NATIONAL CHILD DEVELOPMENT STUDY

STATISTICS TEACHING SET

INTRODUCTORY STATISTICS

PREPARED BY: ORLY MANOR
CITY UNIVERSITY
LONDON

Preface to the teaching data set

Statistics and quantitative methods courses can be made more attractive to students by illustrating analytic methods with examples derived from the contemporary real world and involving students in computer analysis of these data. Although there are numerous data sets which might be appropriate for such purposes, the lecturer giving the course would need to spend quite a lot of time selecting the "right" examples and transforming the data set into a good teaching data set.

This teaching data set has been developed for introductory statistics courses for sociologists, psychologists, teachers, economists and other social scientists The set is based on experiences and characteristics of the National Child Development Study cohort of children growing up in Britain between 1958 and 1981.

The teaching set comprises a data tape and documentation; a booklet providing a general description of NCDS; and this book.

The data set covers 2000 people selected at random from the 12,537 respondents to the survey of the NCDS cohort in 1981, and information on some 32 characteristics obtained in the course of the study. The data include information concerning sex, social class, school attainments, physical measurements etc A complete list can be found in Appendix A The data are available as an SPSS-X save file

Preface to the book

The book covers various statistical topics usually presented in introductory statistics courses The topics are descriptive statistics, elements of probability and statistical inference, frequency tables, t-test, regression, analysis of variance and elements of multivariate analysis.

Each statistical topic encompasses a theoretical presentation followed by a number of examples. Each example includes a suitable SPSS-X Program. This structure should enable students to become familiar with statistical ideas and the methods and uses of SPSS-X

The mathematical presentation throughout the book is kept to as simple a level as possible. The book concentrates on explaining basic concepts using real world examples and on performing statistical procedures using modern software. Mathematical proofs and computational details do not figure prominently.

Two appendices are included Appendix A. contains information concerning the variables included in the data set; Appendix B. contains some statistical tables. The book begins with a description of the National Child Development Study.

A number of persons have helped me in the preparation of this book I would like to thank John Fox and Ken Fogelman for their constructive comments and Peter Shepherd for his help with the data. I am also grateful to Sharon Clarke who skillfully typed the manuscript.

The idea for this book originated from discussions held between the NCDS User Support Group and a number of lecturers interested in the project.

I am indebted to the Cambridge University Press for their permission to print extracts from the New Cambridge Elementary Statistical Tables by D.V Lindley and W.F. Scott (Appendix B, Tables B.2 B.3 and B.4).

ORLY MANOR

CONTENTS

CHAPTER 1

THE NATIONAL CHILD DEVELOPMENT STUDY

1.1 GENERAL

NCDS is a national longitudinal study of all the people in Great Britain who were born in the week 3-9 March, 1958

It has its origins in the 1958 Perinatal Mortality Survey, carried out by the National Birthday Trust Fund in order to study factors associated with stillbirth, early death and congenital handicapping conditions. With the first follow-up, carried out by the National Children's Bureau in 1965, the focus of the study broadened to encompass social, educational and physical development more generally

Subsequent follow-ups of the entire cohort, also by the National Children's Bureau, took place at the ages of eleven, sixteen and twenty-three. In addition details of public examination results were obtained from schools in 1978, and there have been a number of studies of special groups from within the cohort, such as the adopted, the handicapped, the gifted and the socially disadvantaged.

1.2 MAIN STAGES OF THE STUDY

Methods of data collection for the main stages are summarised below:

At birth (1958)

Questionnaires completed by midwives from interviews with the mother and medical records.

At seven (1965)

Parents interviewed by health visitors.
Questionnaires completed by schools
Medical examinations by local authority medical officer.
Educational tests completed by the cohort members.

At eleven (1969)

As at seven, with the addition of a short personal questionnaire completed by the cohort members.

At sixteen (1974)

As at seven and eleven, but with a much more substantial personal questionnaire.

At twenty-three (1981)

Personal interview of cohort members, carried out by professional research interviewers.
Census-based data describing the area in which the cohort member was living in 1974 and in 1981 were also added to the database

It would not be appropriate to attempt here either to describe the several thousand items of information which are now held on each cohort member or to summarise the several hundred publications which have so far arisen from the study. Further information can be found in, for example, Davie et al (1972), Fogelman (1983), NCDS4 Research Team (1987) and Shepherd (1988).

The data from all the above stages have been deposited with the ESRC Data Archive, and some 50 projects based on them are taking place in this country and overseas. There is a NCDS User Support Group at City University which, in addition to encouraging and facilitating use of the existing data, is planning the next stage of the study which it is hoped will take place when the cohort members are in their early thirties.

1.3 RESPONSE

A major feature of the study has been its success in maintaining the goodwill and co-operation of the cohort members and others on whose participation the study has depended. Of some 16,500 survivors since the Perinatal Study over 90% were traced and participated in the seven and eleven-year stages. The comparable figure at sixteen was 87%, and the 12,537 people interviewed at twenty-three represented 78% of the target figure.

General analyses of response have demonstrated only small biases, with a slight under-representation in later stages of some disadvantaged groups. Re-calculation of earlier analyses omitting later non-respondents show that underlying relationships do not appear to be affected by response patterns.

Although attempts were made during the school years to include those people who were born in the same week but who had entered the country since birth, the one severe bias within the study is an under-representation of ethnic minority groups.

1.4 THE NCDS TEACHING DATA SET

The NCDS teaching data set provides data for 2000 individuals who have been selected as a random sample of the 12,537 respondents to the NCDS survey carried out in 1981. In addition to information taken from this survey, data from the earlier surveys carried out at birth, age 7, age 11, age 16 and age 20 have been included The information included in this data set comprises only a small fraction of the vast amount of information available for each member of the NCDS Cohort. A full list of the variables included in the teaching data set is given in Appendix A

CHAPTER II

THE DATA SET

2.1 INTRODUCTION

Information collected in an empirical study is called "data". Data usually include a series of measurements on various observed phenomena. We use the word measurement to represent the operation of assigning numbers to subjects, events or characteristics. Each phenomenon being measured is called a variable, and the basic unit for which measurements are taken is called a case. The collection of all the information gathered, i.e all the variables for all the cases, is called the data set. In our example we have a total of 2000 cases taken from NCDS and for each case 32 measurements or variables. The names of the variables are presented in the variable list (see Appendix A) Our data set therefore includes information on 32 variables taken for 2000 individuals.

2.2 DEFINING THE DATA SET

2.2.1 General

Our data set is held on a computer file. We need to define the way it is stored on the file so that others can read it. To do this we need to specify a number of characteristics of the file such as where variables are stored on the file and how, what they are called and what they represent. The data definitions we consider here are those suitable for SPSS-X, a software package commonly used by social scientists.

2.2.2 The format

The format of the file describes the structure and position of the variables and the numbers of lines per case. The data usually include an additional variable called the ID number (or case ID) which is used as an identifier

The NCDS data set comprises 3 records (lines of information) per case. There are 3 lines with eleven variables on each line. Each variable occupies seven columns where the last two of the seven digits come after the decimal point. All the variables are numenc - that is they contain only numbers. This data set is described by the format statement: (3(11F7.2/)).

As an example we can consider the data recorded for the first case·

```
    100    100    300 365700    100   6300   6700    400   2400   1800    800
  14500   4310   1400   1800    100    400    100   2200    -58   1800   9826
    400    100  17780  19600    800   4000   6976  10626    300      0   9900
```

The first variable is the ID number. It occupies the first seven columns (spaces) where the last two digits come after the decimal point, therefore its value is 1.00. The second variable is SEX and it occupies the next seven çolumns, where again the last two digits come after the decimal point The value of SEX is therefore 1 00. The third variable NATIONO has the value 3 00 and the fourth one BIRTHWT has the value 3657.00. There are altogether 33 variables including the ID number

2.2.3 Variable labels and variable values

The names and the values variables can take are called variable names and variable values. Sometimes information for a particular variable is not available for a case, and when this is the situation a special code is used to indicate that the value is missing for this variable Sometimes we wish to include on printed output more detailed names of the variables and their values Those detailed names are called variable labels and value labels. Consider, for example, the NCDS data. The names of the variables in the NCDS Data Set are given in the variable list, and the variable values including missing values are given in the detailed variable list. Both lists can be found in Appendix A. For example the first variable has the name SEX, it's possible values are 1 (when male) 2 (when female) and -1 (when the value is missing). The label attached to this variable is sex of child and its possible values 1 and 2 will get the labels male and female

2.2.4. SPSS-X system file

The format, the variable names, the missing values, the variable labels and value labels define the data file We only need to define the data file once. All the information can be saved and stored along with the data on an SPSS-X system file One can gain access to the system file for different SPSS-X jobs without re-defining the data file

The NCDS data together with the data file definitions are stored on a file which have been created as an SPSS-X save file ready for you to use. Assume that the file name is NCDS.

2.2.5. Prepare to run your first job

Once the data file is defined we are in a position to run our first job. In order to read our SPSS-X system file (i.e the file NCDS) the two following commands are used:

```
FILE HANDLE NCDS / FILE SPECIFICATION*
GET FILE=NCDS.
```

Example 2.1

Consider that we wish to display all the data file definitions (i.e the names of the variables along with their values). The SPSS-X Program is displayed in job 2.1 Note that each SPSS-X program starts with the command TITLE and ends with the command FINISH.

```
SPSS-X PROGRAM
TITLE "JOB 2.1"
FILE HANDLE NCDS
GET FILE=NCDS
DISPLAY DICTIONARY
FINISH
```

*specification is specific to each computer and operating system

2.2.6 A note concerning the SPSS-X language

Each example in this book includes an SPSS-X program. For each command which is used throughout this book, an explanation and examples are provided. This enables the reader to use the system. We present a brief description of the structure of the SPSS-X commands. The entire language and the full range of specifications for the different commands can be found in the SPSS-X User's Guide upon which this section is based.

Every SPSS-X command begins in the first column of a new line and continues for as many lines as necessary. Commands can be used for example, to get the data, to transform the data and to display the data. Procedures are commands that actually read the data and create tables and plots or produce various statistical analyses of the data. Each command begins with a command keyword which is followed by at least one blank space and then any specification required to complete the command. Many specifications include subcommands. As for example, in the command DISPLAY DICTIONARY a VARIABLES subcommand which limits the display to certain variables can be added

```
DISPLAY DICTIONARY/VARIABLES = SEX, NATIONO
```

In this example, the display is limited to the first two variables in our variable list

The first word of every command keyword must be spelled out in full. All other subsequent keywords that make up the command can be truncated to a minimum of three characters. For example, the DISPLAY command above can be written in the following way.

```
DISPLAY DICTIONARY/VAR = SEX, NATIONO
```

Subcommands are usually separated from each other by a slash. One can add spaces or break lines at any point where a single blank space is allowed, for example· around slashes, parentheses or equal signs and between variable names

2.3 MORE ABOUT THE DATA

In chapters 3-9 the reader will discover how SPSS-X procedures are used to create tables and plots of the data and to calculate various statistics which summarise the data First, we need to introduce different types of measurements

2.3.1 Level of measurements

Measurements can be obtained on one of four scales These are·

(a) Nominal scale- This is the simplest scale in which no assumptions are made about relations between values The values serve only as names for categories Sex of child (SEX) and nation at birth (NATIONO) are examples in NCDS.

(b) Ordinal scale- In this scale it is possible to order all the categories. Father's social class at child's birth (PASCO) and I do not like school at age 16 (LIKES16) are examples of such scales in NCDS.

(c) Interval scale - An interval scale is an ordinal scale in which the distance between values is meaningful. Two examples in NCDS are the draw-a-man test score (DRAW7) and reading test score at age 7 (READ7).

(d) Ratio scale - This is the most complicated scale. It is an interval scale on which it is also possible to locate an absolute zero point. Examples in NCDS include child's height at 11 (HT11) and child's weight at birth (BIRTHWT).

2.3.2 Types of variables

It is also possible to classify variables into two groups - discrete variables and continuous variables - according to the range of results which are possible. Discrete variables can only take a countable number of different values. Continuous variables can take any value within a range. Highest educational qualification (HIGHQUAL) and number of children under 21 in the family at age 11 (NKIDS11) are examples of discrete variables in the NCDS data set. Child's height at age 11 (HT11) and net earnings per week at age 23 (NEARNPW) are examples of continuous variables

CHAPTER III

DESCRIPTIVE STATISTICS - THE FIRST STEP IN THE ANALYSIS OF THE DATA

3.1. INTRODUCTION

Descriptive statistics describes the group of methods used to display the data This should be the preliminary stage in every analysis. The aim is to point out the most interesting or important features of the data. In the following examples we concentrate on the analysis of single variables

3.2. TABLES

We should begin by counting the number of times each value occurs The table containing the frequency with which each value occurs is called the frequency distribution.

Example 3.1

Variables considered

PASCO father's social class at child's birth

The variable PASCO measures the social class of the cohort members' fathers in 1958 when the cohort members were born. The possible values are 1,2,3,4,5,6 and -1 which refer to the categories Professional (1), Intermediate (2), Skilled Non-Manual (3), Skilled Manual (4), Semi-Skilled (5), Unskilled (6) and missing values (-1). We wish to count the number of individuals who at the time of their birth had fathers in professional occupations, intermediate occupations etc

We may also be interested in calculating the percentage of individuals whose fathers were professional.

