in an open-ended class.

The inter-quartile range in particular is a fairly good measure of dispersion, that is reasonably easy to calculate and which most people find fairly simple to understand. It can be used to measure the amount of dispersion and to make simple comparisons between distributions. Quartiles are far enough from the ends of the distribution to make it extremely unlikely that they will fall within an open-ended class. In fact, if there is an open-ended class in a distribution, the inter-quartile range will probably be the only one of our four measures of dispersion which we can calculate accurately. All the other measures will require some assumption to be made about the open-ended class.

In practice, the quartile deviation is often quoted; this is defined as one half of the inter-quartile range and provides a result that is comparable with other measures of dispersion. The major drawback, however, comes when we want to undertake more advanced statistical work. It is difficult to deal mathematically with quartiles, so in practice we tend to concentrate on other measures of dispersion.

4.4.3 Measures of deviation from a specified central value

With this type of measure of dispersion we use every value in the distribution and find the average distance between every observation and some central point. In theory, we could use any central point we like, the median, the mode, or whatever, but in practice we use the arithmetic mean for reasons of mathematical convenience. What we need to do, then, is to find the difference between each observation and the mean, and then calculate the average of these distances. There is, however, one immediate problem which we can illustrate with the following simple set of data:

3, 5, 7, 9, 11, 13

The arithmetic mean is (3+5+7+9+11+13)/6=8 and the differences, or dispersions, of each observation from the mean are:

-5, -3, -1, +1, +3, +5.

The total dispersion is zero, and in fact this will always be true. Because of the way we define the mean, the total dispersion of all the observations from that value will always be zero; it is a check on the accuracy of our calculations. We cannot, therefore, use the values exactly as they are.

What we are interested in, in fact, is the actual size of the dispersion, regardless of the sign, and so one possible solution would be to take the average value disregarding all signs. In our simple example, then, our total dispersion would be:

5+3+1+1+3+5=18, and, since we have 6 observations,
the average will be 18/6=3.

This is a good measure of dispersion, and we call it the mean deviation; it is the average deviation of all observations from the mean, disregarding all signs. For a general sample, $x_1\ x_2\ \ldots x_n$ we can write the formula for mean deviation as

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \qquad \text{or, more simply} \qquad \frac{1}{n}\Sigma\,|x - \bar{x}|$$

The symbol $|\ldots|$ stands for modulus, or mod for short, and it means - take the absolute value, ignore all signs.

Although the mean deviation is a good measure of dispersion, and one that most people find quite easy to understand, we do not use it much. The reason for this is that it is difficult to manipulate the modulus of a number mathematically, which means that the mean deviation cannot be easily used in more advanced statistical work.

4.4.4 Standard deviation

Instead, we calculate the standard deviation, which we obtain as follows:

As before, we work with the deviations from the arithmetic mean; in our example we had:

-5, -3, -1, +1, +3, +5.

In this case we square these deviations, which will give us the following:

25, 9, 1, 1, 9, 25.

All these numbers are positive. We now take the average of these squares, i.e.

(25+9+1+1+9+25)/6 = 70/6 = 11.67

Since we have squared all the deviations, we should return to the magnitude of the original units, and so we take the square root of this result, i.e.

standard deviation = $\sqrt{11.67} = 3.4$

If we are concerned with a population, we use the symbol σ (sigma) to stand for the standard deviation, and in general terms σ is given by:

$$\sigma = \sqrt{\sum_{i=1}^{N}\frac{(x_i - \mu)^2}{N}} \qquad \text{which we will simplify to:} \quad \sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

where μ, of course, is the mean of the population.

For a sample, the situation is a little different; we can use the symbol s to stand for the sample standard deviation and this is given by:

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Here, we use the divisor n-1, where for a population we use N. If the sample size is large, this will make little difference. The reason that we use n-1 is that, in this case, s provides an unbiased estimate of σ. In other words, if we took many different samples and calculated s for each one, the mean of these values would approach the population value. This would not be true if we used the divisor n.

The square of the standard deviation is called the variance, and we sometimes use this to avoid having to take square roots. The population variance is denoted by σ^2, and equals $\frac{\Sigma(x-\mu)^2}{N}$, while the sample variance, s^2 equals $\frac{\Sigma(x-\bar{x})^2}{n-1}$.

As it stands, it is quite a cumbersome procedure to calculate the standard deviation of a large set of numbers. First of all we have to determine the mean of the set, then calculate the deviations of each observation from the mean, square these, add them up and take the square root of the result. Even with a calculator this will require each value to be entered twice, and can take some time.

We can, however, make the calculation much easier by rearranging the formula for the variance. For a sample we have:

$$S = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - \frac{1}{n}(\Sigma x)^2\right)}$$

and for a population

$$\sigma = \sqrt{\frac{1}{N}\left(\Sigma x^2 - \frac{1}{N}(\Sigma x)^2\right)}$$

Although this second formula looks more complicated than the first, it is in fact much easier to use with a calculator. We can observe that in this formula we do not have to start by calculating the arithmetic mean, so we can save one step in the calculation process. Using the memory function on the calculator, we can now calculate the standard deviation without having to write down any intermediate results.

We can use the second version of the formula to give us a fairly simple method for calculating the standard deviation of a frequency distribution. We shall use as an example the yellowfin data to illustrate this. The relevant calculations are as follows:

Weights (kg)	Class Mark (x)	Frequency (f)	fx	fx^2
2.0 - 2.9	2.45	7	17.15	42.02
3.0 - 3.9	3.45	19	65.55	226.15
4.0 - 4.9	4.45	16	71.20	316.84
5.0 - 5.9	5.45	12	65.40	356.43
6.0 - 6.9	6.45	6	38.70	249.61
7.0 - 7.9	7.45	3	22.35	166.51
Total		63	280.35	1357.56

The standard deviation, s $= \sqrt{\frac{1}{\Sigma f-1}\left(\Sigma fx^2 - \frac{1}{\Sigma f}(\Sigma fx)^2\right)}$

$$= \sqrt{\frac{1}{62}(1357.56 - 1247.56)}$$

$$= 1.33$$

We use $\Sigma f-1$ which is another way of writing n-1, as the denominator, because the data are from a sample.

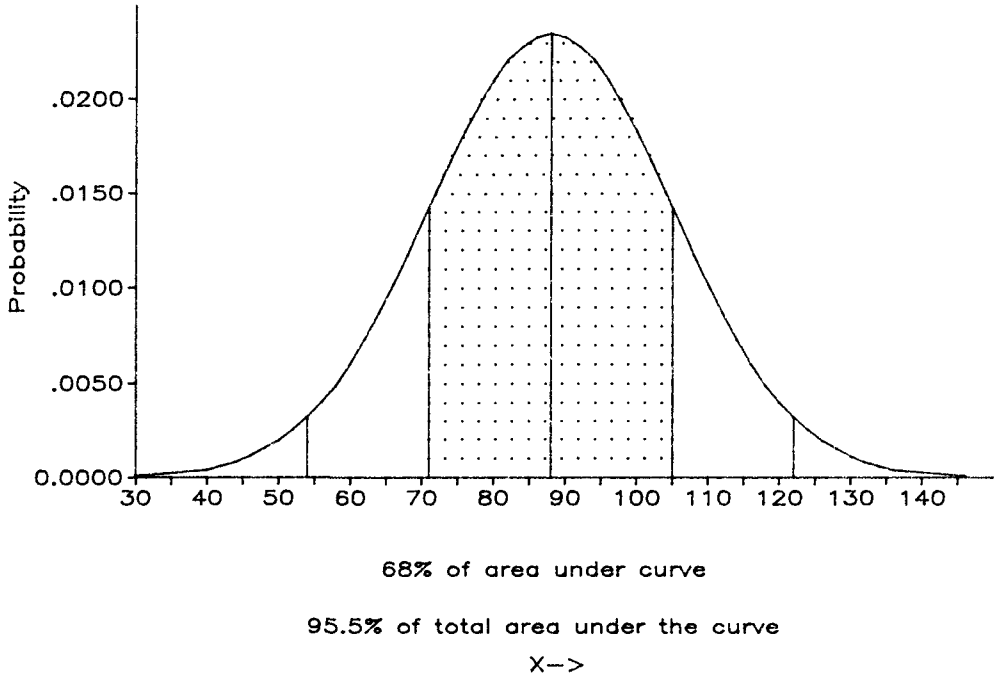
If we wanted to write out the formula for the variance of a frequency distribution in full, it would be:

$$s^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1} \left(\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{\sum_{i=1}^{k} f_i} \left(\sum_{i=1}^{k} f_i x_i \right)^2 \right)$$

in the case of a sample.

One interesting feature of the standard deviation in respect to the normal distribution may be mentioned here. If the population is distributed normally about the mean, then approximately 68 per cent of all values will lie within one standard deviation of the mean, and about 95.5 per cent will lie within 2σ. Thus, for a normal distribution with a mean of 88 and a standard deviation of 17, about 68 per cent of all values will lie in the range 88-17 to 88+17 (i.e. between 71 and 105) and 95.5 per cent will be within the range 54 to 122. We can demonstrate this graphically, as shown in Figure 4.4, by plotting the normal curve, and vertical lines drawn from the x-axis at values 71 and 105 would enclose 68 per cent of the total area under the curve. This particular property will hold true for a normal distribution, no matter how widely spread the values are. It will be very useful in our understanding of standard errors, which will be discussed later in the course.

FIGURE 4.4 : NORMAL PROBABILITY DISTRIBUTION, MEAN 88, S.D. 17



The standard deviation is by far the most widely used of the four measures of dispersion. As we will see it is also used in the calculation of sampling errors. Although it is so widely used, this does not mean that it is superior in every respect. Its main weakness is that it is very

greatly affected by extreme values, much more so than is the mean deviation. This occurs because the deviations from the mean (which are already large in the case of extreme values) become very large indeed when they are squared, as they are in the calculation of the standard deviation.

4.4.5 Summary of the different measures of dispersion

There are two ways we can measure the degree of dispersion in a set of observations: we can look either at the difference between two points in the distribution or at the average deviation of all the observations from some central point. Examples of the first type are the range and the quartile deviation. These are fairly easy to calculate, have an obvious meaning, but ignore quite a large part of the data. The range, in particular, is unstable and is affected by wild observations; the quartile deviation is more stable and better to use in practice. Both measures are difficult to deal with mathematically. Examples of the second type of measure are the mean deviation and the standard deviation. In practice we use the standard deviation because it is, mathematically, more convenient.

Average value and dispersion are not the only properties of distributions we can measure, but we generally concentrate on just these two. The reason for this is that, for a large class of symmetrical or almost symmetrical distributions with a single mode, we can fit a normal distribution quite easily. This will allow us to make many important inferences concerning the data. One very important property of a normal distribution is that the only things we need to know are the mean and the variance (or standard deviation). Once these are determined, the distribution is fixed. So, to fit a normal distribution to a set of data, all we have to do is to calculate its mean and variance.

TOPIC 5 - RELATIONSHIPS : LINKS BETWEEN TWO OR MORE VARIABLES

5.1 Introduction

In the previous two topics, we concentrated entirely on distributions and measures of one variable; but in reality, we normally collect data on several items at once. We are interested in links, or relationships, between the different variables (or, sometimes, between variables and attributes).

For example, the fish catch by a local fishery would be affected by many factors, which may include the following:

- the number of people and boats engaged in fishing
- fishing technique used
- weather conditions prevailing
- the surface temperature of the water
- the effect of other fisheries operating nearby.

No doubt many other items could be added to the list. The mathematics of trying to measure the interrelationships of all of these factors would be very complicated. This is referred to as 'multivariate' analysis, and is beyond the scope of the present course.

We can, however, study the relationship between two variables. Data on two variables are termed bivariate data, and if these are plotted as points on a graph, with one variable on each axis, we have what is known as a scatter diagram. We introduced this briefly in Topic 2. If we look back to Figure 2.1, we see that it demonstrated a relationship between fish catch and number of boats engaged. Similarly, Figure 2.2 showed the relationship between total fish catch and time. In fact in that diagram we drew lines to link up the points on the graph, but we need not have done so. If we had omitted the lines, and plotted only the points, we would have had a scatter diagram of exactly the same type as Figure 2.1.

5.2 Regression

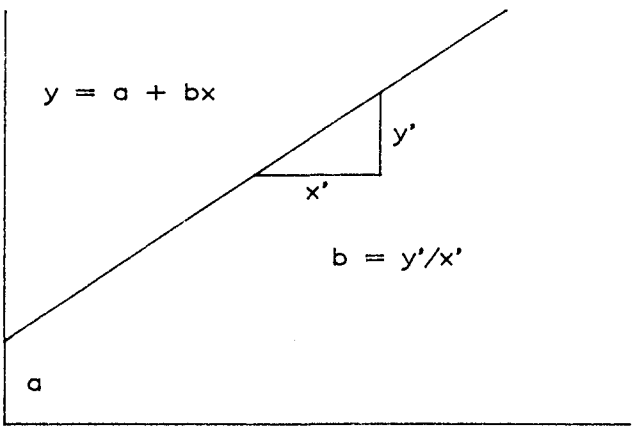
Finding a mathematical formula to describe the relationship

So the purpose in drawing a scatter diagram is to try and get some idea of a simple relationship between two variables. We are not trying to find some mathematical formula that will go through all the points exactly. It is theoretically possible to do this, but the formula would be too complicated to be of practical use. What we would like is some kind of simple formula that 'fits' or describes the data fairly well. If we can do this, then we have some kind of model that tells us something about the underlying process that produced the data and can help us to make predictions or other decisions. Now with two variables, x and y, the simplest kind of relationship between them is shown on a graph as a straight line. This means that if we increase 'x' by a constant amount then 'y' will also increase by a fixed amount.

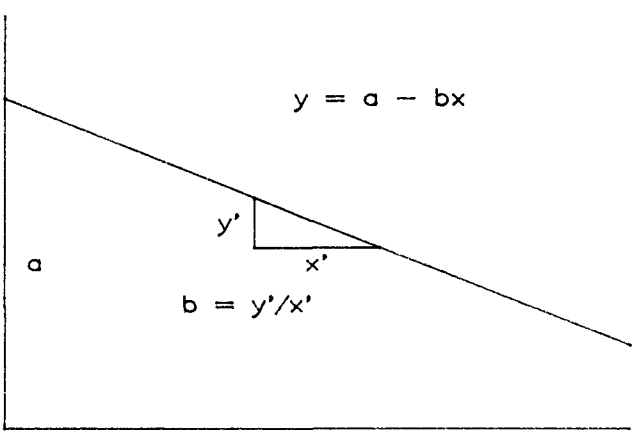
Mathematically, we can represent a straight line by the equation, or formula, $y = a+bx$; a and b are constants where 'a' represents the point at which the line meets the y-axis, and 'b' represents the slope of the line. This is shown in Figure 5.1. By changing the value of a and b we change the position of the line on the graph. If the line goes from the bottom

left hand corner to the top right hand corner, then the slope b will be positive. If it goes the opposite way, from the top left hand corner to the bottom right, then the slope will be negative, and b will be a negative number.

FIGURE 5.1 : THE EQUATION OF A STRAIGHT LINE



Positive slope



Negative slope

When we have a scatter diagram, what we want to do is to find a line which best 'fits' the data, that is, which is closest, in some sense, to all the various data points. This means, effectively, to find values for a and b, since it is these two values that define the line. We can undertake this process by eye. Using a scatter graph and a transparent ruler we can move it until it appears to be the best 'fit' to the data, but this is rather unscientific. We have no guarantee that two different people will produce the same line for the same data. Their ideas of the line of 'best fit' may be rather different, and so it will be very difficult to generalise. Instead of using this method then, we use a mathematical technique in which a and b are calculated from the data values (x_i, y_i).