All the above information can be obtained using the SPSS-X FREQUENCIES procedure. Job 3.1 gives the SPSS-X program for doing this

SPSS-X PROGRAM

```
TITLE "JOB 3.1"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VARIABLE = PASCO
FINISH
```

Job 3 1 will create as an output a frequency table for the variable PASCO The table provides the following information for each value of the variable· the frequency, the percentage, the valid percentage and the cumulative percentage. The valid percentage is the percentage out of all cases for which the value of the variable is not missing. The cumulative percentage for a particular value is the sum of the valid percentages of that value and all other values that precede it in the table. For the final category the cumulative percentage is always 100 percent.

3.3. GRAPHICAL METHODS

3.3.1. Introduction

The data should always be plotted in a number of different ways since a lot of the information contained in the data can often be obtained by eye. Graphical methods are also helpful for detecting observations which are well removed from the main bulk of data. These, so called outliers, are frequently a result of errors in printing, coding or recording

3.3.2. Bar chart

The simplest graph is the bar chart which displays the frequency distribution In bar charts bars are constructed over each value with length proportional to the frequency with which each value occurs

Example 3.2

Variables considered

PASCO father's social class at child's birth.

In SPSS-X bar charts are obtained by the FREQUENCIES procedure displayed in job 3.2

```
SPSS-X PROGRAM
TITLE "JOB 3.2"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VARIABLES = PASCO/ BARCHART
FINISH
```

One FREQUENCIES command can be used to print the frequencies of a number of variables at the same time, as is displayed in Example 3.3.

Example 3.3

Variables considered

MASMOKE	Whether mother smoked in pregnancy
NATIONO	Nation at birth
LIKES16	I do not like school at age 16
MATH11	Maths test score at age 11

The output of the SPSS-X program displayed in job 3.3 contains frequency tables and bar charts for each of the variables

```
SPSS-X PROGRAM
TITLE "JOB 3.3"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VARIABLES = MASMOKE, NATIONO, LIKES16, MATH11/ BARCHART
FINISH
```

Tables and charts such as these can be used to make some preliminary observations about the data.

3.3.3 Histogram

Generally for variables that can take on many different values it is more useful to group the data. First group the values of the variable into non-overlapping intervals and then count the number of cases with values within each interval A histogram is a graph displaying the distribution of grouped data. Usually in a histogram rectangles are constructed over the intervals with areas proportional to the frequency of each interval.

The middle of the interval is called the midpoint and all values within the interval are considered concentrated at this point.

SPSS-X draws a row of asterisks over the midpoint of the interval to represent the frequency of the interval. The number of asterisks is proportional to the frequency. Histograms are produced by the FREQUENCIES procedure.

Example 3.4

Variables considered

MATH11 Maths test score at age 11

In the previous example the frequency table and bar chart for the variable MATH11 were not very informative Here we create a histogram for the variable MATH11. The SPSS-X program is displayed in job 3.4.

SPSS-X PROGRAM

```
TITLE "JOB 3.4"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VAR= MATH11/ HISTOGRAM
FINISH
```

3.3.4 Histogram shapes

Histograms come in various shapes The most common shape is the symmetric bell shape If there are more cases towards one end of the distribution than the other it is called skewed; left skewed if there is a tail towards the left and right skewed if there is a tail towards the right

Example 3.5

Variables considered

PAHT	Father's height in inches
MAHT	Mother's height in inches
DRAW7	Draw-a-man score at age 7
UNEMTIME	Total months ever unemployed
READ16	Reading comprehension test score at age 16
MATH11	Maths test score at age 11
READ11	Reading comprehension test score at age 11

In order to illustrate different shapes of a histogram we consider the variables listed above The following SPSS-X command can be used to print histograms for these variables and to skip the printing of the frequency table

```
FREQUENCIES      VAR= PAHT, MAHT, DRAW7, UNEMTIME, READ16, MATH11,
                      READ11/  HISTOGRAM/ FORMAT= NOTABLE
```

3.4 SUMMARY MEASURES

3.4.1. General

The next step after drawing up some preliminary tables and graphs is to identify some simple measures which summarise the data. Such summaries which are calculated from the data are called statistics There are a number of features of the data which statistics can be used to describe Simple statistics usually describe one of three features· location, dispersion and shape

3.4.2.Measures of location

Measures indicating the "centre" of a set of data are called measures of location. The most commonly used measures of location are the mean, median and mode

The mode describes the most frequent value (or values). It is not unique. It can be used for data measures at any level and is most suitable for nominal and ordered data.

The median describes the mid-point of the distribution, with half of the observations numerically greater than the median and half numerically smaller When the number of observations is even the mean of the two middle observations is taken as the median value The median should not be used for nominal data

The mean is the arithmetic average Suppose n measurements have been taken on the variable under consideration we denote them by $X_1, X_2, \quad , X_n$. The mean of the observations is denoted by $\bar{X}$ and given by $\bar{X}=\sum_{i=1}^{n} X_i/n$. The mean should not be used for nominal or ordinal data

Comments

(1) The mean is the most important and most frequently used measure of location. The reason lies in the statistical properties of the mean. These will be considered in the next chapter.

(2) When the distribution of observations is approximately symmetric

the mean, mode and median will be about the same. This will not be the case for skewed disrtibutions

(3) Outliers influence the mean but not the median.

In SPSS-X location measures are obtained by the FREQUENCIES Procedure, subcommand STATISTICS

Example 3.6

<u>Variables considered</u>

MATH11	Maths test score at age 11
DRAW7	Draw-a-man score at age 7
PAHT	Father's height in inches
UNEMTIME	Total months ever unemployed
LIKES16	I do not like school at age 16

The SPSS-X program displayed in job 3 6 gives the mode, median and mean for each of the variables

```
SPSS-X PROGRAM
TITLE "JOB 3.6"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VAR= MATH11, DRAW7, PAHT, UNEMTIME, LIKES16/
              FORMAT= NOTABLE/
              STATISTICS= MODE, MEDIAN, MEAN
FINISH
```

3.4.3. Measures of dispersion

As well as wanting to know where the centre of the distribution is it is often very important to measure how spread the data are. The most important statistics for measuring the dispersion of a set of data are the range, the inter-quartile range and the variance (and its derivative, the standard deviation).

The range of a set of data is the difference between the largest and smallest observations, as in common English

Deciles and quartiles like the median, break up the distribution into a number of equal parts, in these cases ten and four respectively. So the lower quartile is the value below which one quarter of the observations lie and the upper quartile is the value above which one quarter of the observations lie. The inter-quartile range is the distance between the upper and lower quartiles. The inter-quartile range measures the spread of the middle half of the observations

The variance is a more sophisticated measure of dispersion The variance of a set of n observations $X_1, X_2, \ldots, X_n$ is denoted by S^2 and given by

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

The standard deviation is the square root of the variance and is denoted by S.

Comments

(1) The most important measure of dispersion is the variance (or the standard deviation).

(2) The standard deviation is in the same units as the observations.

(3) The range does not take into account the distribution of the observations between the smallest and the largest

(4) The range is affected by the number of observations.

(5) The variance and standard deviation are not suitable for ordinal data, while the range and inter-quartile range are

(6) None of the statistics mentioned above is suitable for nominal data.

(7) One additional measure of dispersion is the coefficient of variation. This is equal to $S/\bar{X}$. It measures the spread relative to the mean When the data are measured on a ratio scale this coefficient is independent of the units of measurement.

In SPSS-X dispersion statistics are obtained by the FREQUENCIES procedure, STATISTICS Subcommand. The lower and upper quartiles are obtained using the PERCENTILES Subcommand

Example 3.7

<u>Variables considered</u>

As in Example 3.6.

Suppose that we are interested in calculating the following statistics: range, inter quartile range, variance, standard deviation, upper quartile, lower quartile. One can use the program displayed in job 3.6 but replacing the FREQUENCIES command with:

```
FREQUENCIES VAR=  MATH11, DRAW7, PAHT, UNEMTIME, LIKES16/
          STATISTICS= RANGE VARIANCE STDDEV/
          PERCENTILES= 25 75
```

3.4.4 Measures of shape

As already mentioned, the shape of the distribution of a set of measurements can be displayed by a histogram. Examination of histograms provides an indication of skewness. We can also calculate formal measures of skewness. One such measure is Sk It is given by

Sk = 3(MEAN - MEDIAN)/STANDARD DEVIATION

For a perfectly symmetrical distribution the mean and median are identical and the value of Sk is zero. When the distribution is skewed to the left Sk will be negative and when it is skewed to the right, Sk will be positive In SPSS-X a measure of skewness is calculated using the FREQUENCIES procedure subcommand STATISTICS = SKEWNESS.

3.5 MORE GRAPHS

Earlier we suggested that it was advantageous to plot the data in various ways We present two more graphical methods for displaying the distribution of an observed variable. The methods are - stem and leaf and box plots

3.5.1 Stem and leaf

Stem and leaf plots provide a convenient way of examining the distribution of a variable It is similar to a histogram or a bar chart but it displays the values of all the observations Consider the following 20 observations:

1 8, 2.0, 2 1, 2.7, 2.7, 2 8, 3.0, 3 1, 3.1, 3 2, 3.5, 3.5, 3.5, 3.6, 4.0,
4 0, 4 2, 4 8, 4 9, 5.0.

The stem and leaf plot would be·

```
1 | 8
2 | 01
2 | 778
3 | 0112
3 | 5556
4 | 002
4 | 89
5 | 0
```

The numbers to the left of the dotted line are called stem and those to the right are called leaves. The plot is constructed so that every case is represented by a leaf In the plot above the first line represents one case with value 1.8 while the second line represents two cases with values 2.0 and 2 1 The stem, 2, is the same for both cases while the leaves differ. When there are several cases with the same values each is represented by a separate leaf value. The intervals and the scale of the plot are not fixed and can be determined according to the data.

The information contained in a stem and leaf plot is similar to the information contained in a histogram but in a stem and leaf plot we also have information about observations within each interval. In SPSS-X stem and leaf plots are printed by the procedure MANOVA subcommand PLOT.

Example 3.8

Variables considered

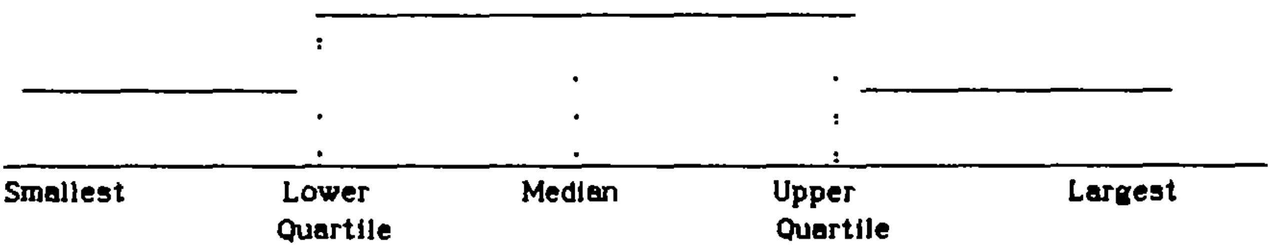
BIRTHWT Child's weight at birth in grams

Consider plotting birthweight in the NCDS teaching data set. The characteristics of the stem and leaf plot can be illustrated better if we take a small number of cases To do this we select a random sample of 200 cases. The command SELECT IF is used to select cases for which the values of BIRTHWT are not missing The command SAMPLE 200 is used to select randomly a sample of 200 out of the cases for which the value of BIRTHWT is not missing (there are 1827 such cases). The SPSS-X program which does this is displayed in job 3.8.

```
SPSS-X PROGRAM
TITLE "JOB 3.8"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (BIRTHWT NE -1)
SAMPLE 200 FROM 1827
MANOVA BIRTHWT/
          PLOT= STEMLEAF/
          DESIGN
FINISH
```

3.5.2 Box plot

To construct a box plot first draw a rectangle whose length is equal to the inter-quartile range and whose width can be chosen arbitrarily. Then divide the rectangle by a line which is locating the median. Then draw lines to connect the rectangle with the smallest and largest observations.



The box plot is easy to draw and understand and gives a good idea where the distribution centres and how spread out it is. It is also helpful for comparing several groups of data

To draw a box plot in SPSS-X use the MANOVA procedure with subcommand PLOT

Example 3.9

Variables considered

READ11	Reading comprehension test at age 11
SEX	Sex of child

In this example we shall draw box plots for the reading test scores at age 11 The SPSS-X program in job 3 9 gives two box plots, one for girls and one for boys.

SPSS-X PROGRAM

```
TITLE "JOB 3.9"
FILE HANDLE NCDS
GET FILE=NCDS
MANOVA READ11 BY SEX (1,2)/
         PLOT= BOXPLOTS/
         DESIGN
FINISH
```

3.5.3 A note concerning outliers

The different methods of descriptive statistics, especially the graphical ones are most helpful for detecting outliers. When an outlier is detected, it should be compared (if possible) with the original records. When rejecting an outlier it can be removed or replaced by another value, for example the mean or the next largest (or smallest) observation. Another possibility is for a particular analysis to use only the middle half of the observations. This is done by trimming one quarter of the observations from each end (so that the outliers are removed) For example, an average calculated on the basis of those observations is called a 25% trimmed mean Other degrees of trimming are, of course, also possible.