Even when we try to develop a mathematical technique, there are some problems which really arise from the basic situation we are dealing with. In Topic 2, we introduced the concept of independent and dependent variables. We saw that we often had the situation where we were interested

in looking at the way one variable changed as the value of some other variable was altered. So in Figure 2.2, we saw how the level of fish catch changed over time. In this example we say that total catch 'depends' upon time, and therefore total catch, being the dependent variable, must be plotted on the y-axis. In similar fashion, in Figure 2.1, fish catch was the dependent variable in its relationship to number of boats engaged. It is this type of relationship, in which one variable is dependent on another, that we are concerned with when we try to find a line that best fits the data. The purpose of finding the equation of the line, of estimating the values a and b, is that we can then estimate different values of y, given the appropriate values of x.

There are many criteria by which we might define what we mean by 'best fit', but the generally accepted method is the <u>least squares</u> criterion. By this we mean that we will seek to establish the formula which expresses y in terms of x in such a way that the sum of the squares of the differences between the observed values of y, and the values calculated by the formula, is as small as possible.

This technique, of estimating the values of an equation of a line, is known as <u>regression</u>. The line y=a+bx is called the 'regression line' and the values 'a' and 'b' are called the 'regression coefficients'. The equation of the regression line is the formula we shall use to predict the values of y we are likely to get, given certain values of x. We also call this the 'regression of y on x'; y is the dependent variable and x the independent variable. In our first example in the previous paragraph, then, we can talk about the regression of fish catch on time. We want to be able to find out what level of catch we might expect at some future time, by use of a mathematical relationship.

Our data consist of a series of pairs of values, x_i and y_i. We calculate our coefficients, a and b, from these observations. We have:

$$b = \left(\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \right) \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2$$

or, in shorthand form:

$$b = \frac{\Sigma (x - \bar{x})\,(y - \bar{y})}{\Sigma (x - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

We shall look at the expression $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ again in more detail in the next section. We notice that the denominator of b is $\sum_{i=1}^{n}(x_i-\bar{x})^2$ and that this also appears in our expression to calculate the standard deviation, or the variance, of x.

It will be recalled that we saw how we could rearrange the formula for the variance to make it easier to find using a calculator. So it is not really surprising to find that we can do the same thing for our expression for 'b'. The alternative formula may look more cumbersome at first sight, but is much more convenient for use, especially with a calculator, so we will always use it from now on. The formula is:

$$b = \frac{\Sigma xy - \dfrac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}$$

For calculating 'a' we retain the expression given above

$$a = \bar{y} - b\bar{x}$$

It will be seen that the formula for the straight line which best fits the data can be calculated if we work out values for Σx, Σy, Σx^2, Σxy and n.

We should make a cautionary note here that although this line is the 'best fit' for our data, according to the criteria we used, this does not mean that it is a perfect fit, or even that it is a good one. There will always be one straight line which fits the data better than any other straight line does, but whether this fit is a good one or a bad one, will depend on how scattered the series of paired observations were. Later in this topic we will develop a measure which will show us how well the line actually corresponds with the data.

Regression analysis is very extensively used in practice in estimating relationships between economic variables, such as demand and supply curves, relationships between income and expenditure, and so on. It also should prove very valuable in the analysis of fisheries statistics. Furthermore, it has great use in time series analysis, whereby we fit a straight trend line to data which is available, in order to estimate data which are missing, and most importantly in order to project forward to make estimates for future periods.

Let us look at one or two practical applications. We will be able to see how the regression line is actually calculated, and how it can be used to make estimates or forecasts.

First, let us revert to our data on boats used (which we denote x) and catch obtained (y) in the artisanal fishery (Table 5.1). We will calculate Σx, Σy, Σx^2 and Σxy, which we will need in order to calculate the coefficient 'b'. We already know that n = 10. We will also calculate Σy^2, which, although not used to calculate the regression coefficients, will be required for another calculation a little later.

TABLE 5.1 : CALCULATION OF THE REGRESSION COEFFICIENTS FOR DATA ON BOATS OPERATING AND CATCH OBTAINED

x	y	x^2	xy	y^2
12	590	144	7080	348100
15	820	225	12300	672400
10	330	100	3300	108900
12	740	144	8880	547600
18	900	324	16200	810000
14	660	196	9240	435600
6	240	36	1440	57600
15	650	225	9750	422500
16	850	256	13600	722500
9	470	81	4230	220900
127	6250	1731	86020	4346100

$\bar{x} = 127/10 = 12.7$ $\quad$ $\bar{y} = 6250/10 = 525$

Now we can substitute in our formula.

$$b = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}$$

$$= \frac{86020 - \frac{127 \times 6250}{10}}{1731 - \frac{(127)^2}{10}}$$

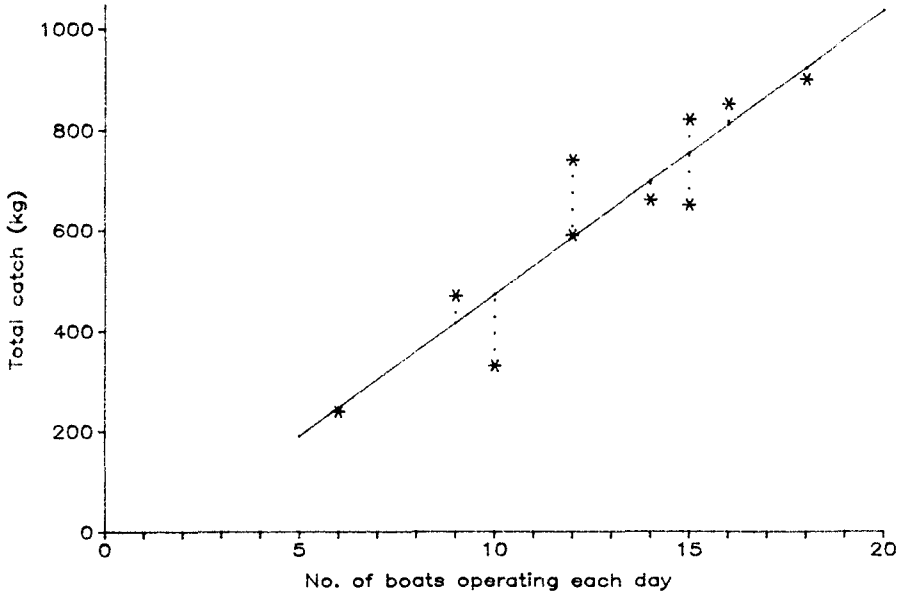
$$= \frac{86020 - 79375}{1731 - 1612.9} = 56.3$$

$$a = 625 - (56.3 \times 12.7) = -90$$

Therefore our regression formula is y=-90+56.3x.

In Figure 5.2, the regression line has been drawn in, and the vertical (y) deviation of each point on the scatter diagram from the regression line is also marked. What we have achieved in calculating the best fitting straight line to the data is to ensure mathematically that the sum of the squares of these 'y' deviations is the minimum possible. For any other line which we try to draw to fit the data, the sum of the squares of these deviations would be greater than for our regression line.

FIGURE 5.2 : CATCH PER DAY BY NUMBER OF BOATS OPERATING SHOWING LINE OF BEST FIT AND DEVIATIONS FROM THE LINE Y=-90 + 56.3 X



If now we want to estimate how much fish would be caught on a day when 14 boats are operating, we simply substitute 14 for x, and we have

$$y = -90 + (56.3 \times 14) = 698$$

So our equation estimates that 698 kg of fish would be caught. It is interesting to note from the actual data that there were 14 boats operating on one day, and the catch was 660 kg, which is very close to the estimate provided by our regression line.

We should also note that the regression line does not provide us with a good estimate for days when very few boats operate. For example, if there is only one boat operating, the equation predicts that the catch would be -90+(56.3x1) = -33.7 kg, which is obviously nonsense. Our original data did not contain any observations for very small numbers of boats operating, so it is perhaps not so surprising that the equation is not good for making estimates when that situation arises.

Another way of interpreting the equation is to note that 'b', the coefficient of the slope of the line, equals 56.3. That means that the regression line estimates that the total daily catch will increase by 56.3 kg for every extra boat engaged. It tells us in effect: multiply number of boats by 56.3 kg, but then deduct 90 kg from the estimate this gives.

For our second illustration, we will look at a time series of annual data. We will use data on fish catch for Country ABC which we have plotted in Figure 2.2, but we will eliminate the observation for 1978. We assume that for some reason statistics were not recorded that year, and we want to use a regression equation to estimate what that year's catch would have been. We will also use the equation to forecast what catch can be expected in 1985 and 1986.

We note that 'year' is plotted on the x-axis in Figure 5.3. We could use 1976, 1977, etc. as values of x, but we would then have to deal with very large numbers. It is far easier to label 1976 as year 1, 1977 as year 2, and so on. This greatly simplifies calculations, and gives exactly the same answer for the predicted values of catch.

The catch figures were not actually listed out in Topic 2, but could be seen fairly accurately from the graph. The actual data, together with calculations of Σx^2, Σxy and Σy^2 (which we require for a later exercise) are shown in Table 5.2. We should note here that n=8, because with data for 1978 excluded from our calculations of the regression line, we have only 8 pairs of observations.

TABLE 5.2 : CALCULATIONS OF THE REGRESSION COEFFICIENTS FOR TOTAL ANNUAL FISH CATCH IN COUNTRY ABC

x	y	x^2	xy	y^2
1	604	1	604	364816
2	552	4	1104	304704
4	677	16	2708	458329
5	621	25	3105	385641
6	875	36	5250	765625
7	880	49	6160	774400
8	774	64	6192	599076
9	869	81	7821	755161
42	5852	276	32944	4407752
$\bar{x}$ = 42/8 = 5.25			$\bar{y}$ = 5852/8 = 731.5	

Substituting in our formula we have

$$b = \frac{32944 - \frac{42 \times 5852}{8}}{276 - \frac{1764}{8}}$$

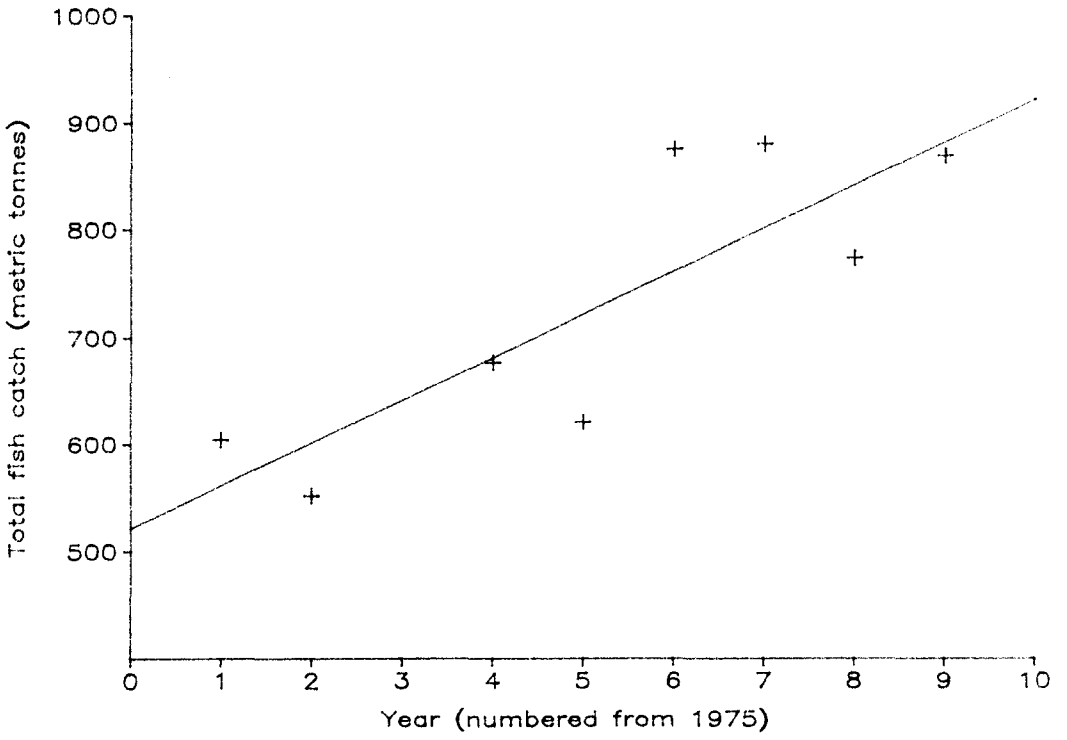
$$= \underline{40.0}$$

$$a = 731.5 - (40 \times 5.25)$$

$$= \underline{521.5}$$

Our equation therefore is y = 521.5+40x. This can be interpreted to mean that the trend line indicates estimated production of 521.5 tonnes in year 0 (i.e. 1975), increasing by 40 tonnes per year. Figure 5.3 shows this.

FIGURE 5.3 : TOTAL ANNUAL FISH CATCH IN COUNTRY ABC SHOWING LINE OF BEST FIT, Y=521.5 + 40.0 X



This allows us to make regression estimates for the other years, namely,

1978 (= year 3)	y = 521.5	+	3 x 40	=	641.5 tonnes
1985 (= year 10)	y = 521.5	+	10 x 40	=	921.5 tonnes
1986 (= year 11)	y = 521.5	+	11 x 40	=	961.5 tonnes

In practice we would have to qualify these estimates by making allowance for factors other than the long-term trend. No doubt weather conditions and other factors are going to have a substantial influence on the actual catch obtained. We can readily see from the graph that 1980 was quite a bad year, and 1976 and 1981 were relatively good years. So we should say that "without making any allowance for outside influences" our regression line makes estimates of annual catch as we showed above. We may hope that local knowledge, or other recorded information, would enable the regression estimates to be adjusted to take account of these outside influences.

We can see indeed from Figure 2.2, which included a value for 1978, that that year was a relatively poor one: catch was well below that recorded in 1976, for instance, and our regression equation is predicting an increase of 40 tonnes every year. If no allowance is made for these external factors (such as weather) and the estimate for 1978 catch is made solely on the basis of the regression equation, then the level of the catch will be over-estimated.

Linear regression estimates are not magical numbers which show exactly what would occur in given situations. They are simply best estimates based on certain available data, and without taking account of any data other than those relating to the two variables used in the regression analysis. Obviously the closer each paired observation lies to the regression line, the more accurate the estimate is likely to be, and the greater the confidence we can have in our estimates. A little later we will develop simple estimates which will indicate how closely our regression line fits the available data.

One final point should be made. Because our regression line in a time series is only an estimate, and because it takes into account only movement in one variable against time, it is of limited value for making forecasts well into the future. In our illustration we projected forward for two years, and perhaps that is as far as we should go. The further we attempt to project beyond the points in our data set, the less reliability we can place on those projections.