3.6 EXERCISE

In the following chapters we shall apply more advanced statistical methods to the following variables:- Social class of current job at age 23 (CURRSOC), Number of children under 21 in the family at age 11 (NKIDS11), Reading comprehension test score at age 16 (READ16), I do not like school at age 16 (LIKES16), Net earnings per week (from main job) at age 23 (NEARNPW) and Child's weight at birth in grams (BIRTHWT). Readers are therefore advised to use the descriptive statistics outlined in this chapter to investigate the distributions of these variables

1. Display the data on each variable in any way you find helpful and informative

2. Calculate measures of location and measures for spread

3. Make some observational comments on any feature worth mentioning.

CHAPTER IV

INTRODUCTION TO STATISTICAL INFERENCE

4.1 INTRODUCTION

Statisticians are frequently concerned with making statistical inferences concerning a population on the basis of partial information. This partial information is provided by drawing samples from the population. The data from the sample is used to deduce or generalize about the population from which the sample was drawn.

Statistical inference may be divided into two major areas: estimation and tests of hypotheses. In estimation a statistical inference is made concerning an unknown population parameter. Examples would be estimating the proportion of voters favouring a specific party in a forthcoming election, or estimating the mean weekly income of young British males. Sample data is used to estimate the parameters of interest. In tests of hypotheses we construct a decision making procedure which leads us to accept or reject a certain claim or hypothesis concerning a population For example, a researcher might be required to decide on the basis of experimental results whether a new teaching method is superior to the one being used, or to decide, on the basis of sample data, whether gender and income are related to each other.

Estimation and tests of hypotheses are discussed in this chapter, but first some basic statistical notions are required. Random variables, the normal distribution and elementary theory of sampling distribution are presented in the first sections of this chapter.

4.2 RANDOM VARIABLES

Most statistical ideas are based on probability theory. This explains in mathematical terms the pattern of observations in the real world. For example, suppose we toss a coin 3 times and are interested in the number of times a tail occurs. The number of tails would be denoted by a capital X and is an example of a discrete random variable. X can take any one of the following values 0,1,2,3 Probability theory can be used to calculate the probability that the random variable, X, will assume any of these values. A discrete random variable takes on various values with various probabilities and the list of all the possible values together with their corresponding probabilities is called the probability distribution of X

Suppose now that we select randomly one student from a large class and record his height If we denote the height of the chosen student as X, then X is now a continuous random variable A continuous random variable can assume any value within a designated range of values. For a continuous random variable one cannot specify in advance all possible values of the variable with corresponding probabilities. However, the distribution of such random variables can be defined by mathematical formulae or by curves on a graph. The curves are such that areas under different sections of the curve represent probabilities.

The distributions of individual random variables can often be characterized by just two parameters, the mean and the variance.

The mean represents the centre while the variance represents the dispersion.

4.3 THE NORMAL DISTRIBUTION

The most important continuous distribution in statistics is the normal distribution. Its graph, called the normal curve, has a bell shape that describes many sets of data that occur in nature.

The normal distribution is symmetric, and the mean which is equal to the median is denoted by μ . The variance of the normal distribution is denoted by σ^2, and σ is the standard deviation. The exact shape of the distribution is defined by the two parameters μ and σ. The location centre is determined by μ while the shape depends on σ. The larger σ is, the more spread the distribution will be Whatever the values of μ and σ, the normal distribution is such that approximately one out of 3 observations will lie more than one standard deviation away from the mean and only one in twenty will lie more than 2 standard deviations from the mean. To denote a random variable, X, which follows the normal distribution with parameters μ and σ we write X is $N(\mu, \sigma^2)$.

The simplest of the normal distributions is the standard normal distribution in which $\mu = 0$ and $\sigma = 1$. The distribution is tabulated and probabilities related to the observed value of a random variable distributed as N(0,1) can be calculated from these tables. The tables are given in Appendix B (Table B.1).

Every random variable, X, having a normal distribution $N(\mu, \sigma^2)$ can be transformed into a random variable, Z , having the standard normal distribution i.e. N(0,1). The transformation is given by:

$$Z = \frac{X - \mu}{\sigma}$$

Example 4.1

Variables considered

READ11	Reading comprehension test score at age 11
MAHT	Mother's height in inches
SEX	Sex of child
BIRTHWT	Child's weight at birth in grams

In the previous chapter we showed the reader how to plot the frequency distributions of the variables READ11, MAHT and BIRTHWT (for females). The histograms of each of these variables look bell shaped. Now let us use the SPSS-X HISTOGRAMS = NORMAL subcommand to print a superimposed normal distribution on the histogram. The computer will choose the normal distribution with the same mean and variance as was calculated from the observations of each of these variables. The SPSS-X program for doing this is displayed in job 4.1.

```
SPSS-X PROGRAM
TITLE "JOB 4.1"
FILE HANDLE NCDS
GET FILE=NCDS
FREQUENCIES VARIABLES= READ11, MAHT / HISTOGRAM= NORMAL
SELECT IF (SEX EQ 2)
FREQUENCIES VARIABLES= BIRTHWT / HISTOGRAM= NORMAL
FINISH
```

4.4 SAMPLES AND POPULATIONS

The universe of objects or people to be studied is called the population. Observations are usually based on a sample drawn from the total population. Statistical inference is concerned with making inferences about the population on the basis of observations made on a sample drawn from the population. To do so successfully we need the sample to reflect well the important characteristics of the population. Random samples are one type of sample which statisticians argue are representative of the populations from which they are drawn. In a random sample, whenever an observation is drawn every individual in the population is as likely as any other individual to be chosen.

As already indicated the most important characteristics of population distributions are the mean and the variance We would therefore like to be able to make inferences about the population mean and variance on the basis of information obtained from a sample The sample statistics provide estimates of the unknown population values. It is important to distinguish between the population values, called parameters, which are fixed characteristics of the population and sample estimates which will vary from sample to sample according to which set of observations were selected for the sample.

The mean of the population is usually denoted by μ while the sample mean is denoted by $\bar{X}$ The variance of the population is denoted by σ^2 while the sample variance is denoted by S^2.

4.5 THE SAMPLING DISTRIBUTION OF $\bar{X}$

Sample estimates of population parameters vary from sample to sample and are therefore random variables and have distributions The distribution of a statistic is called the sampling distribution. We shall now describe the sampling distribution of $\bar{X}$ when a random sample of size n is drawn from a population with mean μ and variance σ^2.

4.5.1 Location, spread and shape

The mean and variance of the sampling distribution are its two most important parameters. It can be shown that the mean is equal to μ while the variance is equal to σ^2/n The standard deviation is $\sigma/\sqrt{n}$ The standard deviation is also called the standard error The standard error gives a measure of the distribution of deviations between the parameter μ and it s estimate $\bar{X}$ As the sample size increases so the standard error gets smaller and we can be more confident of having an estimate of the mean of the population that is not very far from the true mean.

It can also be shown that if the population distribution is normal then the sampling distribution of $\bar{X}$ is also normal. Even when samples are taken from non-normal populations the sampling distribution of $\bar{X}$ is still normal if the samples are large. The theory on which this is based is called the central limit theorem.

To summarize, when a random sample of size n is taken, the sample mean $\bar{X}$ fluctuates around the population mean μ with standard deviation $\sigma/\sqrt{n}$. As n increases the distribution of $\bar{X}$ concentrates more and more around μ and it's shape resembles more and more the normal curve

Example 4.2

Variables considered

MAHT	Mother's height in inches
ATTEN16	Child's school attendance at age 16

An examination of the distribution of each of these variables reveals that MAHT has a bell shaped distribution with mean 63 44 inches and a standard deviation of 2.48 inches ATTEN16 has a very skewed distribution with mean 88 94 percent and standard deviation of 14.98 percent. Assume that the population consists of the cases in our data set.

We will now draw random samples from our population and examine the sampling distribution of the means. The SPSS-X program displayed in job 4 2 draws two samples, one of size 5 and one of 40 For each sample the sample mean for each variable is printed. One can run this program say 30 times in order to generate 30 estimates of the population mean for each variable, for each sample size. One can then calculate the mean for each set of 30 estimates and the standard deviation and can plot the estimates in a histogram. The results can then be compared with the theory about the sampling distribution of $\bar{X}$ presented above.

The TEMPORARY command signals that the exclusion of the cases with missing values and the sampling are in effect only for the next FREQUENCIES procedure. The SET SEED command sets the starting point for the SPSS-X random number generator. If the same seed number is used for each run the computer will generate the same sequence of numbers and will select the same sample. Therefore the seed should be changed at the start of each run

```
SPSS-X PROGRAM
TITLE "JOB 4.2"
FILE HANDLE NCDS
GET FILE=NCDS
SET SEED= 123456
TEMPORARY
SELECT IF (MAHT NE -1)
SAMPLE 5 FROM 1814
FREQUENCIES VAR= MAHT/ FORMAT= NOTABLE/ STATISTICS= MEAN
TEMPORARY
SELECT IF (MAHT NE -1)
SAMPLE 40 FROM 1814
FREQUENCIES VAR= MAHT/ FORMAT= NOTABLE/ STATISTICS= MEAN
```

```
TEMPORARY
SELECT IF (ATTEN16 NE -1 AND ATTEN16 NE -2)
SAMPLE 5 FROM 1546
FREQUENCIES VAR= ATTEN16/ FORMAT= NOTABLE/ STATISTICS= MEAN
TEMPORARY
SELECT IF (ATTEN16 NE -1 AND ATTEN16 NE -2)
SAMPLE 40 FROM 1546
FREQUENCIES VAR= ATTEN16/ FORMAT= NOTABLE/ STATISTICS= MEAN
FINISH
```

4.6 ESTIMATION

4.6.1 Estimates and estimators

The sample information can be used to estimate population parameters. We distinguish between an estimator and an estimate An estimator is a rule that tells us how to determine from any sample a numerical value to estimate a certain population parameter, while an estimate is the actual numerical value obtained from a particular sample

Suppose that in order to estimate the mean height of students in a certain class we select a random sample of 10 students from this class. The sample mean, $\bar{X}$, is an intuitive estimator for the population mean, μ. The sample mean $\bar{X}$ is an estimator and is a random variable that can take on different values from sample to sample. The estimate is the numerical value obtained from one particular sample. Since an estimator is a random variable it's sampling distribution should be considered.

4.6.2 "Good" estimators

The sample mean, $\bar{X}$, is not the only possible estimator for μ, the population mean. The sample median or the average of the smallest and largest observations are also possible estimators.

We choose estimators that have "good" properties. The following two properties are considered as desirable properties for a "good" estimator.

a) Although the estimator value varies from sample to sample, on average, it's value should be equal to the value of the parameter being estimated. Such an estimator is called an unbiased estimator.

b) It is not very helpful for an estimator to be correct "on average" if it fluctuates widely from sample to sample Therefore the sampling distribution of an estimator should be concentrated close to the true value of the parameter. This implies an estimator with a small variance.

An estimator which has the smallest variance among those that are unbiased is called a minimum variance unbiased estimator

It can be shown that when a random sample is taken from a population which is normally distributed, the sample mean and sample variance are minimum variance unbiased estimators for the population mean and variance respectively

Example 4.3

This example is an extension of example 4.2.

Variables considered

MAHT Mother's height in inches

ATTEN16 Child's school attendance at age 16.

The parameter of interest, in each case, is the population mean. Three estimators are considered: the sample mean, the sample median and the sample average between the smallest and largest observations. Job 4.2 can be used after replacing the STATISTICS Subcommand with the following STATISTICS= MEAN, MEDIAN, MINIMUM, MAXIMUM. The sampling distribution of each statistic can then be considered

4.6.3 Confidence limits

The estimators considered so far in this section are point estimators but frequently we are interested in estimating intervals.

Suppose that a random sample of size n is taken from a population which is normally distributed with mean μ and variance σ^2 We learned earlier that the sampling distribution of the sample mean, $\bar{X}$, is normal with mean μ and variance σ^2/n. Therefore $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ follows the standard normal distribution

A characteristic of the standard normal distribution is that 68% of the population lies between -1 and +1 and 95% of the population lies between -1 96 and +1 96 We can, therefore, write

$$\text{Prob.}(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95 \qquad \text{or,}$$

$$\text{Prob}(\bar{X} - 1\ 96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 0\ 95$$

We say that we are 95% certain that μ is between $\bar{X}$ -1 96$\sigma/\sqrt{n}$ and $\bar{X}$ + 1.96$\sigma/\sqrt{n}$ and we call this interval a 95% confidence interval.
A confidence interval should be interpreted in the following way: if we select different samples and from each sample we produce such an interval, then in the long run about 95% of our intervals would include the true mean.

When the population distribution is non normal but the sample size is large this result again tends to be true

The confidence interval for the mean presented in this section is a function of the variance σ^2 which is usually unknown. Constructing a confidence interval for the mean when the variance is unknown and constructing confidence intervals for other parameters is also possible. Nevertheless since the topic of confidence intervals is closely related to the topic of hypothesis testing we will not develop it further in this section

Example 4.4

This example is an extension of example 4.2.

<u>Variables Considered</u>

MAHT	Mother's height in inches
ATTEN16	Child's school attendance at age 16

The output of job 4 2 can be used to calculated a 95% confidence interval for μ, for each of the samples selected. Then the proportions of those intervals which include the value of μ can be calculated.