5.3 Non-linear relationships

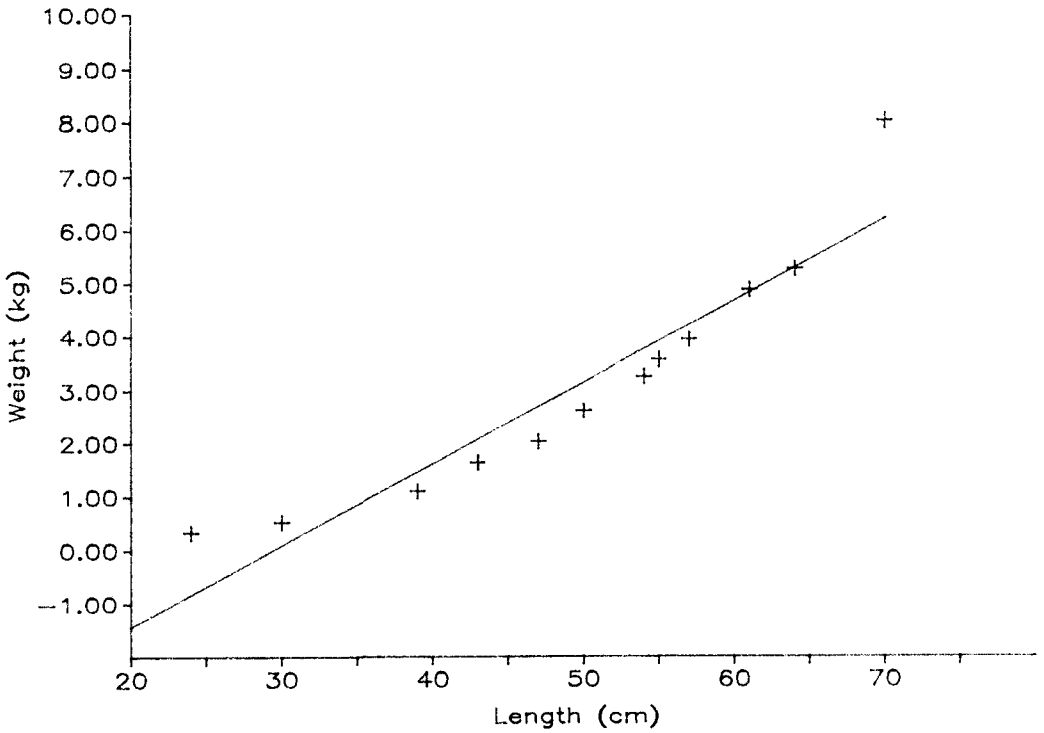
So far we have examined situations where we could reasonably expect to fit a straight line to the available data. All the points lie reasonably close to the regression line we were able to establish.

However, often two variables will bear a clearly non-linear relationship to each other. For instance, it may seem apparent from a visual inspection of a scatter diagram that the points seem to lie more or less along a curve. If we look at the length-weight relationship of skipjack, for instance, it is apparent that the observations (with length plotted on the x-axis and weight on the y-axis) clearly follow a curved path, which slopes upwards more steeply at the right hand side of the graph. In other words, for larger fish there is a large increase in weight for a relatively small increase in length.

Figure 5.4 is a scatter diagram of the length and weight of a sample of 12 skipjack, with points showing the curved pattern we would expect. This does not mean we cannot find a linear regression function which best fits these points. We can do so; there is always one line which fits a set of paired observations better than any other line will. In fact, a quick calculation will show that the equation of that line is $y = -4.5+0.153x$,

and this is plotted on the diagram. It is apparent that the regression line lies above the observed points for skipjack between about 35 and 60 cm in length, and is well below the observation for very small and very large skipjack.

FIGURE 5.4 : WEIGHT BY LENGTH OF SKIPJACK SHOWING LINE OF BEST LINEAR FIT, Y=-4.5 + .153 X



If we substitute in the equation for a skipjack of length 28 cm, we find a predicted value of y (weight) of -0.22 kg, which is clearly nonsense.

In such a situation we can sense that there must be a better way to approach the problem. In our example the points lie so close to a curve we can sketch in that we should be able to find an equation which gives a very good fit for our data.

In fact, it has been established that a relationship between weight (y) and length (x) of fish conforms to the equation $y = a.x^b$. Taking the logarithm of this equation, we have

$$\log_e y = \log_e a + b \log_e x$$

and this is a relationship similar to our equation for a straight line $y = a+bx$. In other words, if we plot a scatter diagram of log y against log x, we should be able to derive values of the coefficients log a (the intercept of the y-axis) and b (the slope), so that we have a regression line which will provide a better fit for our data than the line $y = -4.5+0.153x$ which we drew previously.

The calculations are shown in Table 5.3 and the graph with the appropriate regression line plotted is in Figure 5.5.

FIGURE 5.5 : LN (weight) BY LN (length) OF SKIPJACK SHOWING LINE OF BEST FIT, Ln (Y) = -10.82 + 3.01 Ln (X)

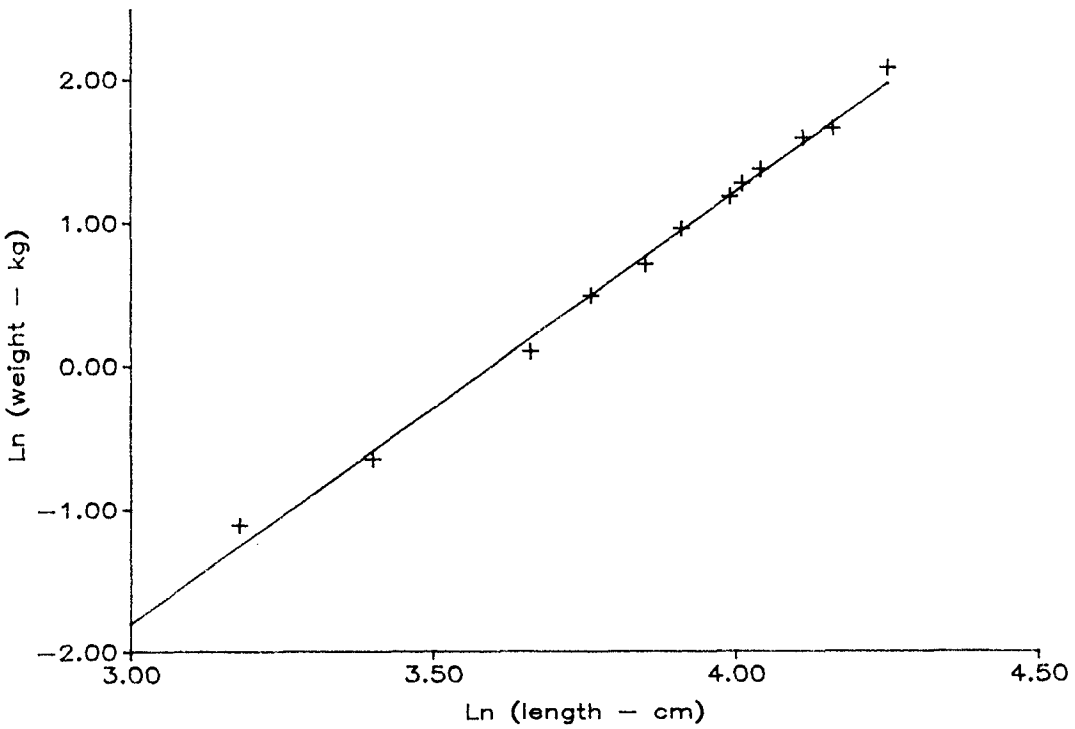


TABLE 5.3 : LENGTH AND WEIGHT DATA FOR A SAMPLE OF SKIPJACK: LOGARITHMIC RELATIONSHIP

Length (cm) x	Log x	Weight (kg) y	Log y
24	3.18	0.33	-1.11
30	3.40	0.52	-0.65
39	3.66	1.10	0.10
43	3.76	1.64	0.49
47	3.85	2.03	0.71
50	3.91	2.61	0.96
54	3.99	3.25	1.18
55	4.01	3.56	1.27
57	4.04	3.94	1.37
61	4.11	4.88	1.59
64	4.16	5.27	1.66
70	4.25	8.01	2.08

$\sum(\log x) = 46.32$ $\sum(\log y) = 9.65$

$\sum(\log x)^2 = 179.9$

$\sum(\log x)(\log y) = 40.57$

and by substituting in our formula for the regression coefficients we find

$$\log y = -10.82 + 3.01 \log x$$

(or $\log y = \log 0.00002 + 3.01 \log x$)

Now if we try this relationship for various values of x, we find the following:

If length = 28 cm, $\log x = 3.33$

$$\log y = -10.82 + (2.01)(3.33)$$
$$= -0.797$$
$$\therefore y = 0.45 \text{ kg}$$

This is obviously a much better estimate than the previous equation provided.

Similarly, if length = 43 cm, $\log x = 3.76$

$$\log y = -10.82 + (3.01)(3.76)$$
$$= 0.498$$
$$\therefore y = 1.65 \text{ kg}$$

and we may observe that this is almost identical with the weight (1.64 kg) of the 43 cm skipjack in our original data set.

In the next section we will provide a measure which clearly shows that this regression line is a far better fit to our data than the first simple formula we derived.

As a generalisation, we can say that when a series of paired observations appearS to follow a simple curve, we should be able to establish some relationship which will permit a much better straight regression line to be drawn than will be obtained by the basic formula of a line, $y = a+bx$. This may involve logarithmic, square, square root or other functions of x and/or y. There are exceptions to this: the relationship between length (or weight) and age of fish is an example where it has been found that the relationship cannot be 'linearised'. In that case, analysis has to be undertaken using techniques which are far more complex than we have discussed here. However, the idea of finding a linear relationship from basic data which is non-linear is a very important one in the analysis of fisheries statistics.

5.4 How well does the mathematical relationship describe the data

In the previous section we have seen how we can fit a straight line to a set of data, where we have n measurements of some independent variable x, and n associated measurements of some dependent variable y. We used a straight line because it is the simplest mathematical relationship we can find, and we have seen that we can still sometimes fit a straight regression line to data which bear a non-linear relationship. It is obvious, however, that for some data sets a straight line is not a good way to describe the data. If we look, for example, at Table 5.4 and Figure 5.6 we can see here that there seems to be very little relationship between the variables.

TABLE 5.4 : EXAMPLE OF NO CLEAR RELATIONSHIP BETWEEN VARIABLES. Catch per unit effort (i.e. catch per boat days fished) by effort (boat days fished) - series of 12 monthly observations.

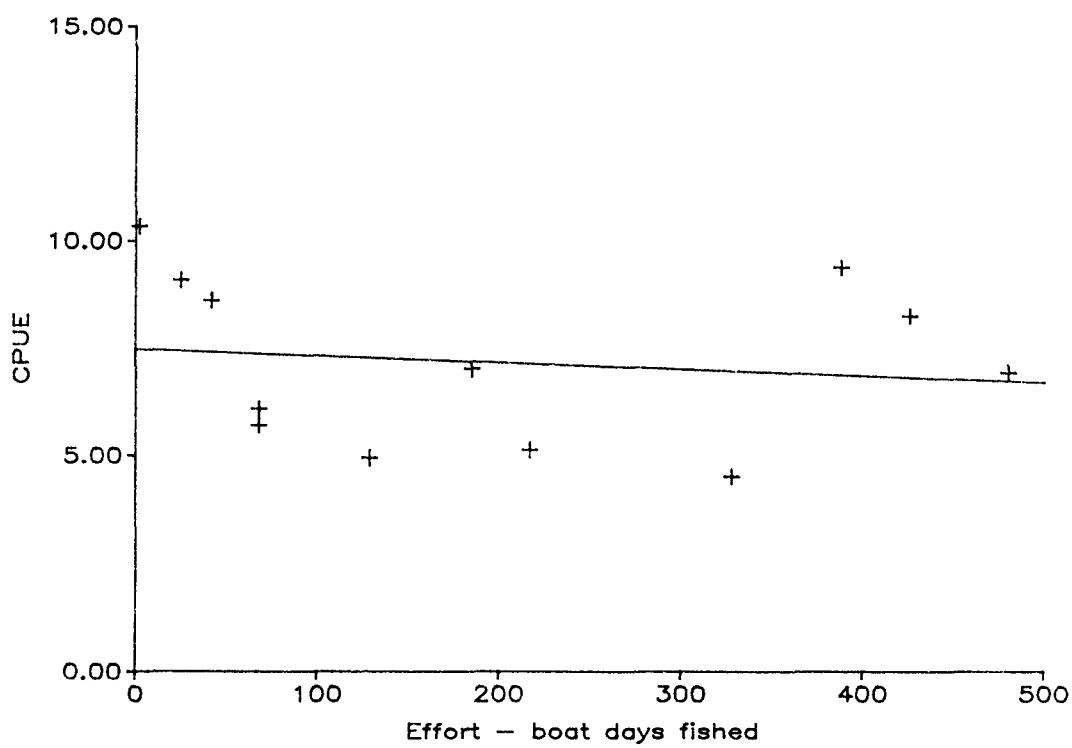
Effort (x)	CPUE (y)	x^2	xy	y^2
129	4.94	16641	637.26	24.40
328	4.52	107584	1482.56	20.43
217	5.14	47089	1115.38	26.42
68	6.12	4624	416.16	37.45
25	9.11	625	227.75	82.99
2	10.35	4	20.70	107.12
68	5.72	4624	388.96	32.72
42	8.64	1764	362.88	74.65
388	9.40	150544	3647.20	88.36
426	8.27	181476	3523.02	68.39
185	7.05	34225	1304.25	49.70
480	6.95	230400	3336.00	48.30
2358	86.21	779600	16462.12	660.95

Substituting in our formula as usual, we find a = 7.48 and b = -0.0015. The regression line therefore becomes

$$y = 7.48 - 0.0015x$$

This is drawn on the scatter diagram, Figure 5.6.

FIGURE 5.6 : CPUE (CATCH PER BOAT DAY FISHED) BY EFFORT SHOWING LINE OF BEST LINEAR FIT, Y = 7.48 - 0.0015 X



We may observe that the coefficient of the slope is negative, and therefore the line slopes downwards to the right. This indicates that the higher the effort in terms of boat-days fished, the lower the return in terms of catch per boat day.

We can see, therefore, that we can always find the equation of a line for almost any set of data. All we have to do is to perform the various calculations and put the values in the equations for 'a' and 'b'. But this is not really sufficient; we have to guarantee that when we have calculated the line for the data that it will provide reliable predictions. Because of the way we have calculated the values of a and b we know that this will be the best line for these data, but we do not know if the data points are closely grouped around the line, or if they are widely dispersed. In Figures 5.2 and 5.3 the data points do seem to be fairly closely distributed around the line we have calculated, but in Figure 5.6 we can see at a glance that this is not so; only 2 of the 12 points lie anywhere near the regression line we have drawn. Common sense would tell us that the predictions in the first two cases are much more likely to be reliable than in the third. What we need therefore is not only a method for finding the best equation of a line through the data, but also some measure of how close the data points are to the line.

5.5 The coefficients of correlation

If we look at any of the scatter diagrams above, we can see that the points do not lie exactly on the line we have drawn. Since we are interested in predicting the y values we can see how far each point is from the regression line in the y direction, that is, vertically. We call the vertical distance from the line to any of the data points the 'residual'. In effect, we can say that each observed value y_i is equal to a value '$a+bx_i$' plus the residual. The smaller the residuals, the closer the points are to the line, and the better the line 'fits' the data. One way, therefore, to see how close our data points are to the line is to measure the residuals. We can do this graphically, but we can also do it mathematically by calculating a value known as the correlation between x and y. The correlation is a measure of how close the relationship between x and y is to a straight line.

Before we go on to see how we can calculate the correlation, it is necessary to go back and look at the expression,

$$\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})$$

which we used when calculating the slope, b, of the regression line. From Topic 4 we remember that to measure the variation among a set of data points x_i, x_2 ...x_n we can calculate the standard deviation or the variance. The formula for the variance was given by:

$$\sum_{i=1}^{n} (x_i - \bar{x}) \ / \ (n - 1)$$

Now this looks somewhat similar to the expression we have just written down. This can be looked at as measuring the joint variation of x and y about their respective means. We call this quantity,

$$\sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y}) \ / \ (n - 1) \ , \quad \text{or} \quad \sum (x - \bar{x})\ (y - \bar{y}) \ / \ (n - 1)$$

the co-variance of x and y. It shows how the two variables change together. If they are closely related, this value will be high; if they are not closely related, it will be small. Notice, however, that the

co-variance can be negative; if the relationship slopes downwards then it will be less than zero.