4.7 TESTING HYPOTHESES

4.7.1 General

The general idea underlying hypothesis testing is that we begin with a designated (hypothesised) value for a population parameter, such as the population mean, and then we test whether the sample data are consistent with the hypothesised value. The sample estimate of the population parameter is compared with the hypothesized parameter and conclusions are drawn.

Assume that the population distribution is normal with mean μ which is unknown and standard deviation σ which is known. Further assume that we have taken a random sample of n observations denoted by X_1, X_2, . .., X_n. The sample mean is denoted by $\bar{X}$ From the assumption it follows that the sampling distribution of $\bar{X}$ is $N(\mu, \sigma^2/n)$

4.7.2 Four steps

The procedure of hypothesis testing can be divide into four steps:

Step 1.Formulate the hypotheses

The problem is presented as a decision problem between two rival possibilities called hypotheses. The null hypothesis is denoted by H_0 and the alternative hypothesis is denoted by H_1. The two possibilities are, for example, either that μ is equal to designated value μ_0 or that μ is larger than μ_0. We can write this as:

H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

We assume that the null hypothesis is true unless the data indicate otherwise. The burden of proof is always on the alternative hypothesis.

Step 2.Select a test statistic and calculate its observed value

We need to select a test statistic which will be used to indicate departures from the null hypothesis In the case of $\bar{X}$ and μ_0 it seems reasonable to base our test on the difference between $\bar{X}$ and μ_0. Since the standard error of $\bar{X}$ is a good measure of the dispersion of the sample mean around the population mean, a good test statistic is·

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

What can we say about this test statistic? If the null hypothesis is true, then as we showed in the previous section, the sampling distribution of $\bar{X}$ is $N(\mu_0, \sigma^2/n)$ The test statistic will consequently follow the standard normal distribution In this case the observed value of the statistic based on our sample of data is equivalent to an observation from the standard normal distribution

Step 3.Calculate the P-value

We now want to calculate how likely or unlikely the observed value of the statistic is, if the null hypothesis is really true. We do this by calculating the probability of getting a result which is as extreme, or more extreme, than the one obtained. This probability is called the P-value. In our example we calculate the probability that a random variable which has the N(0,1) distribution assumes a value which is larger or equal to the observed value. The probability is calculated using tables for the standard normal distribution. A small P-value indicates that the observed value of the statistic is unlikely if H_0 is true

Step 4.Decision making

The decision is based on the P-value, where "small" P-values indicate that H_0 does not seem to hold. We usually establish a clear criterion as to what is "small" before we formulate our hypotheses in step 1. The cut-off point we set is called the significance level and is denoted by α. Alpha is usually set to be either 0.01 or 0.05 The decision is made by comparing the P-value and α. If the P-value is smaller or equal to α then we reject the null hypothesis. If the P-value is larger than α we do not reject the null hypothesis.

4.7.3 General comments

(1) The procedure described above is called a significance test. If the P-value is less than 5 percent, for example, we say that the result is significant at the 5 percent level. This statistical use of the word "significant" should not be mistaken for common usage of the word to mean

"large" or "important". It means that the differences observed are unlikely to have been observed by chance if the null hypothesis is true.

(2) Statistical significance tests cannot be used to prove beyond any doubt that a particular hypothesis is true or false. When the P-value is small the data appear to be inconsistent with the null hypothesis and we will tend to reject it. But there is still a small chance that it is true.

(3) When the P-value is large we do not reject the null hypothesis. We must recognise however that in this case either the null hypothesis is true, that is $\mu = \mu_0$, or $\mu > \mu_0$ but we are unable to detect it The latter situation could arise when the sample is small or when μ is only a little greater than μ_0 If the sample size is small, the variability in $\bar{X}$ is large and even large differences between $\bar{X}$ and μ_0 may not lead us to reject H_0.

(4) The procedure is similar when the alternative hypothesis is formulated differently i e. $H_1 \cdot \mu \neq \mu_0$. In this case the test statistic is the same but the P-value is calculated slightly differently. The P-value is the probability of getting a result which is as extreme or more than the one obtained In this case extreme means both lower and higher and we calculate as follows.

P-value= Prob (Z >|observed value|) + Prob.(Z < -|observed value|)

Where Z denotes a random variable having a standard normal distribution. This case is called a two-tailed test while the cases when $H_1 : \mu > \mu_0$ or when $H_1 : \mu < \mu_0$ are called one-tailed tests.

(5) We began by assuming that the population distribution is normal. In cases where the distribution is non-normal but the sample size is large enough we can still use the method described. According to the central limit theorem the sampling distribution of $\bar{X}$ will be approximately normal.

Example 4.5

In the period after the Second World War it was assumed that the mean birth weight of female babies was 3255 grams with a standard deviation of 495 grams. Researchers assumed that the mean weight increased as a result of the improvements in economic conditions in the late fifties. We can use the NCDS data to test this. Let μ be the population mean of females' birthweights in the late fifties. We assume $\sigma = 495$ and we set $\alpha = 5\%$

Step 1.The null and alternative hypotheses are

$$H_0 \cdot \mu = 3255$$

$$H_1 : \mu > 3255$$

Step 2.The test statistic is $\dfrac{\bar{X} - 3255}{495/\sqrt{922}}$

From the sample data we find $\bar{X}$ = 3285 12 and n = 922. Consequently the observed value of the statistic is 1 85.

Step 3.P-value, P-value = P(Z > 1.85) = 0.032

Step 4.Since the P-value is 0.032, we decide to reject the null hypothesis and to conclude that the mean birthweight is significantly larger than 3255 grams

4.8 THE t-DISTRIBUTION

In the previous section we indicated how we would compare the mean of the sample with a hypothesised mean of the population. The test statistic was $\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$ and, under the assumption that the sampling distribution of $\bar{X}$ was $N(\mu, \sigma^2/n)$, can be shown to be distnbuted according to the standard normal distribution.

In this section we describe how to proceed when σ is unknown. We use the same statistic but replace σ by the sample standard deviation S.

The resulting statistic $\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ does not however follow the standard normal distribution It follows a distribution which is somewhat more spread out than the normal distribution, called the t-distribution The t-distribution is symmetric with mean zero and depends on a parameter called degrees of freedom This is the denominator in the calculation of

S Since

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

the degrees of freedom are n - 1. The number of degrees of freedom indicates the number of values that are free to vary in a random sample.

For large values of degrees of freedom the t-distribution tends towards the standard normal distribution Tables for the t-distribution are given in Appendix B (Table B.2). When testing hypotheses concerning the population mean when the variance is unknown the test statistic is

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and the P-value is calculated using the tables for t-distribution.

4.9 A NOTE CONCERNING DEGREES OF FREEDOM

The term 'degrees of freedom' (d.f.) will occur frequently in the following chapters. The degrees of freedom refer to the number of independent observations in a set. Suppose you know that the sum of 3 numbers is 100 and you are asked to choose the 3 numbers. You may choose only two of the numbers and therefore you have only 2 degrees of freedom. When you choose 3 numbers with no restriction you can choose all the 3 numbers and therefore have 3 degrees of freedom.

When computing the variance, after the mean is fixed, there are only n - 1 degrees of freedom associated with the values of numbers used to compute the variance. This same number is also the number of degrees of freedom associated with the estimated standard deviation and with the t-distribution used for testing hypotheses concerning the mean.

4.10 OTHER CONTINUOUS DISTRIBUTIONS

In later chapters two other continuous distributions are mentioned, the chi-square distribution and the F- distribution

The chi-square, denoted χ^2, distribution is not symmetrical but is positively skewed. The distribution depends on one parameter called degrees of freedom For large values of degrees of freedom the distribution tends towards the normal distribution. Tables of this distribution are given in Appendix B (Table B.3).

The F-distribution depends on two parameters, denoted by v_1 and v_2, both called degrees of freedom. The distribution is not symmetrical and a few percentile points of the distribution are given in Appendix B (Table B 4).

4.11 EXERCISE

Consider the two estimators for the population variance - the sample variance and the sample range. The variables under consideration are MAHT and ATTEN16. Use a similar job to job 4.2 to investigate the sampling distribution of the two suggested estimators.

CHAPTER V

TESTING THE DIFFERENCE BETWEEN THE MEANS OF TWO POPULATIONS

5.1 INTRODUCTION

Often we do not want to test whether a sample is drawn from a population with a given mean but instead we want to compare the means of two populations. For example we might wish to know whether boys and girls differ in terms of school attainment

Usually the sample observations provide the basis for decision whether or not there is a difference between the means of two populations. In this chapter we present a method for testing equality between two population means for variables measured on interval or ratio scales

5.2 ASSUMPTIONS

Let us first consider the situation when the populations being compared have the same variances and we are interested in comparing their means on the basis of random samples of sizes n_1 and n_2 from the two populations We can express our assumptions as

(1) Population one is normally distributed with mean μ_1 and standard deviation σ, population two is normally distributed with mean μ_2 and standard deviation σ

(2) A random sample of size n_1 is taken from population one and the sample mean is $\bar{X}_1$ A random sample of size n_2 is taken from population

two and the sample mean is $\bar{X}_2$. The two samples are independent

5.3 TESTING A HYPOTHESIS CONCERNING THE MEANS OF THE POPULATIONS

Again we need to follow the four stages of formulating the hypothesis, selecting and calculating a test statistic, calculating the P-value and comparing with α.

In this case we formulate our hypothesis in the following way.

H_0. $\mu_1 = \mu_2$

H_1 $\mu_1 \neq \mu_2$

To select the test statistic we draw on the fact that under the assumptions the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is $N(\mu_1 - \mu_2, \sigma^2(1/n_1 + 1/n_2))$ If we assume that H_0 is true, then $\dfrac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{1/n_1 + 1/n_2}}$ follows the standard normal distribution. We can look up the P-value in the standard normal tables (Appendix B, Table B.1).

As when comparing a sample mean with a hypothesised population mean, when the population variance is unknown we have to replace σ by S in the test statistic which becomes:

$$\frac{\bar{X}_1 - \bar{X}_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If the null hypothesis is true this test statistic follows the t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Sp is the sample standard deviation based on the samples drawn from the two populations with assumed equal variances. It is called the pooled standard deviation Sp^2 is given by the expression:

$$Sp^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

5.4 F-TEST

In the situation when the two populations have different variances a slightly different statistic is used (its distribution is approximately t).

When we do not know whether the two populations have the same variance or not we can test H_0· $\sigma_1 = \sigma_2$ against H_1: $\sigma_1 \neq \sigma_2$. The assumptions are as in section 5 2 and the test statistic is based on the ratio of the larger sample variance to the smaller sample variance. If H_0 is true the distribution of the test statistics is F. The F-distribution depends on two parameters, both called degrees of freedom, and these reflect the sizes of the samples on which the larger and smaller variances are estimated. (Tables are given in Appendix B, Table B.4).

5.5 GENERAL COMMENTS

(1) Here again we make the distinction between one-tailed tests and two-tailed tests In the former we pre-specify in the alternative hypothesis that the mean of one population is larger than or smaller than the mean of a second In the two-tailed test we simply state that the two means differ and do not specify which is the larger

(2) We began by assuming that the populations being compared were normally distributed. However, when the sample sizes are large enough, the central limit theorem indicates that the sampling distribution of $\bar{X}_1$ and $\bar{X}_1$ will be approximately normal and so the test procedures are similar

Generally in large samples the t-test is fairly robust against non-normality, but this is not necessarily the case in small samples. The F-test to compare two variances is very sensitive to non-normality

(3) In many cases the observations for the two groups are not selected independently but observations are paired For example, the same individuals might be measured at two times In such a case the test statistic is based on the paired difference X_1 - X_2, where X_1 is the first measurement and X_2 is the second. The test statistic follows the t-distribution

5.6 PERFORMING A T-TEST USING SPSS-X

The procedure used in SPSS-X to compare the means of two populations is the T-TEST procedure

For comparison between independent samples, the two populations have to be specified using the GROUP subcommand. The variables to be tested are named in the VARIABLES subcommand.

The RECODE command is used to combine the categories of a variable into two categories

5.7 EXAMPLES

Example 5.1

Variables considered

CROWD16	Number of persons per room at age 16
MATH16	Maths test score at age 16

Let us consider as an example whether overcrowding is associated with children's attainment at school A household is considered overcrowded if it has more than one person per room The attainment considered is the score on a maths test Both variables were measured when the NCDS sample were aged 16.

The data set suggests that there were 865 individuals in non-overcrowded homes - group 1; and 360 individuals with overcrowded homes - group 2.

Graphical methods can be used as a first stage in the analysis. One can draw a separate box plot for each group and then compare the two plots (see Example 3.9). In the next stage several summary measures are calculated The mean test score and the standard error for each group are presented in the following table·

Table 5.1: Maths test scores by overcrowding

Groups	Sample Size n	Mean Maths Test Score	Standard Error
1 Non-overcrowded	$n_1 = 865$	$\bar{X}_1 = 14\ 39$	0 244
2 Overcrowded	$n_2 = 360$	$\bar{X}_2 = 10.85$	0.353

We see that individuals from group 1 have on average a test score higher than those in group 2 If we are willing to restrict the conclusions to the 1225 individuals included in the sample we can say that on average children from overcrowded homes tend to achieve lower scores than children from non-overcrowded homes However, before making a generalisation as to the relationship between overcrowding and Maths attainment for the population of 16-years old individuals living in Britain in 1974, we have to determine whether the difference between the two samples implies a true difference between the two populations. To do this we use the procedure of hypotheses testing We test H_0. $\mu_1 = \mu_2$ against H_1 $\mu_1 \neq \mu_2$ where μ_1 and μ_2 denote the mean Maths test scores for populations 1 and 2 respectively We set $\alpha = 0\ 05$ We assume that $\sigma_1 = \sigma_2$ where σ_1 and σ_2 denote the standard deviation of populations 1 and 2 respectively

The test statistic is $\dfrac{\bar{X}_1 - \bar{X}_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

and its observed value is 8.04 If H_0 is true, the probability of getting a result which is as extreme or more than 8 04 is very small, less than 0 001. Since the P-value is so small we reject the null hypothesis and conclude that the data indicate that 16-years old individuals living in overcrowded homes tend to have lower Maths attainment

In order to see if the assumption that $\sigma_1 = \sigma_2$ is not inappropriate we test H_0. $\sigma_1 = \sigma_2$ against H_1 $\sigma_1 \neq \sigma_2$ The observed value of the statistic is 1.14 This result is not significant (at the 5% significance level) and so we conclude that the assumption was realistic.

All the above information is produced using the SPSS-X T-TEST procedure as presented in job 5.1

```
SPSS-X PROGRAM
TITLE "JOB 5.1"
FILE HANDLE NCDS
GET FILE=NCDS
RECODE CROWD16 (2 3 4 = 2)
T-TEST GROUPS= CROWD16/
          VARIABLE= MATH16
FINISH
```

Example 5.2

Variables considered

CROWD16	Number of persons per room at age 16
READ16	Reading comprehension test score at age 16

Another aspect of individual's attainment at school which may be related to overcrowding is reading attainment In order to compare reading test scores at age 16 of individuals from overcrowded and non-overcrowded homes, we use the SPSS-X program in job 5 2

```
SPSS-X PROGRAM
TITLE "JOB 5.2"
FILE HANDLE NCDS
GET FILE=NCDS
RECODE CROWD16 (2 3 4 = 2)
T-TEST GROUPS= CROWD16/
        VARIABLE= READ16
FINISH
```

An examination of the output of job 5 2 reveals that there is a significant difference between the variances of the two populations We should therefore use the appropriate statistic, for which the P-value is very small and the conclusion is that there is a significant difference, with respect to mean reading test scores, between the two populations. However, the distribution of the variable READ16 is skewed and therefore the reading scores have been transformed so that the distribution shape resembles the standard normal distribution Repeating the analysis using the transformed scores READT16 shows no significant difference between the variances but a significant difference between the means

Example 5.3

Variables considered

CROWD16 Number of persons per room at age 16

GEARNPW Gross earnings per week (from main job) at age 23

Let us consider now whether overcrowding at age 16 is associated with weekly earning at age 23.

There are 742 individuals who lived in non-overcrowded homes at age 16 (group 1) and 279 individuals who lived in overcrowded homes (group 2). The mean weekly earnings and the standard errors for each group are presented in the following table

Table 5.2: Earnings by overcrowding

Group	Sample Size n	Mean Weekly Earnings	Standard Errors
1 Non-overcrowded	$n_1 = 742$	$\bar{X}_1 = 100\ 71$	1 47
2 Overcrowded	$n_2 = 279$	$\bar{X}_2 = 97\ 51$	2 28

There is a difference of 3.20 pounds between the two sample means. Does this difference imply a difference between the two populations

The observed value of the test statistics is 1 16 and the P-value is 0 248 which is not significant at level $\alpha = 0\ 05$ Note the relatively large standard errors

The above information can also be produced using the T-TEST procedure which can produce several comparisons at the same time. This is illustrated by the following command:

```
T-TEST GROUP= CROWD16/
        VARIABLE= MATH16, READ16, GEARNPW
```

An extension of this example is presented as an exercise in section 5 9

5.8 COMMENTS

(1) When two groups are compared on the basis of the mean of a certain variable, significant results can also occur when marked differences exist with respect to an additional variable which is highly related to the first variable. An example of performing the same type of analysis, while controlling for an additional variable is presented in chapter 9

(2) When comparing two groups on the basis of the means of a few variables one should remember that the comparisons presented here are done one at a time Each comparison is conducted with significance level α. Performing comparisons simultaneously on several variables is possible, but it is beyond the scope of this text

5.9 EXERCISE

This exercise is an extension of example 5 3.

The labour market conditions for males and females are usually very different, therefore, repeat the analysis suggested in example 5 3 separately for each sex Next investigate whether there are gender differences with respect to weekly pay. Conduct the suitable analysis, use box plots and report your findings

The observed distribution of the variable GEARNPW is skewed You may transform this variable in order to achieve a more bell shaped distribution Are there any outliers?

CHAPTER VI

FREQUENCY TABLES

6.1 INTRODUCTION

In earlier chapters we described how graphs and summary statistics may be used to highlight particular features of a data set Another way of doing this is by constructing tables A simple form of table is that in which a series of observations are classified by two or more types of characteristics. These are referred to as frequency or contingency tables They are used to display and summarize data Frequency tables are appropriate when data are measured on nominal scales (e.g. sex), or ordinal scales (e.g. social class) or on either interval scales or ratio scales grouped into intervals (e.g. intervals of age or height). We may also be interested in using frequency tables to test hypotheses about the data. This chapter describes how we create tables and test simple hypotheses.

6.2 CREATING FREQUENCY TABLES

Example 6.1

Variables considered

HIGHQUAL	Highest educational qualification at age 23
SEX	Sex of child

Consider first that we are interested in the relationship between gender and qualifications For this example we may be interested in grouping qualifications into two categories, HIGH and LOW. Since in the data set there are more than two possible categories of qualifications we need to combine some of the original categories This is done using the SPSS-X

RECODE command. Table 6.1 presents the frequency table of gender by qualifications with qualifications 1-8 and 9-15 grouped together. The grouped categories HIGH and LOW correspond to 5 O levels or above and below 5 O levels respectively.

Table 6.1: Sex by qualifications

	Qualifications		
Sex	High	Low	Total
Boy	570 (57 2%)	427 (42.8%)	997 (100%)
Girl	422 (42.1%)	581 (57.9%)	1003 (100%)
Total	992	1008	2000

The table was created by SPSS-X using the procedure CROSSTABS The program is displayed in job 6 1.

```
SPSS-X PROGRAM

TITLE "JOB 6.1"

FILE HANDLE NCDS

GET FILE=NCDS

RECODE HIGHQUAL (1 THRU 8 = 1) (9 THRU 15 = 2)

CROSSTABS TABLES= SEX BY HIGHQUAL

OPTIONS 3

FINISH
```

There are numerous options related to the procedure CROSSTABS. For example option 3 produces row percentages and option 4 produces column percentages

Example 6.2

Variables considered

PASCO	Father's social class at child's birth
CURRSOC	Social class of current job at age 23

For our example let us consider social mobility between generations We might wish to classify every sample member according to his father's social class in 1958 when he was born and by his own social class at age 23. Again we might wish to re-group the categories into 4 broader social classes: professional and intermediate, skilled non-manual, skilled manual, semi-skilled and unskilled. The program is displayed in job 6.2.

```
SPSS-X PROGRAM
TITLE "JOB 6.2"
FILE HANDLE NCDS
GET FILE=NCDS
RECODE PASCO (2 = 1) (6 = 5)
RECODE CURRSOC (2 = 1) (6, 7 = 5)
CROSSTABS TABLES= PASCO BY CURRSOC
OPTIONS 3, 5
FINISH
```

The table created in job 6.2 enables us to examine patterns of mobility between generations

Example 6.3

<u>Variables considered</u>

CURRSOC	Social class of current job at age 23
NKIDS11	Number of children under 21 in the family at age 11
PASCO	Father's social class at child's birth

As a final example in this section let us consider the association between family characteristics in childhood and social class in early adulthood We might, for example, wish to study the relationship between the number of children in the family at age 11 and the current social class in order to see whether children from smaller families achieve higher status in early adulthood. We shall re-group the number of children into small families (1 or 2 children) and large families (3 or more children). Job 6.3 contains the SPSS-X program.

```
SPSS-X PROGRAM
TITLE "JOB 6.3"
FILE HANDLE NCDS
GET FILE = NCDS
RECODE CURRSOC (2 = 1) (6, 7 = 5)
RECODE NKIDS11 (1, 2 = 1) (3 THRU 9 = 2)
CROSSTABS TABLES= NKIDS11 BY CURRSOC
OPTIONS 3
FINISH
```

At first glance the table created by job 6.3 indicates that there may be a strong link between these two variables. However, this may only reflect differences due to social class of origin between people in small families and people in big families. In order to see if this is the case we can create four separate tables for each social class at birth. The SPSS-X command (after the appropriate recoding) is:

```
CROSSTABS TABLES= NKIDS11 BY CURRSOC BY PASCO.
```

An examination of the four tables suggests that the answer to our question "are children from small families doing better as adults?" is not unique and depends on the social class at birth.

The method used in example 6.3 shows how we might consider the relation between two variables after controlling for a third variable

6.3 TESTING THE HYPOTHESIS OF INDEPENDENCE

6.3.1 Expected and observed frequencies

In section 6.2 we gave some examples illustrating how one constructs simple frequency tables. We are often interested in assessing whether two variables in a particular frequency table are independent of each other. For example, we might wish to use Table 1 to test the hypothesis that sex and qualifications are independent.

First some simple theory in order to show what the table would be expected to look like if the two variables were indeed independent.

The general two-way frequency table has r rows and c columns. If n observations are taken, let n(i,j) be the number of observations which fall in the i-th row and j-th columns. We denote by

$n(i,+) = \sum_{j=1}^{c} n(i,j)$ the number of observations in the i-th row and by

$n(+,j) = \sum_{i=1}^{r} n(i,j)$ the number of observations in the j-th column.

The probability that an observation falling into category i of the first variable is estimated by n(i,+)/n. The probability of an observation falling into category j of the second variable is estimated by n(+,j)/n. Therefore if these variables are independent the probability of an observation falling into cell (i,j) is estimated by n(i,+)/n x n(+,j)/n. The expected number of observations (out of n) in cell (i,j) under the assumption of independence is therefore

$$E(i,j) = n(i,+)/n \times n(+,j)/n \times n.$$

The expected number of observations in a cell (i,j) can be compared with the observed frequency in cell (i,j) that is n(i,j). We denote the observed frequency by O(i,j). The difference between the observed and expected frequencies is called the residual.

Consider example 6.1. The probability of being a boy is estimated by 997/2000 and the probability of high qualification is estimated by 992/2000. Thus under the assumption of independence the probability of being a boy with high qualifications is estimated by (997/2000) X (992/2000). The expected number of boys with high qualifications is therefore, (997/2000) X (992/2000) X 2000 = 494.5, and the observed number is 570 The residual for this cell is 570 - 494 5 = 75.5 This would suggest that there are more boys with high qualifications than would be expected under the assumption of independence

6.3.2 The test statistic

We can test the hypothesis that the row and column variables are independent by using the Pearson chi-square statistic

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{residual})^2}{\text{expected number}} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O(i,j) - E(i,j))^2}{E(i,j)}$$

The distribution of χ^2 is approximately chi-square distribution with (r-1) x (c-1) degrees of freedom (Tables are given in Appendix B, Table B.3). One can use this distribution to estimate how likely or unlikely the observed value of the statistic is under the assumption that the null hypothesis stating that the two variables are independent is correct.

For table 6 1 the value of χ^2 is

$$\chi^2 = \frac{(570-494.5)^2}{494.5} + \frac{(427-502\ 5)^2}{502\ 5} + \frac{(422-497\ 5)^2}{497\ 5} + \frac{(581-505\ 5)^2}{505\ 5} = 45.59$$

and the degrees of freedom are (2-1) X (2-1) = 1 The P-value from the table of χ^2 is less than 0 01 consequently the hypothesis of independence is rejected

A few comments concerning the χ^2 - statistic

(1) The distribution of the χ^2 statistic is approximately chi-square if the data are random samples from a multinomial distribution and if the expected values are not too small. A conservative view is that all expected frequencies should be at least 5, while a less conservative one recommends that no more than 20% of the cells should have expected values which are less than 5.

(2) The magnitude of the observed value of χ^2 depends on the magnitude of the residuals and on the sample size. Large values can occur when the residuals are small but the sample size is large

(3) The statistic is useful as a measure of the significance of the association It is not at all useful as a measure of the degree of association.

6.3.3 Calculation using SPSS-X

Option 15 of the procedure CROSSTABS prints the cell residuals

STATISTIC 1 of CROSSTABS prints the χ^2 - statistic and the P-value.