The correlation between x and y is measured by a coefficient, which we denote as 'r', and this is given by the co-variance of x and y divided by the product of the standard deviations of x and y. In terms of a mathematical formula we can write it as:

$$r = \frac{\Sigma(x - \bar{x})\ (y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \ .\ \Sigma(y - \bar{y})^2}}$$

(The values (n-1) can be divided out in both numerator and denominator). If we only used the co-variance to measure the relationship we would have problems in comparing different sets of data. For example, if we changed the units of measurement of one, or both, of the variables, we would change the value of the co-variance. We get round this problem by dividing by the product of the standard deviations; this means that r can only be a value between plus and minus one. If all the data points lie exactly on an upward sloping line, then r will be +1; if they all lie on a downward sloping line, r will be -1. Values in between, then, tell us how strong the relationship is between x and y.

If r is very close to +1, we say there is a strong positive correlation: y increases as x increases, and the relationship is good. If r is close to -1, there is a strong negative correlation: y decreases as x increases. When r is close to zero (either positive or negative) there is very little relationship between the two variables.

As with all similar measures we have studied, we find that in practice our alternative version of the formula is more suitable for ordinary use, especially with a calculator. This is

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \ .\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

We may note that r can be calculated if we can obtain the values of n, Σx, Σy, Σx^2, Σxy, all of which were used in our earlier calculations of the regression coefficient 'b', and one additional value, Σy^2. It was for this purpose that we obtained the value of Σy^2 in our earlier examples in this topic.

If we revert to our data in Table 5.1, and substitute in our formula for 'r', we have

$$r = \frac{10(86020) - 127(6250)}{\sqrt{10(1731) - (127)^2} \ .\ \sqrt{10(4346100) - (6250)^2}}$$

$$= \frac{66450}{\sqrt{1181} \ .\ \sqrt{4398500}} = 0.92$$

This is a positive value, quite close to +1, so we can say that there is a good positive relationship between the two variables, according to our data.

We can go through a similar process to calculate 'r' for our time series data on annual fish catch (Table 5.2), and we find r = 0.84. In other words there is a good positive correlation (i.e. our catch is moving

upwards over time), but the relationship is not as strong as in the previous example.

The calculation of the coefficient of correlation for our two alternative formulae for the length-weight relationship of skipjack is quite interesting. We would expect a strong positive relationship, because obviously weight increases as length increases. In fact, in our first equation, which was $y = -4.5+0.153x$, we can calculate that $r = 0.94$. This shows that, despite the poor estimates which the regression line gave for very small and very large fish, the fit of this line to the data we had available is good.

However, when we examine the second equation, $\log y = \log (0.00002)+3.01 \log x$, we find that $r = 0.997$. This is very close to 1, and clearly shows that we were able to find a much better regression line by our special technique of using a 'log log' relationship.

Finally, we can look at the relationship of 'effort' to 'catch per unit effort', portrayed in Table 5.4 and Figure 5.6. We noted at the time that y was decreasing as x increased, so we must expect r to be negative; we also observed that the relationship appeared to be very weak, so we should anticipate obtaining a value for r which is closer to 0 than to -1. When we make the appropriate calculation we find that $r=-0.13$, which is so close to zero that hardly any relationship at all can be established. We could have very little confidence at all in any conclusions we attempted to draw from this regression equation.

We can summarise what we have been discussing by saying that the regression coefficients measure the linear relationship between two variables, and the correlation coefficient tells us how closely the data fit this relationship. The two are clearly related to each other, but it is important not to confuse them because they measure different things.

It is possible to go much further than this in analysis, and calculate 'confidence limits' for our estimates. Essentially this is another way of expressing the goodness of fit of a relationship, but this is beyond the scope of this course. However, we may note that in our next topic, Sampling, we will be calculating confidence limits for estimates derived from samples, and there is a close parallel between the two.

5.6 Seasonal variation

Earlier in this topic, in examining links between two variables, we looked at an illustration of one of the most important relationships - that between the values of a characteristic and time. This relationship is referred to as a time series .

That particular example concerned a time series of annual data, and we obtained a regression line which best fitted the observations and we referred to this as the trend line. The line endeavours to show how the value of the characteristic is changing in the long term.

However, often we are vitally interested, not just in the trend in annual values, but in movements in the shorter term - from month to month, for example. We then may find that our study is complicated by a pattern of peaks and troughs in the value of our observations. If this pattern tends to repeat itself each year, we call this seasonal variation. By this we mean that the value of our variable tends to vary according to the time of the year; at some times it will be high, at others low, but the pattern

will tend to repeat itself regularly. This kind of variation occurs quite often in time series, particularly those related to production or climatic factors.

If we look at the figures of tuna catch in Fiji given in Table 5.5, we will see that there is a very high, and quite regular, seasonal variation. In each of the four years covered by our data, the catch is highest in the first five months of the year, from January to May, and is lowest around August to October, being almost zero each September.

TABLE 5.5 : IKA CORPORATION, FIJI - ESTIMATED TUNA CATCH, MONTHLY 1979-1982. Calculation of 12-month moving averages.

Month	Tuna Catch (tonnes)	12-month Total	24-month Total	Moving Average
Jan. 79	594			
	488			
	535			
	468			
	566			
	354			
	190	3358		
	18	3094	6456	269
	2	2980	6074	253
	0	2911	5891	245
	57	2751	5662	236
	86	2257	5008	209
Jan. 80	330	1988	4245	177
	374	1927	3915	163
	466	2009	3936	164
	308	2038	4047	169
	72	2123	4161	173
	85	2205	4328	180
	129	2335	4540	189
	100	2824	5159	215
	31	3408	6232	260
	85	4001	7409	309
	139	4365	8366	349
	216	4888	9253	386
Jan 81	819	5267	10155	423
	958	5568	10835	451
	1059	5553	11121	463
	672	5522	11075	461
	595	5522	11044	460
	464	5542	11064	461
	430	5657	11199	467
	85	5467	11124	464
	-	5288	10755	448
	85	4991	10279	428
	159	4892	9883	412
	331	5015	9907	413
Jan. 82	629	4939	9954	415
	779	4659	9598	400
	762	4624	9283	387
	573	4624	9248	385
	718	4666	9290	387
	388	4618	9284	387
	150	4555	9173	382
	50			
	-			
	127			
	111			
	268			

Source: Annual Report 1982, Fisheries Division, Ministry of Agriculture and Fisheries, Fiji

We would like to look beyond this seasonal variation, and try to find how the catch, as a whole, is changing over time. Obviously we cannot just compare data for consecutive months. It would be quite unreasonable to conclude, for example, that the tuna catch is falling, just because the amount caught fell each month from May to September 1982. The catch falls in that period every year, and what we would need to establish, in order to get any picture of a longer-term trend, is whether the fall in these months of 1982 was greater or less than the fall which is usually recorded at that time of the year.

There are various techniques available to calculate this seasonal variation. If we can obtain a measure of this variation we can eliminate it from our data, to give us a more meaningful trend line. This process is referred to as seasonal adjustment. There are now quite sophisticated computer programs which are widely used by statisticians all over the world, to seasonally adjust (or 'deseasonalise' as it is sometimes called) any time series data. We will not go into this topic in detail during this course, but we will look at the first step in the process, the moving average, and see how this can assist in eliminating seasonal patterns from data, in order to highlight the trend.

To see how this works consider the following sets of numbers:

4, 5, 7, 3, 6, 4, 5, 3, 7, 6, 3, 4.

The average (arithmetic mean) of the first three numbers is: (4+5+7)/3 = 5.3. We could then 'move' the average along, and find the average of the second three numbers, 5, 7 and 3; this will be (5+7+3)/3 = 5.0. We can repeat the process by moving the average along the series, one observation at a time. This then gives us the following situation.

Original series	4	5	7	3	6	4	5	3	7	6	3	4
Moving average of order 3		5.3	5.0	5.3	4.3	5.0	4.0	5.0	5.3	5.3	4.3	

Since our first moving average is the mean of the first three terms (4, 5 and 7) we can place this underneath the middle value and then move the average along one value each time. We have calculated a moving average of three terms; we call this an average of order 3. An average of order 5 would include 5 terms and so on.

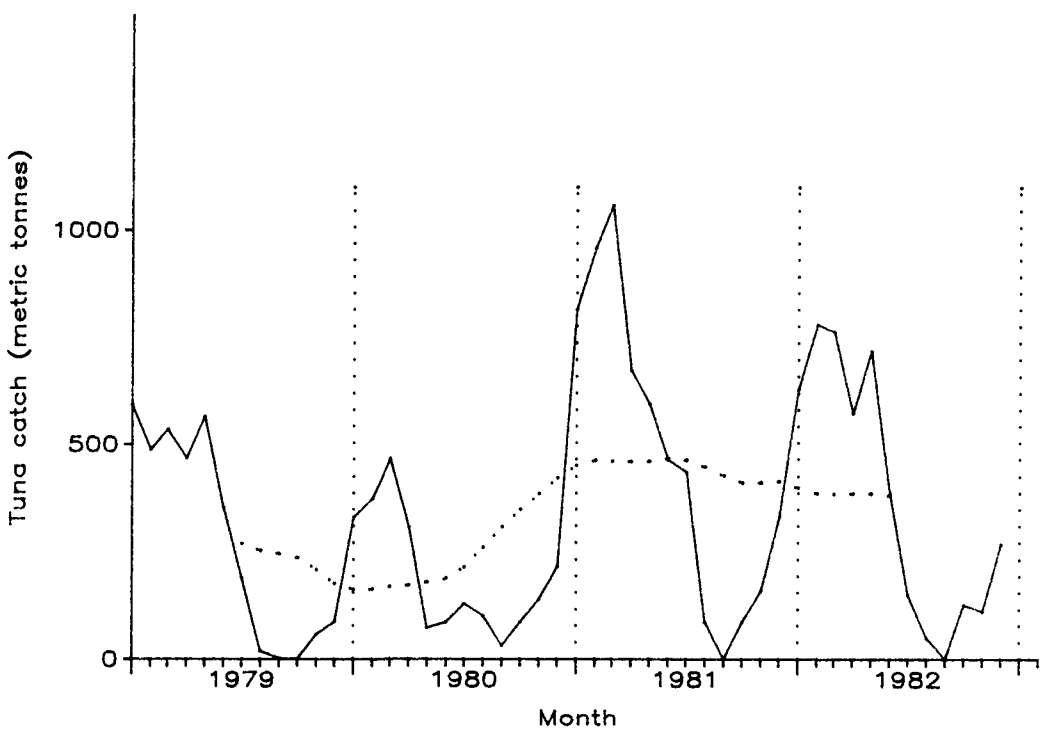
One of the reasons that we calculate a moving average is to reduce the variation in the original series; in our example the moving average is less variable than the original series. We also see that the series of averages is shorter than the original; we have "lost" terms at the beginning and the end. This is inevitable, because of the way we calculate the moving average. The greater the order, the more variation will be smoothed out, but the more terms that will be "lost" at the beginning and end of the series.

A moving average, therefore, will smooth out random variation in a time series, and if we choose an appropriate order it will also eliminate the seasonal variation. What will be left will be the trend values. We have defined the seasonal variation to be that which varies with the seasons, so which will repeat itself annually. If we have monthly data, therefore, we shall need a moving average of order 12 to eliminate the

seasonal variation; with quarterly data we will use an average of order 4. There is, however, one extra problem when we use an average with an order which is an even number. If we calculate an average of order 12 and start with observations from January 1979 to December 1979 (as is done in Table 5.5), this will be centred half-way between June and July 1979. The average for February 1979 to January 1980 will be half-way between July and August 1979, and so on. This is obviously inconvenient; we want to know the trend value of June and July and not some mid-point. What we have to do, once we have calculated the first average, is take a second average of two terms which will 'centre' our trend values. We call the resulting moving average as one of order 2x12.

The calculations of the 2x12 moving average for the Fiji tuna catch were shown in Table 5.5; the moving averages, or trend values, have been plotted on Figure 5.7. We can see that the moving average has removed most of the random and seasonal variation and so allows us to get a much better idea of the trend.

FIGURE 5.7 : MONTHLY TUNA CATCH (FIJI) AND 12-MONTH MOVING AVERAGE (broken line)



Of course moving averages do not eliminate the effects of different conditions from year to year: in fact they help to highlight these effects. It is very easy to see from the moving average line on the graph that 1980 was a bad year, and 1981 a good year, for example.

The example we have given relates to repeated fluctuations which occur during a year, and this is probably the most common seasonal variation we will encounter. However, we can also encounter seasonal patterns over different periods - for example:

Monthly

There may be a repeating pattern during each lunar month. Studies have shown that catches of baitfish are regularly highest at the time of new moon, and lowest at time of full moon.

Daily

Some variables may change regularly at different times of the day. The price of fish in the local market, for instance, may be highest early in the morning, and may be lower later in the day as vendors reduce their prices in order to get rid of their unsold stock.

The same techniques can be applied in these circumstances, in order to remove the repeating pattern, and to obtain a more realistic trend line for data.

TOPIC 6 - SAMPLING : HOW TO GET SOMETHING FOR A LITTLE

6.1 Introduction

'Sampling' is the process of choosing a portion of a population to represent the whole population. It contrasts with a 'census' when every member or unit of a defined population is included. Almost everyone is familiar with situations in which judgements are made about a whole group of items when information is available for only a few of them. If we want to find out about the quality of a sack of rice we would probably only pick out a couple of handfuls and look carefully at these; it would not be necessary to investigate the whole sackful. Similarly, when testing the grade of a shipment of copra, only a small part of the shipment is actually tested, it being assumed that the remainder will be similar.

These simple situations are examples of a common statistical technique known as sampling. We are selecting a number of units from a population, observing some characteristics for these units and then using the results from the sample to estimate values for the whole population. Obviously this will be an important, practical technique, because if we can achieve reasonable results from a sample of observations, then this will be much cheaper and more efficient than having to observe every unit in the population.

In many situations, including the examples we looked at above, the procedure of sampling is quite simple and straightforward. We know that rice selected from one part of the sack will be very similar to that taken from some other part. We will not obtain a very different result if we take two handfuls or twenty. Similarly, if a doctor wishes to take a sample of blood from a patient in order to test for the presence of some disease, he knows that wherever in the body he takes the sample, he will get the same results so he can take just one blood sample.

The important thing about all the examples we have looked at so far is that the population of items under consideration were well mixed up; they did not vary very much. Technically, we can say that these populations are relatively homogeneous, which means that the variability between units is small.

We must realise that, to obtain fisheries statistics, particularly information on artisanal and subsistence fishing which is of interest to many governments in the region, we will have to use sampling methods. It would be too expensive and too time-consuming to try to run a continuous collection of data about all the fishing effort and catch in a country. Unfortunately when we look at the kind of populations we have to deal with, we will expect to find that they are far from homogeneous, and sampling will not be nearly as easy as in our simple examples above. Some villages will no doubt catch far more fish than others, so we cannot use data from one or two villages to tell us about the total fish catch of the country. The catch will no doubt vary from day to day, and from one time of the year to another, so we cannot easily choose data for one day or one week to estimate catch accurately for a whole year; different types of boat, different fishing techniques, etc. will produce different results, and we will have to make allowance for this in designing a sample to derive our estimates.

The populations we usually have to deal with, then, are fairly heterogeneous; there is considerable variability. In this case, sampling

is more difficult. Very often we have very little prior information about the population we want to study; all we know is that it is heterogeneous.