Job 6.1A contains a program which will print a frequency table with cell residuals as well as the χ^2 - statistic with its P-value. The variables considered are as in example 6.1 that is SEX and HIGHQUAL

SPSS-X PROGRAM

```
TITLE "JOB 6.1A"

FILE HANDLE NCDS

GET FILE=NCDS

RECODE HIGHQUAL (1 THRU 8 = 1) (9 THRU 15 = 2)

CROSSTABS TABLES = SEX BY HIGHQUAL

OPTIONS 3 15

STATISTICS 1

FINISH
```

6.4 MEASURES OF ASSOCIATION

6.4.1 General

Frequently we wish to describe relationships between two cross-classified variables using simple summary measures. Such measures are generally called measures of association and can provide a useful description of the structure displayed in a two-way table Many such measures of association have been suggested. Here we will consider only two types of measures: measures of prediction and measures of association for ordered categories.

6.4.2 Measures of prediction

Measures of prediction quantify the improvement in prediction of one categorical variable in the table when the value of the second categorical variable is known, relative to when the value of the second variable is not known. The Goodman and Kruskal - Lambda and the uncertainty coefficient - u, are examples of such measures. Let us consider

Lambda. Lambda is suitable in situations when we wish to predict optimally the category of one variable from the category of the second. Consider the relationship between sex and qualifications in Table 6.1. Qualifications can be predicted either (a) by assuming that sex is unknown, or (b) by assuming that sex is known. Lambda measures the proportional reduction in error (PRE) which is defined by:

$$\text{PRE} = \frac{\text{probability of error in (a) - probability of error in (b)}}{\text{Probability of error in (a)}}$$

If we did not know the sex we would predict the most frequent qualifications; in this case LOW. The probability of error would be 992/2000 = 0.496. If the sex were known, we would predict HIGH qualifications for boys and LOW for girls The probability of error would then be 427/2000 + 422/2000 = 0 425. From these figures we can derive Lambda as:

$$\text{Lambda} = \text{PRE} = \frac{0.496 - 0.425}{0.496} = 0\ 143$$

Thus, a 14% reduction in error is obtained when sex is used to predict qualifications.

Lambda ranges between 0 and 1 For each table two Lambdas can be computed depending on which variable is used as the predictor. STATISTIC 4 of procedure CROSSTABS of SPSS-X print Lambdas.

6.4.3 Measures of association for ordered categories

Another group of statistics measures the association and correlation for two-way tables when both variables of the table have ordered categories These use the information about the ordering of categories of variables by considering every possible pair of cases in the table. Each pair is checked to see if their relative ordering of the first variable is the same as their relative ordering on the second, or if it is reversed. If two individuals in a pair happen to have the same value for one or both the variables, then the pair is said to be tied. Gamma, Tau b, Tau c and Somers' d are examples of such measures. The main difference between these measures is the way in which ties are allowed for. Here we will consider Gamma

Let P be the number of pairs for which the relative ordering in both variables is the same. Let Q be the number of pairs for which the relative ordering on one variable is opposite from the relative ordering on the second. Gamma is defined by:

$$\text{Gamma} = \frac{P - Q}{P + Q}$$

Gamma can be interpreted as the probability of similar (dissimilar) ordering on two variables among cases with different values for both variables. Gamma ranges between -1 and +1 and is 0 when the variables are independent. STATISTIC 8 of procedure CROSSTABS of SPSS-X prints Gamma.

Example 6.4

Variables considered

PASCO	Father's social class at child's birth
CURRSOC	Social class of current job at age 23
NKIDS11	Number of children under 21 in the family at age 11

Consider again example 6.3. Both variables, current social class and number of children in the family, have ordered categories We created four separate tables each corresponding to a different social class at birth. A comparison of the level of association in each of the tables can be made using summary measures like Gamma and Tau-c. Job 6.4 contains the SPSS-X program for doing this

```
SPSS-X PROGRAM
TITLE "JOBS 6.4"
FILE HANDLE NCDS
GET FILE=NCDS
RECODE PASCO (2 = 1) (6 = 5)
RECODE CURRSOC (2 = 1) (6, 7 = 5)
CROSSTABS TABLES= NKIDS11 BY CURRSOC BY PASCO
STATISTICS 1 7 8
FINISH
```

6.5 EXERCISE

Doctors claim that a mother to be who smokes during pregnancy risks having a relatively tiny baby. Does the NCDS data support this statement?

Group the values of the variable BIRTHWT into two categories - slim babies and non-slim babies. (As a cut-off point you may use the lower quartile or the lower decile) and analyse the association between the birth weight and the mother's smoking habits. Analyse each sex separately

A T-test could also be used to answer this question. Carry on the relevant analysis and report your finding. Refer also to the general problem of analysing a continuous variable in a categorical framework.

CHAPTER VII

ANALYSIS OF VARIANCE - ONE WAY

7.1 INTRODUCTION

In chapter 5 we considered testing the differences between two means. In this chapter we consider an extended problem of comparing the means of several populations at the same time. For example we may wish to test the hypothesis that there is no difference in average male height between social classes. Analysis of variance is a statistical procedure commonly used to test the hypothesis that several population means are equal. In this chapter we describe the assumptions, present the test statistic and give some examples of analysis of variance

7.2 NOTATION AND ASSUMPTIONS

Assume there are k - populations which constitute the entire set of populations about which conclusions are described. Now assume that k random samples are selected, one from each of the k populations. The sample sizes are n_i, $i = 1, . . ., k$; and $n = \sum_i n_i$. The samples are assumed to be independent and the k populations are assumed to be normally distributed with means $\mu_i, \mu_i, . . ., \mu_k$, and common variance σ^2

The hypothesis of interest is that the population means are equal. That is:

H_0: $\mu_1 = \mu_2 =, . . ., = \mu_k$

H_1: at least two means are not equal.

Let Y be the dependent or criterion variable. We assume Y is measured on an interval or ratio scale.

The observations of the first sample are denoted by

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ and their mean is denoted by $\bar{Y}_1$.

The observations of the second sample are denoted by

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ and their mean is denoted by $\bar{Y}_2$.

The observations of the i-th sample are denoted by

$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ and their mean is denoted by $\bar{Y}_i$.

Let $\bar{\bar{Y}}$ be the mean of the entire sample (i.e. of all samples added together).

7.3 DECOMPOSITION OF VARIATION

The basis of analysis of variance is the decomposition of variation. The total observed variation (in Y) is divided into two components - variation of observations within groups and variation between group means, i e

$$\text{Total variation} = \underset{\text{groups}}{\text{variation between}} + \underset{\text{groups}}{\text{variation within}}$$

In mathematical terms we can show that

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{\bar{Y}})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{\bar{Y}})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

The quantity $\sum_i\sum_j(Y_{ij} - \bar{\bar{Y}})^2$ measures the sum of squared deviations of the observed Y_{ij} from the mean $\bar{\bar{Y}}$. It represents the total variation in Y and is denoted by SSy. The quantity $\sum_i n_i(\bar{Y}_i - \bar{\bar{Y}})^2$ measures variation between the means of the groups. It is denoted by SS Between. The quantity $\sum_i\sum_j(Y_{ij} - \bar{Y}_i)^2$ measures variation within the groups. It is denoted by SS Within. We can therefore write SSy = SS Within + SS Between.

7.4 ANALYSIS OF VARIANCE TABLE

The equation presented in the previous section is the basis for creating an analysis of variance table as below

Table 7.1: ANOVA Table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Squares (MS)
Between groups	SS Between	K - 1	SS Between/(K - 1)
Within groups	SS Within	n - K	SS Within/(n - K)
Total	SSy	n - 1	

The first column indicates the source of variation The next column indicated the sum of squares associated with each source The third column indicates the number of pieces of information needed to calculate the sum of squares, otherwise called the degrees of freedom. For example, the degrees of freedom for the total sum of squares is n-1, because there are n observations but one degree of freedom is needed to represent the estimate of the mean. The final column in the table gives the ratio of the sums of squares to the degrees of freedom. These ratios are known as the mean squares.

7.5 THE TEST STATISTIC

If there is no difference between the groups (with respect to their means) SS Between and SS Within came from the same source. Therefore when H_0 is true both MS Between and MS Within represent (estimate) the same thing - the sample variation in Y - the statistic is based on their ratio. The test statistic is

$$\frac{\text{MS Between}}{\text{MS Within}}$$

If H_0 is true, the statistic follows the F-distribution with (K - 1) and (n - K) degrees of freedom. The P-value is calculated by considering the probability of obtaining an F statistic at least as large as the one observed (tables of the F-distribution are given in Appendix B, table B.4). The P-value is compared with the predesignated significance level and a decision is made accordingly.

The analysis described here is called one way analysis of variance since there is only one grouping variable. An example of a model which includes two grouping variables is presented in chapter 9

7.6 ONE WAY ANALYSIS OF VARIANCE USING SPSS-X

The SPSS-X procedure ONEWAY produces a one way analysis of variance. The command ONEWAY is used to name the dependent variable (s) as well as the group variable followed by its minimum and maximum values The output includes: an ANOVA table, the F - statistic and the P-value.

7.7 EXAMPLES

Example 7.1

Variables considered

HT23	Height at age 23 in cm
CURRSOC	Social class of current job at age 23
SEX	Sex of child

7.7.1 Introduction

Let us consider the difference in average male height between social classes. In the first stage of the analysis we use some graphical methods for example, we plot an histogram of the variable HT23 (for males only) as well as several box plots, one for each social class An examination of the outputs reveals that there are a few outliers; more specifically there are 8 observations for which the height at age 23 is more than 5 meters. Such measurements are impossible and probably are a result of errors in coding. The analysis, is therefore, repeated after the exclusion of these observations. Table 7.2 below contains some basic descriptive statistic for male height at age 23. (The re-grouping of the variable CURRSOC is as in example 6.2).

Table 7.2: Male height (age 23)

Social class	n	MEAN Height	Standard Deviation	Standard Error
Professionals & intermediate	232	178 52	6 53	0.43
Skilled non-manual	150	178.44	6 83	0 56
Skilled manual	365	177.22	6.97	0.37
Semi-skilled & unskilled	192	176.16	7 25	0.52
Total	939	177.52	6.95	0 23

There seems to be a clear gradient of social class in height where males in the first social class category are on average more than 2cm taller than males in the last category.

The sample means ($\bar{Y}_i$) are good estimates for the population means (μ_i). We want to ask whether the observed differences between the sample means can be attributed to chance or do they indicate true differences between the four social classes.

7.7.2 Analysis and conclusions

We are interested in testing

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

against

H_1: at least two means are not equal.

Where μ_i is the mean height of males in the i-th social class

$i = 1, . \quad ., 4$

It is clear that the observed height of males varies - the sample standard deviation is 6.95 cm. The question is whether the variation between the different social classes is of the same magnitude as the variation within the social classes. This can be examined from the ANOVA table

Table 7.3: ANOVA Table

Source	Sum of Squares	Degrees of Freedom	Mean Squares
Between groups	750.31	3	250.10
Within groups	44568.83	935	47.67
Total	45319 14	938	

From the table we can also calculate the value of the F - test statistic which is 250.10/47 67 = 5 246 The P-value is less than 0 01 and therefore we can conclude that males in the four social classes do not have the same mean height

Both the descriptive statistics and the ANOVA table can be printed using the SPSS-X procedure ONEWAY Job 7 1 contains the program.

```
SPSS-X PROGRAM
TITLE "JOB 7.1"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (SEX EQ 1)
SELECT IF (HT23 LT 300)
RECODE CURRSOC (1 2 = 1) (3 = 2) (4 = 3) (5 6 7 = 4)
ONEWAY HT23 BY CURRSOC (1, 4)/
STATISTICS 1
FINISH
```

Example 7.2

Variables considered

LEFTED Age completed full-time education (month no.)

VOTED Party voted for in 1979 general election

Next let us compare individuals grouped according to their voting patterns in the 1979 general election. The comparison is done with respect to the age, in months, individuals left full-time continuous education (the time spent in full-time education seems positively correlated to high social class and higher earnings). Some general statistics are displayed in table 7.4

Table 7.4: Age (in months) respondent left full-time continuous education

Voted	n	Mean Age	Standard Error
Conservative	507	215.40	1.13
Labour	569	205.48	0 84
Liberal	172	213.69	1 87

The table displays sample differences in the mean age for leaving full-time education where the conservative voters stayed longer in continuous full-time education. To print the related ANOVA table use the SPSS-X command:

```
ONEWAY LEFTED BY VOTED (1, 3).
```

Separate analysis for males and females should also be considered.

7.8 MULTIPLE COMPARISONS

In both examples presented the null hypothesis was rejected and we conclude that the data indicate that the population means are unequal Typically, we are not only interested in determining whether there are differences in the means but also in pinpointing where the differences lie. For example, given that height differs significantly among social class categories, we would next want to determine which of the pairs of social class categories differ from one another

When considering several population means it is not appropriate statistically to carry out a sequence of tests of differences between pairs of means based on the t-distribution (i.e. t-test). If one does so, the probability of incorrectly rejecting any one of several null hypotheses is greater than the level of significance associated with each test individually. However, a number of types of tests are available for multiple comparison following the rejection of the null hypothesis in the analysis of variance. Examples of such tests are: LSD (least significant difference) test, Tukey's Honestly Significant Difference test and Scheffe's test. All these are systematic procedures for comparing all possible pairs of group means while "protecting" against calling too many differences significant A number of multiple comparison tests are available in SPSS-X. The subcommand is RANGES, as for example RANGES = LSD or RANGES = SCHEFFE.

7.9 CHECKING THE ASSUMPTIONS

As with other statistical procedures the analysis of variance relies on a number of assumptions. If any one of these assumptions is false some of the analysis may be invalid. It is therefore desirable to check whether or not the assumptions hold.

The assumption of normality can be checked, for example, by plotting a histogram. A second look at examples 7.1 and 7.2 reveals that the assumption of normality is inappropriate in example 7.2. Fortunately the F-test is reasonably robust to departures from the normality assumption.

Nevertheless, a transformation of the data would be valuable and one can try the logarithmic transformation.

The assumption of equality of variances can be tested by several test procedures. Three such test procedures are available in SPSS-X. The F-test is not very sensitive to departures from the equality of variance assumption when the sample sizes of the groups are similar. STATISTICS 3 of procedure ONEWAY prints the tests for equality of variance. The assumption of homogeneity of the variance is inappropriate in example 7 2 As a result one may need to consider another way of analysing example 7.2, for example grouping the data and constructing a frequency table.

Example 7.3

<u>Variables considered</u>

SEX	Sex of child
SCHLT16	School type at age 16
MATH16	Maths test score at age 16

For the last example in this chapter we compare educational attainments for cohort members who studied at different types of schools The measure of educational attainment we use is the Maths test scores at age 16. We consider comprehensive, grammar, secondary modern and independent schools. The last category includes direct grant schools.

The null hypothesis states that there is no difference in the mean Maths score between the different types of schools. The analysis is done separately for boys and girls.

The output of the SPSS-X program displayed in job 7.3 includes the following:

- General descriptive statistics for Maths test scores in each type of school.

- An ANOVA table for testing the null hypothesis.

- A histogram of Maths test scores, for checking the assumption of normality.

- 3 test procedures for testing homogeneity of the variance.

- Multiple comparison test (Scheffe) to pinpoint significant differences between all pairs of types of schools

```
SPSS-X PROGRAM
TITLE "JOB 7.3"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (SEX EQ 2)
FREQUENCIES VAR= MATH16/ HISTOGRAM/
            FORMAT= NOTABLE
ONEWAY MATH16 BY SCHLT16 (1, 4)/
            RANGE= SCHEFFE/
STATISTICS 1 3
FINISH
```

The completion of the analysis of this example is left as an exercise.

7.10 EXERCISE

Complete the analysis of the problem presented in example 7.3 and report your findings. In your report address yourself to the following:

- The null and alternative hypothesis, the test statistic, the conclusions, the assumptions and the differences between males and female

CHAPTER VIII

REGRESSION ANALYSIS

8.1 INTRODUCTION

In this section we consider the situation in which simultaneous measurements are taken on two variables and interest is focused on examining the effect one variable exerts on the other

For physical processes there is often a functional relationship between two variables We are concerned here with the cases where these functional relationships may be approximated by a linear function.

When no physical relationship exists between two variables, we may also be interested in expressing the relationship between them by a linear equation. Such an equation can be useful for predicting the value of one variable from the knowledge of the other This chapter explains how we go about estimating or predicting the value of one variable on the basis of knowledge of a second variable

8.2 PRELIMINARY EXAMPLES

Let us label the two variables as X and Y Often it is of interest to examine the effect one variable (X) exerts on another variable (Y). For example we may be interested to see how changes in height (X) affect the weight (Y), how qualifications (X) affect income (Y); and how school attainment at age 11 (X) predicts school attainment at age 16 (Y).

In all these examples and in many other situations the question of interest is how changes on one variable (X) affect another variable (Y). The X- variable is called the independent variable. It can be set to a particular value or else take values that can be observed but not controlled. As a result of changes in the independent variable an affect may be introduced into the Y-variable, called the dependent variable. The dependent variable is assumed to be measured at least on an interval scale.

8.3 SCATTER PLOT

Suppose n pairs of measurements $(X_1\ Y_1)$ $(X_2\ Y_2)$, . . ., $(X_n\ Y_n)$ are taken on two variables X and Y

The first step in investigating the relation between the two variables is to plot the data on a scatter plot. In this plot each dot represents one observation, showing the values for this observation on the two variables. A scatter plot can reveal various types of associations between the two variables. If the observed points cluster, more or less, around a straight line we say that a linear relationship exists between the two variables. Scatter plots are also useful for detecting outliers.

Example 8.1

Variables Considered

HT11	Child's height at age 11 in cm
WT11	Child's weight at age 11 in kgs
SEX	Sex of child

Let us first construct a scatter plot of males' height and weight Scatter plots can be printed using the SPSS-X SCATTERGRAM procedure as presented in job 8 1

```
SPSS-X PROGRAM
TITLE "JOB 8.1"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (SEX EQ 1)
SCATTERGRAM WT11 WITH HT11
FINISH
```

Inspection of the plot suggests that weight increases with height It is clear that for a given height there is a variation in the observed weight. This variation is due mainly to variation between individuals Although no unique linear equation appears to relate the observed weight to height, we can notice that the average value of weight (for a given height) increases with the height. Thus we can fit a line to the data in order to predict the mean weight for a given height. The line is called a regression line

8.4 FITTING A STRAIGHT LINE

As in the example of height and weight, a straight line can frequently express the dependency of one variable upon another. The problem is to use the data to find the optimal line

Let us assume that the line has the form $Y = \alpha + \beta X$.

Our model is $Y = \alpha + \beta X + \epsilon$.

The model implies that for a given X, the corresponding Y consists of two elements: the value $\alpha + \beta X$ and the amount ϵ. The value $\alpha + \beta X$ is the linear part while ϵ represents the amount by which any individual Y may fall off the line α and β are called the model's parameters and they are unknown. The data can be used to estimate the parameters

8.4.1 Estimation of the parameters

Suppose n pairs of observations $(X_1\ Y_1)$, $(X_2\ Y_2)$, . .., $(X_n\ Y_n)$ are given. The model for the i-th individual is: $Y_i = \alpha + \beta X_i + \epsilon_i$.

One approach is to choose estimates of α and β which minimise the sum of squared deviations from the line. This approach is called the method of least squares and the estimates are called least squares estimates.

Least squares estimates of α and β are denoted by $\hat{\alpha}$ and $\hat{\beta}$ and by means of differential calculus they can be shown to be:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

We can now write our fitted line as: $\hat{Y} = \hat{\alpha} + \hat{\beta}X$.

Where $\hat{Y}$ is the predicted (or fitted) value of Y, given X The differences between the observed and fitted values $Y_i - \hat{Y}_i$ for $i = 1, . \quad ., n$ are called the residuals.

8.5 ASSUMPTIONS IN REGRESSION ANALYSIS

There are some questions one might want to ask about our model

$Y_i = \alpha + \beta X_i + \epsilon_i \quad i = 1, \quad ., n.$

For example how close are the estimates of parameters to the population values of those parameters? Do the estimates support or contradict hypotheses about the population parameters? How good is the fit of this model?

As a basis for doing this we assume that the ϵ_i are independent normally distributed random variables with mean zero and variance σ^2

The model and the assumptions made may be written in the following way: for any fixed value of X, the distribution of Y is normal with mean $\alpha + \beta X$ and constant variance σ^2 and the Ys are independent from each other. We can therefore conclude that for a fixed value of X, Y varies and the mean of the variable Y is $\alpha + \beta X$. Thus the regression line $Y = \alpha + \beta X$ joins the mean values of the Y-distributions.

8.6 THE ANALYSIS

8.6.1 The estimates

The least squares estimates were presented in Section 8.4. Given the assumptions about the distribution of the error terms, the mean and standard error of the estimates can be shown to be

$$\text{mean } (\hat{\alpha}) = \alpha$$

$$\text{S.E. } (\hat{\alpha}) = \sqrt{\sigma^2 \frac{\sum_i X_i^2}{n\sum_i (X_i - \bar{X})^2}}$$

$$\text{mean } (\hat{\beta}) = \beta$$

$$\text{S.E. } (\hat{\beta}) = \frac{\sigma}{\sqrt{\sum_i (X_i - \bar{X})^2}}$$

Given the assumptions about the distribution of the error term, it can also be shown that the least squares estimates are the best among a large group of possible estimates.

8.6.2 Testing hypotheses

We are frequently interested in testing the hypothesis that the slope of the line is zero as would be the case if there were no linear relationship between X and Y

We can formulate the null and alternative hypotheses as

H_0. $\beta = 0$

H_1. $\beta \neq 0$

It can be shown that (given the assumptions made) $\hat{\beta}$ follows a normal distribution with mean β and variance $\sigma^2/\sum_i(X_i - \bar{X})^2$. The appropriate test statistics is:

$$\frac{\hat{\beta}}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

which under the null hypothesis has a t-distribution with (n - 2) degrees of freedom.

The unknown variance σ^2 is estimated by $\hat{\sigma}^2$ which is given by.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n - 2}$$

8.6.3 The fit of the line

Now let us consider how well the estimated regression line fits the data. We do this by splitting the total variation in Y into the variation explained by the regression and the unexplained variation, i.e

Total variation = variation explained + Unexplained
by regression variation (= residuals).

In mathematical terms we can show that:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

The term $\sum_i (Y_i - \bar{Y})^2$ measures the sum of squared deviations of the observations from the mean It represents the total variation in Y. The term $\sum_i (\hat{Y}_i - \bar{Y})^2$ measures variation of $\hat{Y}_i$ and represents the variation explained by the linear relationship. The term $\sum_i (Y_i - \hat{Y}_i)^2$ is the sum of the squared residuals.

Our model fits well if the explained variation is much greater than the unexplained variation or if the ratio of the explained variation to the total variation is near to one This ratio is called the coefficient of determination and is denoted by R^2 The coefficient R^2 lies between 0 and 1 and measures the prediction accuracy.

8.6.4 Analysis of variance table

The equation presented in the previous section is the basis for creating an analysis of variance table as below. (See section 7.4).

Table 8.1: ANOVA table

Source of variation	Sum of Squares	Degrees of freedom	Mean Squares
Regression	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2/1$
Residual	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	n - 2	$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2/(n - 2)$
Total	$\sum_{i=1}^{n} (Y_i - \bar{Y})^2$	n - 1	

It can be shown that if $\beta = 0$ then the ratio

$$F = \frac{\text{Mean square regression}}{\text{Mean square residual}}$$

follows the F-distribution with 1 and (n - 2) degrees of freedom. This fact can be used to test if $\beta = 0$. The F statistic can be used also to test the null hypothesis that R^2 in the population equals zero. (Tables of the F-distribution are given in Appendix B, table B.4).

8.7 REGRESSION ANALYSIS USING SPSS-X

The SPSS-X procedure for doing regression is REGRESSION. Three subcommands are necessary. A VARIABLES subcommand which includes all the variables to be analysed; a DEPENDENT subcommand that identifies the dependent variable; an ENTER subcommand that identifies the independent variable.

The output includes: the least square estimates, the test statistics and the P-values to test whether each parameter is equal to zero, the coefficient of determination R^2 (R is referred to as multiple R) and an analysis of variance table Other statistics are optional.

Example 8.2

Variables considered

PAHT	Father's height in inches
HT23	Height at age 23 in cm.
SEX	Sex of child

Height of both children and adults serves as a reasonable measure of the individual's general health In this example we consider the heredity effect in height We investigate how changes in the father's height affects son's height

As the father's height was recorded in 1958 and the son's in 1981 it is not surprising that the former is in inches and the latter in centimetres. For simplicity we transform the father's height into centimetres The transformation is done using the SPSS-X COMPUTE command.

In an earlier analysis of variables HT23 (see section 7.7) it was found that there are a few observations for which the height at age 23 is more than 5 metres. These observations are excluded from the analysis. The program is displayed in job 8.2.

```
SPSS-X PROGRAM

TITLE "JOB 8.2"

FILE HANDLE NCDS

GET FILE=NCDS

SELECT IF (HT23 LT 300)

SELECT IF (SEX EQ 1)

COMPUTE PAHT = 2.54* PAHT

SCATTERGRAM HT23 WITH PAHT

REGRESSION VAR= HT23, PAHT/ DEPENDENT= HT23/ ENTER PAHT

FINISH
```

The following results can be found in the output.

The estimates

$\hat{\alpha}$ = 107 043

$\hat{\beta}$ = 0.404

Each parameter is significantly different from zero

The fitted values

The equation for the fitted regression line is

Son's height = 107.043 + 0.404 (father's height)

This equation enables us to predict the height of an individual given his father's height We see that a change of one cm in the father's height results in a change of 0.4 cm in the son's height Note that the prediction is only meaningful for such values of the independent variable which are in the range of the values of this variable found within this sample. It would be wrong to predict that the height of a son is 147.4 cm given his father's height is 1 metre since we have no father as small as this and cannot assess whether a linear relationship continues down to this level

<u>The fit of the regression model</u>

The decomposition of the variance (presented in the analysis of variance table) is as follows

total variation	=	Variation Explained by regression	+	Variation Unexplained
39343 77	=	6750.59	+	32593.18

The coefficient of determination R^2, is equal to 0.17, which implies that 17% of the total variation about the mean is explained by the regression. The F-statistics is (6750.59/1) / (32593.18/815) = 168.8 which is significant at the 0 01 level

We would therefore reject the null hypothesis that there is no linear relationship between son's height and father's height.

Example 8.3

Variables considered

MAHT	Mother's height in inches
HT23	Height at age 23 in cm.
SEX	Sex of child

Consider the same question, but this time for female cohort members and their mothers The dependent variable is HT23 and the independent variable is MAHT. The SPSS-X program is displayed in job 8.3.