In this topic we will look at different types of samples we can design to give us the estimates we need; a little of the mathematical theory of sampling; ways to measure the accuracy or reliability of the estimates we obtain from our sample; and how to assess the size of sample we will need in order to achieve an acceptable level of reliability in our results.

6.2 Some concepts and definitions

First, it will be useful to define some new concepts which we shall use. These follow on from our previous definitions in Topic 2. There we looked at the terms: statistical unit, observation, characteristic, and population.

Finite and infinite populations

We sometimes need to distinguish between 'finite' and 'infinite' populations. A finite population is one which has some limit to its size, e.g. the number of foreign longline vessels operating in a country's waters; the total catch by artisanal fisheries in a country, and so on. An infinite population is one which has no limit (or is so large that we cannot identify a limit); for example, all the fish in the sea. It is interesting to observe that for infinite populations there is no alternative to sampling in order to make estimates of population characteristics. It is impossible to measure the average fork length of the whole population of skipjack tuna; to do that we would have to catch every skipjack in the sea. All we can do is make estimates based on a sample of fish.

Sampling unit

Elementary units, or groups of units, which are convenient for purposes of sampling, are called sampling units. For example, in a subsistence fishing survey we may find it most convenient to make a selection of villages from which to collect data. The village is the sampling unit.

We must be very careful here to distinguish between the terms 'sampling unit' and 'statistical unit'. These may be the same, but this is not necessarily so. For instance, if we select a sample of villages, then 'village' is the sampling unit. But if we then collect our data of catch and effort for each boat in the village, then 'boat' is the statistical unit.

Thus, we define 'statistical unit' to refer to the element we wish to collect information about, and the 'sampling unit' to refer to the element or group of elements which we use as a basis for sample selection.

As we shall see later, we often select samples in two or more stages. For example, we might select a number of villages at the first stage, and then within each selected village, we could choose a sample of households from which to collect data on fishing. In this case we refer to the village as the 'primary sampling unit' and the household as the 'secondary sampling unit'.

This multiple usage of the word 'unit' can be quite confusing. We also talk about a 'fishing unit' which we define as the smallest discrete, complete unit necessary for a fishing activity. In practice, we can expect the 'fishing unit' to equate to the 'statistical unit'. We will be trying to collect information about the 'fishing unit'.

Sampling frame

In order to select a sample from a finite population we need a list of all the sampling units. We call such a list the sampling frame. The frame must be complete and up-to-date; if any unit is not included then it has no chance of being selected in the sample and this may well lead to inaccuracies in the results.

The preparation of the sampling frame can be one of the most difficult and time-consuming tasks in a sample survey. We also often find that the information that we have available to provide the frame will limit the kind of sample we can select and the results that we can obtain. Information from previous investigations is sometimes suitable, e.g. records from a population census can provide a valuable source of data for sampling frames.

Sample size

The 'sample size' is simply the number of sampling units we select to be in the sample. Obviously the sample size must be less than the number of units in the whole population.

Sampling fraction

The 'sampling fraction' refers to the proportion of the population which is included in the sample. It is usual to refer to a population of 'N' units, and the sample as consisting of 'n' units. The sampling fraction is then: n/N. For infinite populations, the concept of a sampling fraction does not exist.

Parameter

We use the word 'parameter' (or population parameter) to mean the true value of the characteristic of the population which is being estimated. Thus, for instance, in a sample survey of local fisheries in a country, the population parameters we are interested in will probably be the total catch of fish in the country, the average catch per boat day for all boats operating in the country, the proportion of all households which are engaged in subsistence fishing, and so on.

Sampling error

We use results from a sample to make estimates of population parameters. The word 'estimate' indicates that we do not know the exact value of the parameter, but that we hope to be quite close. Obviously, if we do not measure or collect data from every sampling unit within the population we cannot expect to obtain the value of a population parameter exactly. We refer to the difference between the estimate and the true value as the sampling error. In a sense this is the price we have to pay

for only observing part of the population. With a large sampling error then, the estimate obtained from our sample will be inaccurate; if the sampling error is small, the estimates will be close to the true value.

There is, however, a problem, because normally we do not know the value of the population parameter. There would be little point in undertaking a survey to collect results to estimate a value which we already know. It is, therefore, impossible to calculate the sampling error exactly. What we can do is to calculate how large we expect the sampling error to be, provided we select the sample according to certain, well-defined rules. To help understand the idea of sampling error more clearly it will be useful to look at Figure 6.1.

FIGURE 6.1 : REPRESENTATION OF SAMPLING ERROR

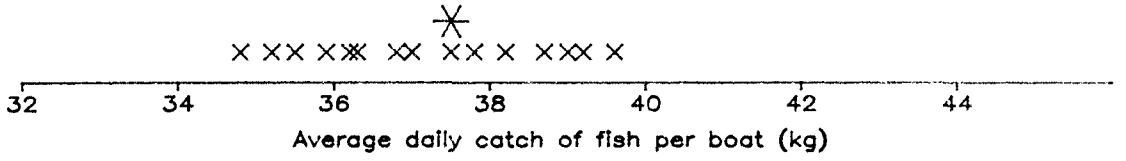


Figure 6.1 represents the results of a sample survey to measure the average daily catch of fish per boat in a country. We use a scale to represent catch and we can plot the value we obtain from a sample as a small cross. The true population value, is represented by the large asterisk. In practice, of course, this value will not be known. Let us assume that the population we are considering consists of 500 boats and that we are taking a sample of 20. With a different sample we get another estimate and we can plot this as another cross. On the diagram we have plotted the results from 15 different samples, but we could have plotted many more. In fact, there are 267×10^{33} different possible samples (that is, 267 followed by 33 zeros), and each of these may well produce different estimates. We could plot all these estimates as a frequency distribution and we would find that the shape would be almost exactly the same as a Normal distribution which we looked at in Topics 3 and 4. This will always be the case, provided the sample size is large enough (say, 20 or more), more or less regardless of the actual distribution of the population.

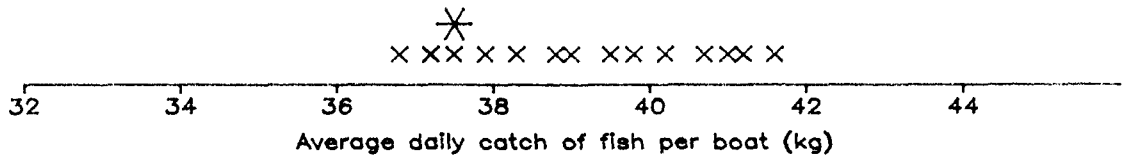
Now, as long as the sample results are clustered around the true value, or, putting it technically, that the mean of the Normal distribution is the population mean, then a measure of the accuracy of the sample estimates is provided by the degree of dispersion of the distribution. We measure dispersion by the standard deviation, but to distinguish between the standard deviation of the population and the dispersion of the sample estimates around the mean, we call the latter the <u>standard error of the estimate</u>.

This is a very important concept in statistics. We will come back to it a little later in this topic, and will give the formulae for calculating the standard error for some principal types of samples.

Bias

In the previous section we saw how the standard error of an estimate is a measure of its precision, provided that the distribution of all the different possible estimates is centred round the true population value. If this is not the case, we say that our sampling scheme is 'biased' and an example of this is given in Figure 6.2.

FIGURE 6.2 : EXAMPLE OF A BIASED SAMPLE



We are using the same situation as before, but with this sampling scheme we can see that the different estimates are not clustered round the true population mean but around some higher level. If we looked at the distribution of all the possible sample estimates, it would still look like a Normal distribution, but this time the mean of the Normal distribution would not be the same as the population mean. The difference between the two is the bias.

With a biased sample the accuracy of the estimate is not measured just by the standard error; it also includes the bias. Technically, we can measure the accuracy by the square root of the sum of the squares of the bias and the standard error. This will not matter very much as long as we know the value of the bias. In a very few situations we do use biased samples because they may be more accurate in the end, but the important thing is that we know what the bias will be.

In practice, however, bias may well be introduced and we do not realise it, and so we do not know how large it will be. It can arise in several different ways: from problems in preparing the sample frame, from the way the sample is selected, from the way observations are made, from non-response, from mistakes in calculations and also, in some cases, from the way we make the population estimates.

In our example of catch per boat, bias could arise because we have faulty scales which give a wrong reading; because people collecting the data decide to guess the weights, instead of measuring them accurately; because we fail to observe all the fish caught, e.g. by missing out on part of the catch at night; because we have failed to include in our frame some boats which have a different average catch from the boats we have included; or for a variety of other reasons.

Bias also introduces another drawback into samples. In general, we hope that as we increase the size of the sample we improve the estimates by reducing the sampling error. In other words we get more accurate results at the cost of having to make more observations. If the sample is biased, however, it does not matter how large the sample is made; the bias will still be present.

Generally speaking, we may say that it is very important to try to reduce bias, or eliminate it altogether from our surveys.

Non-sampling errors

We use the general term non-sampling errors to refer to all the types of errors and mistakes that can occur when we undertake a survey, other than the basic inaccuracy that is a result of the sampling process itself. Non-sampling errors can arise because of mistakes by enumerators, wrong answers given by respondents, problems with the sampling frame, poor data processing techniques, and many other reasons. They will, of course, happen in complete censuses as well as sample surveys and, in fact, can be a serious problem in these cases because of the much larger nature of the operation. In a sample enquiry, however, it is especially important to try to control these errors, because each sampling unit 'represents' many others in the population; just as we multiply our sample results to estimate population totals, so we multiply the effect of each error. Most statistical textbooks tend to concentrate on techniques for reducing the sampling error and to overlook the operational difficulties of actually carrying out the survey. This is mainly because the effect of non-sampling errors is very difficult to estimate and varies considerably from survey to survey. There is no mathematical technique we can use, as we can to calculate the sampling errors of estimates. When detailed research has been undertaken, however, it has been found that non-sampling errors in some surveys can be at least three times larger than the sampling error. What we have to do, when carrying out any investigation, is to build in as many checks and controls as possible.

6.3 Methods of selecting a sample

6.3.1 Random and non-random samples

First we should differentiate between two types of sample selection - random and non-random. A random method of selection is one which gives each of the units in the population a specified, or calculable (and non-zero) probability of being selected. This is sometimes referred to as probability sampling.

Other methods of sample selection are referred to as non-random, or non-probability. For example, suppose we wish to collect a sample to estimate the total subsistence and artisanal fish catch in a country. If, for reasons of cost and convenience, we were to restrict our sample selection to boats operating within 20 km of our urban centre, we would have a non-random sample. We cannot really expect that estimates obtained from within and near an urban centre are truly representative of the whole population.

A rather similar situation arises with 'judgement' sampling where the sample selected is one which we believe, or feel intuitively, would be representative of the whole population. We may feel confident that (say) two particular islands are 'typical' of the whole country; in other words we make a subjective judgement that we can estimate population parameters by selecting a sample from those islands only. This may be true; on the other hand it may not. There is no way of knowing, or calculating, how well such a sample does in fact represent the population.

All non-random samples suffer from one very serious drawback; there is no mathematical way to calculate the sampling error. A sample based on the

laws of chance, on the other hand, can provide a measure of how precise these estimates are. Thus, we have an objective means of evaluating the results of a survey. This is a most important characteristic, and is the biggest single reason why statisticians prefer to use random, or probability, sampling methods whenever possible. For the rest of this chapter we will concentrate our discussions entirely on random sampling.

There are several different types of random sample, which will be discussed a little later. In some types every unit in the population has an equal chance of being selected; in others some have more chance than others. What all random samples have in common, however, is that every unit has some chance of selection which is known or can be calculated. In a non-random sample this is not so.

6.3.2 The use of random numbers

In order to select units at random we need some kind of random process that produces results with no order or pattern, but where each unit has a known probability of selection. Some examples of such processes are:

(a) tossing a coin;
(b) throwing dice;
(c) selecting numbers out of a hat.

Any of these methods could be used to select a sample, but, in practice, they will be rather cumbersome to use, particularly if the sample size is at all large. Therefore, most people use random numbers from a computer, a calculator, or already prepared tables. An example of a table of random numbers is given in Table 6.1. This table consists of a number of digits and there is absolutely no pattern or order in the way these digits are written down. The table can be read horizontally or vertically. The gaps do not mean anything; they are simply there to make the table easier to read.

TABLE 6.1 : EXAMPLE OF A TABLE OF RANDOM NUMBERS

87 08 83 09 40	14 39 15 99 24	21 85 00 45 54	19 36 18 03 88
88 33 78 20 40	40 24 73 77 70	00 31 84 59 25	06 50 30 95 96
22 50 09 11 00	37 36 51 55 95	83 97 13 75 46	22 77 50 11 72
48 70 56 57 16	24 21 74 91 53	18 05 59 61 74	97 31 82 77 68
93 45 40 93 12	80 88 63 26 93	85 06 19 87 84	37 59 76 16 65
50 76 72 02 39	19 40 69 57 23	09 33 20 70 86	45 13 94 98 39
91 64 01 34 67	13 11 00 32 09	39 76 21 64 29	85 65 14 51 74
33 20 63 71 95	94 13 77 12 94	91 04 41 83 79	72 44 08 12 44
90 59 65 46 78	82 16 45 97 85	57 75 79 96 79	08 16 83 43 99
05 10 93 57 80	32 86 65 26 90	27 45 34 94 46	33 65 35 56 84
92 85 63 26 69	69 81 54 70 56	17 62 43 17 86	78 99 62 34 15
08 50 36 45 82	11 26 54 76 88	86 67 82 21 65	00 82 89 06 09
59 36 77 09 83	78 81 77 93 77	48 44 88 30 37	21 74 02 93 10
05 85 86 43 25	50 76 70 36 32	26 68 54 92 84	90 02 38 77 40
13 46 99 31 30	29 71 70 91 10	99 84 55 31 95	20 90 28 49 78
56 27 09 33 66	79 33 29 50 54	76 94 27 01 45	78 29 66 23 15
54 14 52 11 22	33 39 39 58 30	73 43 59 32 26	43 76 12 99 10
83 01 86 58 89	77 68 87 29 71	49 50 46 53 41	53 52 20 56 53
00 28 17 33 81	42 24 33 55 75	42 70 73 65 16	96 47 17 42 69
52 29 69 59 32	59 40 30 89 12	11 07 18 53 27	13 46 54 85 40
54 43 09 80 68	29 86 65 60 27	87 70 77 45 31	69 12 31 21 79
80 68 13 48 80	84 25 33 70 89	76 61 03 41 57	89 97 07 56 12
28 72 57 80 54	05 80 92 82 65	25 01 74 58 89	39 25 05 57 66
23 48 49 96 00	17 88 90 63 67	02 64 71 12 21	02 29 86 88 54
04 41 27 70 10	49 13 76 99 28	64 14 90 60 69	75 10 97 16 60

Let us look at an illustration of how we use this table, to select a random number between 1 and 63. We choose a random starting point on the table - say the 11th row of the 17th column of paired digits. We will observe that this starting point is the number 99. It is too large for our requirements (i.e. it is greater than 63) so we must reject it and take the next number. If we are working vertically the next number is 82 and the next 74; both are too large so they too must be rejected. The next number, 02, since it falls within our specified range, becomes our random selection. If we need another selection we must continue from immediately after our last selection. Thus, we would have to reject 90 as being too large, and our next selection would be 29.

If we had been working horizontally from our starting point, we would have rejected the first number 99 as before, and would have selected the next number, 62, which falls within our specified range. Our second selection would be the next random number, 34.

We may see from this that we really need rules about using random number tables. For our purposes we will work vertically until reaching the foot of a column, then continuing at the top of the next column, and so on. If random numbers are being extensively used, we would need more precise rules than that.

6.4 Types of random sample

There are many different types of random sample, and we will concentrate on a few of the best known and most commonly used.

6.4.1 Simple random sample

The most basic type of random sample is known as a 'simple random sample', which can also be written as srs, for short. From a population of N units a sample of n is selected. This is done in such a way that any one of all the possible samples that could be used, is equally likely to be chosen. In effect this also means that every one of the N units has the same chance of being in the sample.

Simple random sampling can be realised by selecting units one by one with equal probability (i) replacing units already selected before the next draw, so that in fact the same unit may be selected more than once, or (ii) without replacing the selected units before the next draw. The former is termed 'srs with replacement' and the latter 'srs without replacement'. The latter can be shown to provide a more efficient estimate than the former.

It may be noted that srs is not widely used in practice, mainly because some information or other is usually available for all the units in the population and this information can generally be utilised in the selection schemes which are discussed below, to increase the efficiency of the sample design.

6.4.2 Systematic sample

A systematic sample is one in which the sample is selected from a list of the population according to some pre-determined systematic pattern.

Perhaps the most commonly used method is to make selection at regular intervals from the list. For example, to draw a 10 per cent sample we would select every 10th unit, and would do this by drawing a random number between 1 and 10 to choose the first unit. If this were, say, 5, then the units selected in the sample would be the 5th, 15th, 25th, 35th, and so on. With this method there is no chance for various combinations of units to be selected, e.g. it is impossible to select both the first and the second units on the list, as could occur with simple random sampling.

This fact can be turned to our advantage, and systematic sampling can provide a more efficient and more representative sample than could be obtained by simple random sampling. This is achieved by first arranging the units in the list in a suitable order. For instance, if we wished to draw a sample of villages, we might list the villages geographically. This systematic procedure guarantees a very good geographic 'spread' of selections. It avoids the possibility, which is always present in a simple random sample, that by chance we might select, for example, a higher proportion of villages near an urban centre than is actually present in the population. So, provided the list is ordered in a satisfactory way, we can be more confident of drawing a representative sample. However, the list must not be prepared so that it contains a regularly repeating pattern, as this can lead to a most unrepresentative selection.

Another form of systematic sample, which has been shown to be very efficient, is called a 'Balanced Systematic Sample'. With this method, the population is listed in a suitable order, and selections are made at equal distances from each end of the list. For example, in a list of 100 units, if the first unit is selected by random means, then this would be balanced by also selecting the 100th unit; if the 12th unit were selected, then the 89th unit (i.e. the 12th from the other end of the list) would also be selected, and so on.

In general we can say that systematic sampling is a good method in many circumstances, since it is unbiased, is easy to understand and to operate, and gives us an efficient sample.

6.4.3 Stratified random sample

In simple random sampling the selection of the sample is left to the luck of the draw. No use is made of any knowledge that we possess about members of the population. If we have such knowledge, we should be able to improve upon simple random sampling by using the knowledge to guide us in the selection of the sample.

For example, suppose we wish to estimate the average daily landing per vessel from a fishing fleet at a particular port, by taking a random sample of the fishing boats. If all of the fishing boats are similar, then we can proceed as described before. But if the fleet consists of, say, 100 small canoes and four large motorised boats, obviously our answer will depend greatly upon whether our sample happens to include one or more of these large boats.

In circumstances like this we can often improve the accuracy of our estimates by dividing the population up into groups, or strata, and we can then take a sample from each stratum separately.

In the example we gave above, we would stratify by type of boat - with powered fishing boats in one stratum, and canoes in another. This is the principle behind stratification: we try to have each unit in a stratum as

similar to each other unit as possible (in terms of the characteristic we are measuring). Thus, no matter what unit we happen to select will be representative of other units in the stratum. However, we can make the difference between strata as great as we like, and indeed it is to our advantage to do so. We refer to this as 'low within-stratum variability' and 'high between-stratum variability'. The problem in practice is that we need a suitable sampling frame in order to make the stratification, and this may be a limiting factor.

The most common, and the most obvious, method of stratification is geographic. For a fisheries survey we would almost certainly wish to stratify between high islands and atolls, and between rural and urban areas, for instance.

As well as improving the precision of our overall population estimates, stratification is important for other reasons. We may well need, for example, estimates for different districts or provinces as well as for the whole country. By making each district a stratum we automatically get results for the district. Using stratification also allows us to change the size of the sample in each strata. If one area is very expensive to survey then the sample size can be reduced, and if another area seems to be very variable then it can be sampled more intensively.

Once the strata have been defined, a separate sample is taken from each one, using simple random or systematic sampling.

6.4.4 Multi-stage sampling

The two main types of random sample that we have looked at so far, simple random samples and stratified samples, while being very useful techniques, do have two drawbacks when used in many Pacific countries. Firstly, the cost of collecting the data for each selected sampling unit can be very high, which means that, because the overall budget is usually limited, the sample size has to be reduced. The main reason for this high 'unit cost' is the amount of time that is required to reach scattered units which can often be quite isolated. Obviously this will be much more of a problem with surveys in rural areas. Choosing a simple random, or stratified sample, could well mean that it would be necessary to visit a large number of villages. In some countries this can mean a journey of two or three days simply to obtain one observation. Since a large part of the cost of any rural survey is accounted for by salaries and transportation and since time taken to locate a unit is not productive, it is clear that we need some way of organising the sample in order to reduce the amount of travelling required.

The second major disadvantage is that both types of sample require a complete sampling frame covering the whole population. To carry out a fishing survey in a country in which the sampling unit is to be the boat (or the fishing unit), we would need to prepare a list of every boat (or every fishing unit) in the country for our frame, and this is likely to be quite impractical. In many data collection exercises one of the most difficult problems can be the preparation of a sampling frame.

Multi-stage sampling has been developed to help overcome these two serious drawbacks, although, as we shall see later, this can only be done at the expense of a certain amount of accuracy in our estimates. The basic idea is that, instead of selecting a sample of our final sampling units, we combine these units into groups and then select a sample of these groups.

For example, in our fishing survey we could first of all list villages and select a sample of these. Since we only select a sample of villages, we have immediately cut down on the amount of travelling required to go from place to place. In addition, to select this sample all we need is a list of villages and not a complete frame of all fishing boats. We would then need to make a list of boats in our selected villages, but this is obviously a far simpler task than making a list of all the boats in the country.

We can then undertake a second stage of sampling, selecting boats within the villages already picked out. We have illustrated here two-stage sampling; the first stage was selecting villages and the second stage choosing boats. In principle, we can have any number of stages; and then we refer to multi-stage sampling. For example, we could select islands at the first stage, villages within each selected island at the second stage, boats within those villages at the third stage, and certain days of the month on which to collect data at the fourth stage.

In practice it is probable that some combination of multi-stage (or at least two-stage) and stratified sampling will prove to be the most efficient and cost-effective system we can devise.

6.4.5 Sampling with probability proportional to size

We will not attempt to examine more sophisticated designs in this course, but will mention one particular technique, namely, sampling with probability proportional to size (often referred to as 'p.p.s.' sampling), because this technique is very widely used, especially in the first stages of multi-stage sampling. With p.p.s. sampling, instead of giving each unit an equal chance of selection, we adopt procedures which give larger units a greater chance of selection than smaller units.

We will use an illustration to show how to make selection with probability proportional to size. Suppose in a multi-stage sample we wish to select one of five villages in a district, with the intention of selecting certain fishing units within the selected village at the next stage. Using the selection methods we have described so far we would simply take a random number between 1 and 5 to choose the village.

With p.p.s. sampling, we would need to know the population (or the number of households, or some other measure which is suitable for use as a measure of size) of each village. Even if we do not have a precise measure of size, there are often records available which will be adequate. For example, data from the last census would give us the comparative populations at that time, and would be good enough to be used as estimates of the current population. When we use estimated measures of size we may refer to p.p.e.s. (i.e. probability proportion to estimated size) sampling. We would make the village selection in the following manner:

Village	Population	Cumulative population	Selection range
A	226	226	1- 226
B	705	931	227- 931
C	339	1,270	932-1,270
D	104	1,374	1,271-1,374
E	295	1,669	1,375-1,669

The total population of five villages is 1,669, so we would choose a random number between 1 and 1,669. Any number between 1 and 226 would select village A, and so on, according to the figures shown in the column 'Selection range' above. With this system we are far more likely to choose the largest village, B, than the smallest village, D, because there are 705 different random numbers (from 227 to 931) which would select B, and only 104 numbers (from 1,271 to 1,374) which would select D. In fact the chances of selection are exactly proportional to the population of the villages, as shown.

We should note here that the measure of 'size' which we really wanted is presumably the catch of fish, since that is the characteristic we are trying to measure. We would really like to give probability of selection of villages in proportion to the amount of fish they catch, but that is presumably not known. However, in the case of subsistence fishing, it may be reasonable to expect that fish catch would be roughly proportional to population (at least within a certain geographic area), so we can use population figures as a useful substitute for fish catch as the basis on which to make p.p.s. selections.

The use of p.p.s. sampling, at the first or second stages of multi-stage sampling, is very common. When several units are being selected at once (e.g. if we are selecting 4 villages out of 50), it is the usual practice to make selections <u>with</u> replacement, thus giving a village the chance of being doubly-selected. This is slightly less efficient than selection without replacement - that is to say, it has slightly higher sampling errors, but it makes the estimation process and the calculation of standard error easier.

Although in p.p.s. sampling the probability of selecting sampling units at this stage is unequal, we can still make subsequent selections in such a way that the probability of selection for any final-stage sampling unit is exactly equal, if we wish.

6.5 <u>Principles of calculating standard error</u>

We will introduce this discussion by referring to the calculation for simple random samples, where the population is infinite (e.g. the total number of fish in the sea).

Earlier in this topic, in introducing the concept of sampling error, we noted that if we took a number of separate samples, and calculated $\bar{x}$ in each case, we would finish up with a <u>distribution</u> of values of $\bar{x}$, each of which is a separate estimate of μ. If we were to repeat this process a very large number of times (which, fortunately we will not have to do in practice) we would obtain the theoretical distribution of $\bar{x}$, and we could calculate the standard deviation of this distribution.

We will give the result of this as a theorem, i.e. "If random samples of size n are taken from an infinite population, the theoretical distribution of $\bar{x}$ has a standard deviation of $\sigma/\sqrt{n}$".

We write this as $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and we refer to the measure $\sigma_{\bar{x}}$ as the <u>standard error of the mean</u>. To avoid confusion in use of the symbol σ, it is more usual to write $Se(\bar{x})$ to denote the standard error of the mean. The standard error of the mean plays a very important role in statistics, because it measures the variation of the theoretical sampling distribution of $\bar{x}$. In other words, it tells us how much sample means can be expected to vary from sample to sample. We can see that, since the divisor is $\sqrt{n}$, the

standard error of the mean will decrease as we increase the sample size. So the larger we make n, the more reliable will our $\bar{x}$ be as an estimator of μ.

Of course all this is theoretical, and in practice we do not have a large number of samples, each giving a calculation of $\bar{x}$; normally we have one sample only. More importantly, we cannot ever know the value of σ (the standard deviation of the population) of an infinite population. Therefore, we need to modify our formula by replacing σ by an estimate of σ. Fortunately we can do this, because we can calculate s, the standard deviation of the sample. Provided the sample is random and unbiased and the sample size is significantly large, we can expect s to approximate quite closely to σ. However, this may not hold for small samples, for large samples then, we estimate the standard error of the mean to be equal to $s/\sqrt{n}$.

For example, s can be calculated for our yellowfin data as 1.37 kg. Now we can say that from our sample we estimate the mean weight of the total population of fish to be 4.43 kg and that we estimate the standard error of this to be $\frac{1.37}{\sqrt{63}} = 0.17$ kg.

6.5.1 The finite population correction factor

So far we have been discussing samples drawn from infinite populations. When we are sampling without replacement from finite populations we have to make an adjustment to our formula, and this becomes the following:

$$Se(\bar{x}) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$$

$\sqrt{\frac{N-n}{N-1}}$ In other words we have multiplied our previous formula by the factor This is known as the 'finite population correction factor' (or f.p.c. for short). The f.p.c. serves to reduce the standard error of our estimate from the value it would have had if we had been dealing with an infinite population, or if we had used sampling with replacement. To demonstrate the effect which this factor has on our estimate of the standard error of the mean, let us consider what the value of this factor would be if we had (a) a sample of 200 observations out of a total population of 40,000 and (b) a sample of 200 observations out of a population of 400 units.

In the first case f.p.c. would be equal to:

$$\sqrt{\frac{40{,}000 - 200}{40{,}000 - 1}} = \sqrt{\frac{39{,}800}{39{,}999}} = 0.998$$

This is so close to unity that multiplying by it will have virtually no effect on the answer we obtain. In practice, where the sampling fraction is small (say less than 5%), we can ignore the f.p.c. altogether, and we can say that for very large finite populations, or with very small sampling fractions, we will treat $Se(\bar{x})$ as being the same as for our infinite population.

However, in the second case, this factor equals:

$$\sqrt{\frac{400 - 200}{400 - 1}} = \sqrt{\frac{200}{399}} = 0.701$$

Thus, multiplication of our unadjusted calculation by this factor will reduce our estimate of the standard error of the mean by almost 30 per cent, and this is so significant that it certainly cannot be ignored.

In practice it is usual to modify this correction factor slightly. We can note that

$$\sqrt{\frac{N - n}{N - 1}} = \sqrt{\frac{N - 1 + 1 - n}{N - 1}} = \sqrt{1 - \frac{n - 1}{N - 1}}$$

and this is almost equal to

$$\sqrt{1 - \frac{n}{N}}$$

or in other words the square root of one minus the sampling fraction. If we also replace σ by s in our formula (as we did before), because we normally will not know the value of σ, our formula for the estimate of the standard error of the mean can be written as:

$$Se(\bar{x}) = \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

We will find that this is the most usual presentation of the formula for practical purposes.

6.5.2 Confidence intervals

We must next ask ourselves what this really means. It is all very well to say that we estimate the mean weight of all fish in the area as 4.43 kg with a standard error of 0.17 kg, but such a statement on its own will have limited value. Users of the statistics will probably understand perfectly well what the mean of 4.43 kg denotes, but how are they to interpret the value of the standard error?

Fortunately, as we noted earlier, the distribution of calculations of $\bar{x}$ will approximate very closely to a Normal curve, and this will hold true even if the population itself was not distributed normally. It follows that the properties of the standard deviation in respect to the normal distribution, which we mentioned in Topic 4, will apply. Therefore we can say that about 68 per cent of all sample estimates will lie within one standard error either side of the mean, and over 95 per cent will lie within two standard errors.

Of course we usually have only one single sample estimate, not a large number of them, so we need to put our statement into a different form to make it more useful. In practice we say that, provided $\bar{x}$ is an unbiased estimator of the population mean, there is a probability of about 68 per cent (or that we are "68% confident") that the sample mean plus or minus one standard error will include the population mean, and over 95 per cent probability that the sample mean plus or minus two standard errors will do

so. If we assume that our 63 yellowfin comprised a sample representing the total yellowfin population of the area, then there is a 68 per cent probability that the mean weight of all yellowfin in the area is within the range (4.43-0.17) kg and (4.43+0.17) kg, i.e. between 4.26 and 4.60 kg, and that there is over 95 per cent probability that it is within the range 4.09 to 4.77 kg. These ranges are referred to as confidence intervals, and the different probabilities (i.e. 68%, 95%, etc.) are called confidence levels.

Statisticians make most use of the 95 per cent confidence levels in practice, because this is a high enough figure for us to be "fairly sure" of being correct. Thus for the fish data, we interpret our results to mean that we are 95 per cent confident that the true mean weight of the fish lies somewhere in the range 4.09 to 4.77 kg. However, we can never be really sure, and we must never assert that the true mean is within this confidence interval.

If it were decided that for some purposes a 95 per cent confidence level is inadequate, we can make a similar calculation for other levels. For example, the 99 per cent confidence level pertains to an interval of 2.6 times the standard error on either side of the sample mean. So we could say we are confident at the 99 per cent level that the true mean weight of the fish is between (4.43 - 2.6x0.17) and (4.43 + 2.6x0.17) kg, i.e. between 3.99 and 4.87 kg. A similar calculation can be made for any desired level of confidence, and we can look up special tables to find out the appropriate confidence interval for any level.

6.6 Principles for estimating population parameters from sample data

Next we must look at the techniques for making the estimates themselves - that is to say, the way in which we 'expand' the data obtained from the sample to obtain estimates for the whole population. This is in fact a very difficult subject and for surveys with a complex design, involving stratification, multi-stage selection, and probability proportional to size, the computation can become so complicated that processing will almost certainly have to be undertaken on a computer. However, the principles in making unbiased estimations still apply, even though the applications may be difficult.

The estimators commonly used for estimating population parameters are of the form

$$\hat{X} = \sum_{i=1}^{n} w_i x_i$$

where $\hat{X}$ (pronounced X hat) represents the estimate of the characteristic for the population X; x_i is the value of this characteristic for the ith selected 'final-stage' sampling unit; and w_i is the 'weight' applicable to that unit. It is this 'weight' - which is variously referred to as the multiplier, the expansion factor or the raising factor - which provides the main difficulty in the estimation phase.

Incidentally, it will be noted that we referred to x_i as the value for the 'final-stage' sampling unit; it is necessary to have this qualification because in multi-stage sampling we can have different sampling units at different stages, e.g. the village at the first stage and the fishing unit at the second stage, so we have to be careful to define just what we mean by 'sampling unit'.

For the simplest type of samples, where all units have an equal chance of selection, the expansion from sample data to population estimates is quite straightforward. In these circumstances, e.g. in simple random sampling or simple systematic sampling, the multiplier is the same for all selected units and is equal to N/n, which is the inverse of the sampling fraction - sometimes referred to as the sampling interval. All that is needed is to 'raise' sample values by this factor, to obtain population estimates, i.e. $\hat{X} = \frac{N}{n} \Sigma x$.

6.7 Population estimates and sampling error for the main types of sample

We are now in a position to look at the formulae whereby we estimate population parameters and calculate sampling errors, for the main types of sample we have been discussing.

6.7.1 Simple random sample

The notation we will use is:

N = Number of sampling units in population
n = Number of sampling units in the sample
X = Population value
$\hat{X}$ = Estimate of population value
$\hat{\bar{X}}$ = Estimate of population mean

(We could also write this, as $\hat{\mu}$, since μ was our symbol for the population mean, but in practice it is more common to write the estimate as $\hat{\bar{X}}$, pronounced X bar hat.)

$Se(\hat{X})$ = Standard error of population estimate
$Se(\hat{\bar{X}})$ = Standard error of estimate of population mean

We then have, for estimates of the population mean:

$\hat{\bar{X}} = \bar{x}$, or $\frac{\Sigma x}{n}$ — In other words we simply estimate the population mean to be equal to the sample mean.

$$Se(\hat{\bar{X}}) = \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

and for estimates of the total population :

$\hat{X} = N\bar{x}$, or $\frac{N}{n} \Sigma x$ — That is, we estimate the value of the population total to be the value of the sample total, multiplied by the sampling interval.

$$Se(\hat{X}) = N \cdot \frac{S}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

For large and infinite populations, the f.p.c. can be omitted. So we have

$$Se(\hat{\bar{X}}) = \frac{S}{\sqrt{n}} \quad \text{and} \quad Se(\hat{X}) = N \cdot \frac{S}{\sqrt{n}}$$

6.7.2 Stratified random sample

k = Number of strata

N_c = Number of sampling units in stratum c, so we have $N = \sum_{c=1}^{k} N_c$

w_c = N_c/N, i.e. the weight, or proportion, of total sampling units in stratum c

n_c = Sample size in stratum c

$\hat{\bar{X}}_c$ = Estimate of the population mean in stratum c

$\bar{x}_c$ = Sample mean in stratum c

Then $$\hat{\bar{X}} = \frac{1}{N} \sum_{c=1}^{k} N_c \, \bar{x}_c$$

or, perhaps more commonly this would be written

$$\hat{\bar{X}} = \sum_{c=1}^{k} w_c \, \bar{x}_c$$

That is to say, we have a weighted mean, with the means of each individual stratum weighted by the proportion of the total number of sampling units in that stratum.

The formula for the standard error is given by :

$$Se(\hat{\bar{X}}) = \frac{1}{N}\sqrt{\sum_{c=1}^{k} \frac{N_c (N_c - n_c)}{n_c} s_c^2} \quad \text{or} \quad \frac{1}{N}\sqrt{\sum_{c=1}^{k}\left(\frac{N_c^2 s_c^2}{n_c}(1 - \frac{n_c}{N_c})\right)}$$

This may be written, in terms of the weights, w_c, as

$$Se(\hat{\bar{X}}) = \sqrt{\sum_{c=1}^{k}\left(\frac{w_c^2 \cdot s_c^2}{n_c}(1 - \frac{n_c}{N_c})\right)}$$

In this formula the f.p.c. $(1 - n_c/N_c)$ can be omitted if the sampling fraction is small (e.g. less than 5%) for every stratum. In practice in stratified samples we quite often need to have fairly large sampling fractions for some strata, and the f.p.c. is an important component in the calculation of the standard error. For example, if there are a few large boats and many canoes in a local fishery, we would endeavour to stratify by type of boat, and would probably include a fairly high percentage of the 'large boat' stratum in the survey.

Let us look at an example where we want to estimate the average fish catch on a certain day from a population of 350 boats, and that we have resources to collect data from 50 boats. We may take a stratified sample by type of boat, and get the following results.

Type of boat	Total No. of Boats (N_c)	Sample No. of Boats (n_c)	Survey Results: Est. Av. Catch ($\bar{x}_c$)	Survey Results: Est. Standard Deviation (S_c)
Canoes	278	28	18	7
Small power boats	56	14	32	10
Large power boats	16	8	112	36
Total	350	50		

To estimate average catch we simply substitute in this formula, and we have

$$\hat{\bar{X}} = \frac{1}{350} \sum_{c=1}^{3} N_c \, x_c$$

$$= \frac{1}{350} \quad (278 \times 18 \quad + \quad 56 \times 32 \quad + \quad 16 \times 112)$$

$$= \frac{8588}{350} \qquad = \quad 24.5 \text{ kg}$$

This is a simple weighted mean. The figure of 8588 kg of course represents the total catch of all boats that day.

In similar fashion we can substitute in our formula to calculate the standard error of our estimate, as follows:

$$Se(\hat{\bar{X}}) = \frac{1}{350} \sqrt{\sum_{c=1}^{3} \left(\frac{N_c \ (N_c - n_c)}{n_c} \ S_c^2 \right)}$$

We must note that the sampling fraction is above five per cent in each stratum, so we certainly cannot ignore the f.p.c. We have

$$Se(\hat{\bar{X}}) = \frac{1}{350} \sqrt{\frac{278(250)}{28} \cdot 7^2 \quad + \quad \frac{56(42)}{14} \cdot 10^2 \quad + \quad \frac{16(8)}{8} \cdot 36^2}$$

$$= 1.14$$

So our estimate from the stratified sample is a catch of 24.5 kg per boat, with a standard error of 1.14 kg. In other words we are 95 per cent confident that the true average catch was 24.5 $\pm$ 2.3 kg, i.e. in the range 22.2 to 26.8 kg.

In is worth noting here that the predominant part of the total standard error arose from the first stratum, viz canoes. This may give us a clue that, if we were undertaking another such survey, we might reduce sampling error by increasing the sample size of this stratum. This idea will be explored in more detail in section 6.9.

6.7.3 Multi-stage sampling

Estimation of population parameters for multi-stage samples becomes more complicated because we have to use different weights, or expansion factors, at each stage of the sampling process. We have to build up our estimates at one stage before we can go on to estimating at the next stage.

The situation will be different depending on whether selection at the first stage is made with probability proportional to size or not, and whether first stage selections were made with or without replacement. We will give formulae here only for the simplest situation, but in practice it is more likely that p.p.s. would be used.

Suppose we wish to estimate the total landings of fish along a section of coastline by sampling the landings from the fishing fleet. But we know that there are a number of landing places and many vessels fishing along the coast, and we cannot visit all places. In this case we could resort to Two-Stage Sampling. First, we select at random a convenient number of landing sites from the total sites available along the coast, e.g. suppose there are 8 sites and we select 3 of these. Then:

N = Number of 1st stage units = 8
n = Number of 1st stage samples = 3

Next we select at each of these 3 sites a convenient number of boats from the total number of boats landing at these ports. Then:

M_i = Number of 2nd stage units available in 1st stage unit 'i'
m_i = Number of 2nd stage units sampled from those available in 1st stage unit 'i'

We will assume our data is as follows :

Landing site (n_i)	1	2	3
Number of boats present (M_i)	6	9	7
Number of boats sampled (m_i)	3	3	3
Landings (tonnes)	13	5	12
	9	7	8
	6	10	13
Total landings by sample vessels	28	22	33
S_i	5.3	4.7	2.6

The notation we will use is:

$\frac{n}{N}$ = 1st stage sampling fraction

$\frac{m_i}{M_i}$ = 2nd stage sampling fraction for the ith landing port

Then, if (y_{ij}) is the landing of a particular vessel, i.e. the jth vessel at the ith port, we have

$\bar{y}_i$ = Average landing per vessel at site$_i$ = $\frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$

$\hat{y}_i$ = Total estimated landing at site$_i$ = $M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij}$

$\hat{y}$ = Total estimated landings for all selected sites = $\sum_{i=1}^{n} \left(M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \right)$

$\hat{Y}$ = Total estimated landing for entire coast = $\frac{N}{n} \left[\sum_{i=1}^{n} \left(M_i \cdot \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} \right) \right]$

Estimation of the standard error of this total estimate is a little more complicated, because we have two sources of variation in our calculation. Firstly, we have the variation in landings by the vessels at any landing site, and then we have the variation in landings <u>between</u> the landing sites. Our total sampling error is therefore the sum of these two factors.

It is for this reason that we said earlier that the multi-stage sampling is only achieved at the expense of some loss in accuracy of our results. We have saved in costs and ease of data collection by concentrating all our efforts on 3 out of 8 landing sites, instead of taking observations at all of them, as we would in other types of sample design. Now we have to pay the penalty for using this grouping, or 'clustering' of observations, by making allowance for the variation between landing sites.

Our formula for the variance, i.e. the square of the standard error, becomes

$$V(\hat{Y}) = \frac{N(N-n)}{n} \frac{1}{n-1} \left(\sum_{i=1}^{n} \hat{y}_i^2 - \frac{\left(\sum_{i=1}^{n} \hat{y}_i\right)^2}{n} \right) + \frac{N}{n} \sum_{i=1}^{m_i} \frac{M_i(M_i-m_i)}{m_i} \cdot S_i^2$$

where $S_i = \frac{1}{m_i-1} \left(\sum_{j=1}^{m_i} y_{ij}^2 - \frac{\left(\sum_{j=1}^{m_i} y_{ij}\right)^2}{m_i} \right)$

Substituting the data from our example in this formula we have

$$V(\hat{Y}) = \left(\frac{8x5}{3} \times \frac{1}{2} \times 221\right) + \frac{8}{3} \left[\left(\frac{6x3}{3} \times 12.3\right) + \left(\frac{9x6}{3} \times 6.3\right) + \left(\frac{7x4}{3} \times 7.0\right) \right]$$

$$= 1473 + 673 = 2146$$

The first term in this expression represents the variation between landing sites. It will be noted that the contribution to variance due to this term is much greater than that which is due to difference among second-stage units within the first-stage units. This means that if we were going to carry out the exercise again, it would be preferable to increase the number of fishing sites sampled, even if it meant that we had to reduce the number of vessels sampled at each site.

We have in our example $Se(y) = \sqrt{2146} = 46.3$, therefore

95% confidence limits are $531 \pm 2\ (46.3)$
$= 531 \pm 92.6$

i.e. our estimate of the total landings, with 95 per cent confidence, is between 438 and 624 tonnes.

6.8 Ratio estimation

Ratios of population totals of two characteristics are as important as, and sometimes more important than, the population totals themselves. For example, we may obtain from a sample survey information on total catch and on total effort, and we may be more interested in estimating catch per unit effort than we are in estimating either of the totals. For surveys covering two different points of time, we may be more concerned with finding out whether total catches have gone up or down, than with measuring the level at any one point of time.

We use the term ratio estimations to refer to the method of estimating a ratio of the population by means of a ratio of the unbiased estimators of two characteristics. Thus, if $\hat{Y}$ and $\hat{X}$ are unbiased estimators of Y and X respectively, then an estimator of the population ratio $R=Y/X$ is given by the ratio estimator $\hat{Y}/\hat{X} \rightarrow R$.

In situations where the actual population value for the denominator (X) is known, it might be felt that to estimate the desired population ratio (R) all that is necessary would be to estimate the numerator (Y); thus $\hat{Y}/X \rightarrow R$. However, if the estimators of the numerator and the denominator are approximately proportional (that is, if the two characteristics are highly linearly related with the line passing through the origin), then an estimator based on the ratio of the estimators of the numerator and the denominator is a more efficient method.

The method has a possible application in fisheries statistics, if a country manages to conduct a complete census of all local fisheries to measure total catch in one year, and then wishes to monitor changes to the total catch in following years by a sample survey of fishing units.

If we start by discussing a simple random sample, the approach we have discussed so far would be to select n fishing units out of the total of N units in the population, measure the catch, y, of each sample unit, and estimate total fish catch as $\hat{Y}=\frac{n}{N}\Sigma y$.

In the ratio estimate, we would ascertain the catch by each of the n fishing units in the present survey, and also the catch which those same units obtained in the census year.

We use the notation here

X = total catch by all fishing units in the census year
y = catch by sample units in the present year
x = catch by those same units in the census year
$\hat{Y}$ = estimated total catch in the present year

Then we could say

$$\hat{Y} = \frac{\Sigma y}{\Sigma x} \cdot X$$

This is a ratio method of estimating total catch by measuring for each sample unit the ratio of catch between two different periods.

To put this in simple terms, we might come to a conclusion from our survey as follows: The total catch at last year's census was 420 tonnes. The catch by our sample boats has increased by five per cent since then, so we will estimate that the catch by all boats has increased by five per cent. Therefore, we estimate the catch this year to be 420+21 = 441 tonnes. This is the line of reasoning we follow in ratio estimation: we calculate some ratio derived from a sample, in order to estimate population parameters.

If we examine the estimator $\frac{\Sigma y}{\Sigma x}.X$, it is clear that X is not derived from a sample, so the sampling error will depend solely on the sampling error of the ratio $\frac{\Sigma y}{\Sigma x}$, with X having only the effect of a constant multiplier.

The formula can be modified to give a different weight or expansion factor to the selected units. It becomes

$$\hat{Y} = \left(\frac{\sum_{i=1}^{n} w_i \, y_i}{\sum_{i=1}^{n} w_i \, x_i} \right) \cdot X$$

where w_i is the weight, or expansion factor of the ith unit.

The actual formula for calculating standard errors for ratio estimates is complicated, and is beyond the scope of this course. We will simply observe that in some circumstances the ratio method of estimation leads to substantially lower standard errors. However, it is a biased method. Fortunately the bias tends to be negligible for moderately large samples. In many practical applications indeed, it is so small compared with the advantage gained in reducing the sampling error, that the ratio estimate is preferred to the unbiased estimate.

6.9 Determining sample size

A very important part of planning a sample survey is deciding how many units to sample, i.e. the size of the sample. The size of the sample will depend upon the resources we have available, i.e. the number of trained collectors, money, data analysis facilities, time, and on the degree of precision we need in the results. If we require only approximate estimates, a small sample may suffice, but if we require more exact estimates, a large sample may be needed.

In order to determine the size of a sample we generally need the following kinds of information:

(a) the total resources (money, manpower, etc.) available for the investigation;

(b) the cost of collecting data from one unit;

(c) the expected variability in the population;

(d) the required precision.

It is unlikely that we will ever be able to have exact information on any of these, but we can often use approximations, estimates and data from previous surveys in order to gain some idea of the size of the sample we need. In some cases, results from a pilot survey will provide estimates of costs and also give an idea of variability. Assuming, therefore, that we have some idea of this kind of information we shall look at different types of sample to to see how we can determine the sample size.

6.9.1 Simple random sample

Let us start by looking at a simple example. Suppose we wish to estimate the average per capita fish consumption per week from a simple random sample of people. We would like to have 95 per cent confidence that our estimate will be within plus or minus 0.2 kg per week. From last year's study we may have an estimate that the standard deviation is about 0.5 kg. Now we know that the range $\hat{\bar{X}}-2Se(\hat{\bar{X}})$ to $\hat{\bar{X}}+2Se(\hat{\bar{X}})$ equates to our 95 per cent confidence limits. For a simple random sample, we have approximately $Se(\hat{\bar{X}})=s/\sqrt{n}$ where s is the estimated population standard deviation. Therefore, we have

$$2Se(\hat{\bar{X}}) = 0.2$$

and $$Se(\hat{\bar{X}}) = \sqrt{(0.5^2/n)}$$

Thus $$\sqrt{(0.5^2)/n} = 0.2/2$$

which gives $n = 0.25/0.01 = 25$, i.e. we need to sample about 25 people.

Quite often, instead of specifying a tolerance value (such as the 0.2 kg above), we say that we want the true result to be within a certain percentage of our estimate. For example, we might want to have 95 per cent confidence that the true value will be within the range $\hat{\bar{X}} \pm 5\%\hat{\bar{X}}$. In general, we can write this as $\hat{\bar{X}} \pm p\hat{\bar{X}}$ and we have to specify p.

We know that

$$2Se(\hat{\bar{X}}) = p\hat{\bar{X}} \qquad \text{and} \qquad Se(\hat{\bar{X}}) = \sqrt{\left[(1-\frac{n}{N})\ \frac{S^2}{n}\right]}$$

If we solve these two equations for n, we have

$$n = \left(\frac{2s}{p\hat{\bar{X}}}\right)^2 \cdot \frac{1}{1 + \frac{1}{N}\left(\frac{2s}{p\hat{\bar{X}}}\right)^2}$$

To see how this works, we shall use the following example: N=430; we know from previous studies that $\hat{\bar{X}}=19$ and $s^2=85.6$ and we specify p=0.10 (or 10%).

Then we have

$$n = \frac{2^2 \times 85.6}{(0.10 \times 19)^2} \cdot \frac{1}{1 + \frac{1}{430} \frac{2^2 \times 85.6}{(0.10 \times 19)^2}}$$

$$= 78$$

If we needed a precision of one per cent instead of ten per cent, we would have:

$$n = \frac{4 \times 85.6}{(0.01 \times 19)^2} \cdot \frac{1}{1 + \frac{1}{430} \frac{2^2 \times 85.6}{(0.01 \times 19)^2}}$$

$$= 411$$

that is, we would have to sample nearly the entire population. In practice, we would not take a sample, but rather a census, if the required sample size is calculated to be as close to the total population as that.

We note that the second part of the expression for n, $\frac{1}{1 + \frac{1}{n}\left(\frac{2s}{\hat{\bar{x}}}\right)^2}$ is the finite population correction factor. If $\frac{1}{n}\left(\frac{2s}{\hat{\bar{x}}}\right)^2$ is less than 0.05, then we can safely approximate n by the expression $\left(\frac{2s}{\hat{\bar{x}}}\right)^2$ because then the term $\frac{1}{1 + \frac{1}{n}\left(\frac{2s}{\hat{\bar{x}}}\right)^2}$ is very close to 1.

6.9.2 Stratified sample

In stratified random sampling we have to decide both the total sample size and how to allocate sample size in each stratum. We have four main methods for choosing overall sample size and strata sample sizes. Which one we use depends on how much prior information we have about the variability in the population and strata and also on the costs of sampling each unit. We will examine each of the four methods.

We introduce the symbols

C_c = the cost of sampling one unit in the cth stratum

d = maximum acceptable error (such as the 0.2 kg used in our example for simple random sampling)

z = a variable whose value depends on the degree of precision required, expresed in number of standard deviations from the mean. If we want confidence limits of 95 per cent, we use the value 2 for z, or if we want confidence limits of only 68 per cent, we use the value 1. Other values of z, for different confidence limits, can be found in tables of the Normal probability distribution.

Equal allocation

In equal allocation we take the same number of samples from each stratum. Therefore we have only to determine the overall sample size n, and we sample $n_c = n/k$ units from each stratum.

The formula for calculating n is:

$$n = \frac{k \sum N_c^2 s_c^2}{N^2 \frac{d^2}{z^2} + \sum N_c s_c^2}$$

We use the method of equal allocation in the following situations:

(i) When the total numbers of sample units N_c in each of the k strata are more or less equal;

(ii) When the stratum variances (s_c^2) and cost per sampling unit (C_c) do not vary much from stratum to stratum;

(iii) When there is no prior knowledge of stratum variances (s_c^2) or cost per sampling unit (C_c).

Proportional allocation

The total sample size is allocated among strata in proportion to the size of each stratum. For example, if stratum 3 contains 25 per cent of the population, then 25 per cent of the overall sample will be taken in stratum 3. The formula for calculating the allocation per stratum is

$$n_c = \frac{N_c}{N} \cdot n = w_c \cdot n$$

To calculate the overall sample size, n,

$$n = \frac{N \sum N_c s_c^2}{N^2 \frac{d^2}{z^2} + \sum N_c s_c^2}$$

We use proportional allocation when the stratum total number of units, N_c, varies from stratum to stratum, and when either the stratum variances and cost per sampling unit do not vary much from stratum to stratum or when we do not have any prior knowledge about stratum variances and costs.

Neyman allocation

The Neyman method is named after the famous statistician who developed the method. We use the Neyman method when the stratum variances s_c^2 vary from stratum to stratum. The formula for allocation of sample size in stratum c is

$$n_c = \frac{N_c s_c}{\sum N_c s_c} \cdot n$$

The overall sample size, n, is calculated thus :

$$n = \frac{(\sum N_c\ s_c)^2}{N^2 \frac{d^2}{z^2} + \sum N_c\ s_c^2}$$

Optimum allocation

To use optimum allocation we need some prior knowledge of the stratum variances and cost per sampling unit in each stratum. If both stratum variance, s_c^2, and cost per sampling unit, C_c, vary from stratum to stratum, then we will obtain the greatest precision in our estimates if we use the formulae for optimal allocation. The sample size in each stratum is

$$n_c = \frac{N_c\ s_c}{\sqrt{C_c}} \cdot \frac{1}{\sum \frac{N_c\ s_c}{\sqrt{C_c}}} \cdot n$$

and the formula for overall sample size n is

$$n = \frac{(\sum N_c\ s_c\ \sqrt{C_c}) \cdot (\sum \frac{N_c\ s_c}{\sqrt{C_c}})}{N^2 \frac{d^2}{z^2} + \sum N_c\ s_c^2}$$

6.9.3 An example of sample size allocation

To see how we may apply the four methods in practice, we will calculate sample sizes using each method for the following problem.

Suppose that along a certain coast, the 100 places where fish are landed can be roughly graded into three classes according to the weight of fish landed. During a typical week, the weights landed are

Large landing places : 45, 59, 87, 41, 71, 25, 9, 69, 10, 7

Medium landing places: 17, 13, 19, 26, 1, 8, 27, 11, 12, 26
5, 8, 10, 16, 16, 4, 16, 16, 13, 29
14, 25, 29, 27, 20, 25, 2, 7, 3, 12

Small landing places : 2, 6, 7, 0, 1, 2, 1, 5, 4, 7
8, 9, 3, 2, 5, 4, 2, 0, 2, 8
5, 3, 8, 9, 8, 9, 1, 6, 5, 3
3, 4, 7, 5, 5, 3, 2, 4, 6, 1
6, 2, 5, 1, 0, 3, 8, 0, 4, 3
3, 5, 5, 0, 7, 0, 9, 7, 9, 0

Calculations on the complete census of weights landed show the following

		N_c	s_c	$\bar{X}_c$
c=1:	Large landing places	10	28.91	42.30
c=2:	Medium landing places	30	8.57	15.23
c=3:	Small landing places	60	2.81	4.20

In our formulae, we will use k=3, since landing places are divided into three strata. We want to have 95 per cent confidence that our final estimate from the stratified random sample will be within two units of the true total landed weight, therefore d=2 and z=2.

If we suppose for the purposes of optimal allocation that each unit in strata 1 and 2 (large and medium landing places) costs \$10.00 to sample, and each unit in stratum 3, the small landing places, costs \$20.00 to sample, then C_1=10, C_2=10, C_3=20. Of course, with different costs we will get different sample size allocations from those worked out in this example.

We now have all the information we need for the formulae to find n, n_1, n_2, and n_3 for each of the allocation methods. We will not show the calculations but the results are given in Table 6.2.

TABLE 6.2 : TOTAL AND STRATA SAMPLE SIZE FOR FISH LANDINGS USING FOUR DIFFERENT METHODS OF ALLOCATION

Method	n_1	n_2	n_3	n
Equal	8	8	8	24
Proportional	5	16	32	53
Neyman	9	9	6	24
Optimal	10	10	5	25

The equal and proportional methods do not take into account the differences in standard error among strata. In the present example, the differences are large and, in proportional allocation, the overall sample size is more than twice that of all other methods. The large sample size is required because no weighting is given to sampling from the strata with the largest standard errors and so the overall sample size has to be increased to obtain the required degree of precision. Equal allocation does not require an exceptionally large sample size because rather large samples are taken from the strata with smallest numbers (N_c), but greatest standard errors, i.e. strata 1 and 2.

Neyman allocation takes into account the standard errors of the strata. We see that 9 units out of 10 should be sampled from the large landing sites because of the large standard error in this strata, 9 out of 30 in the medium landing sites and only 6 out of 60 in the small landing sites where the standard error is small. The particular example of optimal allocation given here produces a similar sample allocation to that of the Neyman method, except that the relatively high cost of sampling the small landing places reduced the sample size in stratum 3 and slightly increased that of the other 2 strata. Different cost values would give different allocations.

In practice, where the calculations of sample size indicate that 8 or 9 out of 10 units should be sampled in a stratum, we would not sample but take a census of the stratum. Our overall estimated total landings would then have sampling errors due only to the strata which were sampled. Another example of a case where certain strata are completely enumerated and others are sampled only is the estimation of tuna catch. We try to get a census of large-scale foreign and domestic fishing vessel catches but we usually have to be content with sample estimates only of small-scale local catches.

In general, the recommended method of allocation is that method which uses the maximum information available. Therefore, if we have some estimates of the standard errors and per unit sampling costs in each stratum, we should use optimum allocation. If we have no idea of sampling costs, but we do have estimates of standard errors, then Neyman allocation is recommended. Equal and proportional allocation are used when costs and standard errors are not known. In the landing place example, equal allocation produced quite a reasonable allocation because of the inverse relationship between size of stratum and standard error. In other cases where the relationship is different, e.g. where standard error is proportional to stratum size or where no relationship exists, equal allocation will not be as good as proportional allocation.

6.9.4 Some conclusions about sample size

We usually undertake a sample survey because the costs of a complete census are likely to be too great. Usually money and other resources are very limited, and if we are undertaking any kind of statistical investigation we have to make sure that it is carried out as efficiently as possible. In this section we have seen how, for different types of samples, we can decide how large a sample we need for different purposes. The various formulae are just techniques for using prior information for better planning. Very often the information we have is vague, but these formulae can still provide a means for using this.

We have been looking at the problem of deciding on the best sample size by using information about just one variable. The problem is that in most of the surveys we undertake we are interested in collecting data on several characteristics. We can make the sample design optimal for one of them, but this does not guarantee that it is optimal for the rest. In practice, there is not a great deal we can do about this; it would obviously be impossible to choose a different sample for every variable. If we have sufficient information for several variables, then we can try to find a sample design that is almost optimal for all of them. This will be a compromise for each variable, but, otherwise, all we can do is base the sample on the most important variable and hope that it is not too inefficient for the others.

6.10 Concluding remarks

One final point on the whole topic of sampling needs to be made, and it is a very important one. All our calculations and formulae in this topic have been based on the assumption that the sample is random and unbiased. When we claim 95 per cent confidence that the true population mean is within two Se either side of $\bar{x}$, we have implicitly assumed that $\bar{x}$ is an unbiased estimator of X.

We discussed bias earlier in this topic, and we know that in some surveys there are substantial biases which we cannot eliminate. In the rather unlikely event that we are able to measure the bias, we can still quote confidence limits. Thus, if a bias B exists, i.e. if we know that the amount of displacement of the distribution of $\bar{x}$ away from the true position of the population mean is +B, then our 95 per cent confidence limits become $(\bar{x} - B - 2Se)$ to $(\bar{x} - B + 2Se)$.

However, in the great majority of situations we will not have any measure of bias, even if we are aware that it exists. It follows that if we know or suspect that there are substantial non-sampling errors in a

survey, it is a dangerous and misleading practice to express the results in terms of the mean and standard error, without any other qualification. The very fact that we publish a figure for standard error is likely to lead users of the statistics into believing that this is an accurate portrayal of the extent of errors in the results, and they are likely to assume that no other errors are present, unless we make it clear that this is not so.

USEFUL REFERENCE MATERIAL

COCHRAN, W.G. (1963). Sampling techniques. John Wiley & Sons, Inc., New York, N.Y. 413 p. --best general reference--

CUFF, W. & N. COLEMAN (1979). Optimal survey design: lessons from a stratified random sample of macrobenthos. J. Fish. Res. Board Can. 36:351-361.

JESSEN, R.J. (1978). Statistical survey techniques. John Wiley & Sons, Inc., New York, N.Y. 520 p.

RAJ, D. (1968). Sampling theory. McGraw-Hill Book Co., New York, N.Y. 302 p. ---very good general reference---

SAILA, S.B., R.A. PIKANOWSKI & D.S. VAUGHAN (1976). Optimum allocation stategies for sampling benthos in the New York Bight. Est. Coast Mar. Sci. 4:119-128.

SCHWEIGERT, J.F. & J.R. SIBERT (1983). Optimizing survey design for determining age structure of fish stocks: an example from British Columbia herring (Clupea harengus pallasi). Can. J. Fish. Aquat. Sci. 40:588-597.

SMITH, P.J. & J. SEDRANSK (1982). Bayesian optimization of the estimation of the age composition of a fish population. J. Am. Stat. Ass. 77:707-713.

YAMANE, T. (1976). Elementary sampling theory. Prentice-Hall, Inc., Englewood Cliffs, N.J. 405 p.