SPSS-X PROGRAM

```
TITLE "JOB 8.3"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (SEX EQ 2)
COMPUTE MAHT= 2.54* MAHT
SCATTERGRAM HT23 WITH MAHT
REGRESSION VAR= HT23, MAHT/ DEPENDENT= HT23/ ENTER MAHT
FINISH
```

An inspection of the output reveals that in this case there are also a few outliers which seem to result from errors in coding. One should repeat the previous analysis after removing or replacing these observations

8.8 THE ASSUMPTIONS - A RECONSIDERATION

8.8.1 Introduction

The regression analysis described here and the various significance tests are based upon the model outlined in section 8 5 and the assumptions about the error terms

The model together with the assumptions imply that

(1) There is a linear relationship between X and Y

(2) ϵ_i are independent random variables which follow the normal distribution with mean 0 and constant variance σ^2

An important part of every regression analysis is to check the appropriateness of these assumptions. We check the assumption of linearity by a scatter plot. The assumptions about the error terms, ϵ_i, are assessed by examining the residuals from the fitted line.

8.8.2 Examination of residuals

The residuals are defined as the n differences $Y_i - \hat{Y}_i$ $i = 1, \quad , n$ The residuals are the amount not explained by the regression equation and if the model is correct, represent the observed errors of the model

If our fitted model and the assumptions are appropriate the residuals should not display characteristics which appear to contradict the assumptions Residual plots are extremely useful for checking whether this is so Some of the ways of plotting the residuals are

(1) Plot of the residuals against the fitted value - if the assumptions are met, a horizontal band of randomly distributed residuals is expected If the spread of the residuals increases or decreases with the value of $\hat{Y}$, the assumption of constant variance does not hold

(2) Plot of the residuals against the independent variable - here again a horizontal band of residuals is regarded as satisfactory while often patterns indicate that the assumptions have been violated.

(3) Histogram - the errors ϵ_i were assumed to be normally distributed with mean zero and variance σ^2 Therefore the ϵ_i/σ should follow the standard normal distribution. We can examine the standardized residuals, that is residual/$\hat{\sigma}$, to see if they resemble observations from the standard normal distribution This can be done by constructing a histogram of the standardized residuals

Residual plots are also very useful for detecting outliers However, one should be careful as a large residual can also be the result of a wrong model

8.8.3 Constructing residual plots using SPSS-X

A variety of residual plots are possible within the REGRESSION procedure The RESIDUALS subcommand can be used to print a histogram of the standardized residuals (labelled ZRESID). The histogram is presented with a superimposed normal curve. The SCATTERPLOT subcommand can be used to construct different scatterplots.

Example 8.4

Consider the two regression models presented in examples 8 2 and 8.3 (after removing the individuals who are taller than 5 metres!). We will now examine each of these models for violations of the assumptions. The SPSS-X program is displayed in job 8.4

```
SPSS-X PROGRAM

TITLE "JOB 8.4"

FILE HANDLE NCDS

GET FILE=NCDS

SELECT IF (SEX EQ 1)

SELECT IF (HT23 LT 300)

COMPUTE PAHT= 2.54* PAHT

REGRESSION VAR= HT23 PAHT/ DEPENDENT= HT23/ ENTER PAHT/

           RESIDUALS= SIZE (SMALL) HISTOGRAM (ZRESID)/

           SCATTERPLOT= (*RES, *PRED) (HT23, PAHT) (*RES, PAHT)

FINISH
```

Note that two subcommands were added to the REGRESSION procedure The first requests a small histogram of the standardized residuals while the second requests 3 scatter plots (asterisks denote temporary variables). The program for the second model (i e the females) is quite similar.

The outputs indicate that the linear relation seems not to be very strong in both cases, although it is slightly stronger in the females' case

The plot of the residuals against the fitted values seems satisfactory in both cases This is also true for the plot of the residuals against the independent variable. The histogram of the standardized residuals seems more satisfactory in the males' case than in the females' case Note also the outliers in the females' case

Example 8.5

<u>Variables considered</u>

MATH11	Maths test score at age 11.
MATH16	Maths test score at age 16.
READ11	Reading Comprehension test score at age 11
READ16	Reading Comprehension test score at age 16

In this example we aim at predicting school attainment at age 16 using school attainment at age 11 We consider separately Maths attainment and reading attainment. The SPSS-X program is displayed in job 8.5.

```
SPSS-X PROGRAM
TITLE "JOB 8.5"
FILE HANDLE NCDS
GET FILE= NCDS
REGRESSION VAR= MATH11, MATH16/ DEPENDENT= MATH16/ ENTER MATH11/
      RESIDUALS= SIZE (SMALL) HISTOGRAM (ZRESID)/
      SCATTTERPLOT= (*RES, *PRED) (MATH16, MATH11) (*RES, MATH11)

REGRESSION VAR= READ11, READ16/ DEPENDENT= READ16/ ENTER READ11/
      RESIDUALS= SIZE (SMALL) HISTOGRAM (ZRESID)/
      SCATTERPLOT= (*RES, *PRED) (READ16, READ11) (*RES, READ11)
FINISH
```

The output of job 8 5 displays the estimated parameters for each model, which can be used for prediction. The output reveals that both for reading and Maths the linear relationships between the scores at age 11 and the scores at age 16 are very strong.

For the Maths scores the check of the model's assumptions is satisfactory but this is not the case for the reading scores The scatterplot of READ16 against READ11 indicates that the test at age 16 was less sensitive mainly at the high levels. When we consider the histograms of MATH16 and READ16 it is clear that, although neither plots resemble exactly a normal plot, the deviations are larger for READ16.

When the assumptions seem to be incorrect one should consider transforming the data. The reading scores at age 16 have been transformed to follow a normal distribution The transformed scores are given in the variable READT16 The completion of this example using the transformed scores is left as an exercise (see section 8.10).

8.9 CORRELATION

8.9.1 The correlation coefficient

Regression analysis is concerned with estimating or predicting the value of the dependent variable given the value of the independent variable. We now consider a situation in which measurements are taken simultaneously on two variables and the degree of association between them is of interest. Both variables are random variables

An important measure of the degree of linear association between two variables is Pearson's correlation coefficient. Consider n pairs of measurements $(X_1\ Y_1)$ $(X_2\ Y_2)$, . . ., $(X_n\ Y_n)$ the observed correlation coefficient is denoted by r and given by

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

The correlation coefficient, r, takes values in the range -1, +1. A positive correlation coefficient indicates that large values of one variable are associated with large values of the other, while a negative correlation coefficient indicates that the relationship is inverse. The absolute value of the coefficient indicates the "strength" of the association, that is the degree to which variation in one variable is related to variation in the other The coefficient can be used to compare the strength of the relationship between two pairs of variables

The correlation coefficient is directly related to the coefficient of determination in regression analysis where $r = R$.

8.9.2 Testing hypothesis

We are frequently interested in testing the hypothesis that the population correlation coefficient is zero, i.e that there is no linear relationship between the two variables

Under a certain assumption (i.e that the pairs of measurements follow a bivariate normal distribution) the test statistic is

$$t = r\sqrt{\frac{n - 2}{1 - r^2}}$$

Under the null hypothesis this statistic follows a t-distribution with n - 2 degrees of freedom

8.9.3 Computing correlation coefficient using SPSS-X

In SPSS-X the correlation coefficient can be calculated using the REGRESSION procedure or the PEARSON CORR procedure. The latter enables us to print a matrix of the observed correlation coefficient between every pair of variables mentioned.

Example 8.6

Variables considered

MATH11	Maths test score at age 11
MATH16	Maths test score at age 16
READ11	Reading comprehension test score at age 11
READ16	Reading comprehension test score at age 16
ATTEN16	Child's school attendance at age 16
DRAW7	Draw-a-man score at age 7

We consider the association between several variables, all of them related to education

The output of the SPSS-X program which is displayed in job 8.6 contains A matrix of observed correlation coefficients between every pair of variables; and the P-value for testing the hypothesis that the population correlation coefficient is zero

SPSS-X PROGRAM

```
TITLE "JOB 8.6"
FILE HANDLE NCDS
GET FILE=NCDS
PEARSON CORR MATH11, MATH16, READ11, READ16, ATTEN16, DRAW7
FINISH
```

8.9.4 Comments concerning the correlation coefficient

(1) A high correlation coefficient does not necessarily indicate that two-variables are related as there might be a third variable which is associated with both variables

(2) The correlation coefficient can be misleading when the relationship between the two random variables is not linear

8.10 EXERCISE

This exercise is an extention of example 8 5 Conduct a regression analysis where the dependent variable is READT16 and the independent one is READ11 and report your findings. In the report address yourself to the following· the estimated parameters, the fit of the model and the appropriateness of the assumptions made.

CHAPTER IX

MULTIVARIATE ANALYSIS

9.1 INTRODUCTION

In chapters 7 and 8 we analysed a dependent variable measured on at least an interval scale In the simple regression model we analysed the effects of one continuous independent variable. In the one way analysis of variance the effects of one categorial variable called also a factor, were of interest.

Both models can be extended and combined The extension of a regression model so that it includes two or more (continuous) independent variables is called multiple regression The extension of the analysis of variance so that it includes n factors is called n-way analysis of variance When interest is focused on the effects of both categorial and continuous variables the analysis is called analysis of covariance and the continuous variable is called covariate

All the three models - multiple regression, n-way analysis of variance and analysis of covariance - are too complicated to be described in a short and compact way. A full and comprehensive description is beyond the scope of this text We, therefore devote to every topic an example and a related SPSS-X program

9.2 MULTIPLE REGRESSION

Example 9.1

Variables considered

NEARNPW	Net earnings per week at age 23
MATH16	Maths test score at age 16
HIGHQUAL	Highest educational qualification at age 23
WORKHRS	Current job hours of work at age 23
UNEMTIME	Total months ever unemployed
SEX	Sex of child

A considerable amount of research has been devoted in the last decade to the issue of unemployment. For this example we consider the effects of the length of unemployment, the level of qualifications and the Maths test score (age 16) on net weekly earnings. In the analysis we include only those individuals who were in full-time employment at age 23. The analysis is done separately for males and females.

The dependent variable (Y) is NEARNPW. The independent variables are: UNEMTIME (X_1) HIGHQUAL (X_2) and MATH16 (X_3). The model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

where α, β_1, β_2, and β_3 are the model's parameters and the assumptions concerning the error terms (ϵ's) are as in the simple regression model

An examination of the observed distribution of the variable NEARNPW, the dependent variable, reveals that the distribution is skewed and does not resemble the shape of the normal curve The logarithmic transformation is used to overcome this problem when analysing the females (note the outliers when analysing the males).

In the analysis we might be interested in some or all the following issues

- Is there a linear relationship between the dependent variable and the independent variables?

- Estimation of the model parameters

- Testing the hypothesis that a certain parameter is equal to zero

- Determining the goodness of fit of the model

- Determining the relative importance of each independent variable

- Checking the appropriateness of the model's assumptions.

The output of the SPSS-X program displayed in job 9.1 will help us in the analysis and in answering many of these questions

```
SPSS-X PROGRAM
TITLE "JOB 9.1"
FILE HANDLE NCDS
GET FILE=NCDS
SELECT IF (SEX EQ 2)
SELECT IF (WORKHRS GT 29)
COMPUTE NEARNPW= LN(NEARNPW)
FREQUENCIES VAR= NEARNPW/
            HISTOGRAM/
            FORMAT= NOTABLE
PEARSON CORR NEARNPW, HIGHQUAL, UNEMTIME, MATH16
REGRESSION VAR= NEARNPW, HIGHQUAL, UNEMTIME, MATH16/
           DEPENDENT= NEARNPW/
           STEPWISE/
           RESIDUALS= SIZE (SMALL) HISTOGRAM, (ZRESID)/
           SCATTERPLOT= (*RES, *PRED)
FINISH
```

In the REGRESSION procedure we used the subcommand STEPWISE Stepwise is a procedure for selecting independent variables to the model. With a given set of independent variables one can construct a variety of regression models. There are some commonly used procedures for selecting the variables, stepwise, forward and backward are examples of such procedures Generally in stepwise selection the first variable considered for entry into the model is the one with the largest (positive or negative) correlation with the dependent variable. The second variable considered for entry into the model is the one with the largest correlation with the dependent variable while adjusting for the effect of the first independent variable and so on

9.3 TWO WAY ANALYSIS OF VARIANCE

Example 9.2

Variables considered

PARINT7	Parental interest in child's education at age 7
TENURE11	Housing tenure at age 11
READ11	Reading comprehension test score at age 11

The NCDS data show that parental interest in the child's education (at age 7) varies There are only 19 couples who were over concerned while 194 couples showed very little interest An interesting question is whether the level of parental interest affects the child's attainments at a later age, for example his reading ability at age 11

School attainments are highly related to social class or tenure, therefore, we will consider the effects of both tenure (age 11) and parental interest (age 7) on the reading test score at age 11.

The following questions are of interest

- Is tenure related to reading score?

- Is parental interest related to reading score?

- Is there an interaction between the effects of parental interest and tenure?

- Estimating the mean reading test score for each category of parental interest, with and without adjustment for tenure

The statistical procedure used to answer such questions is 2 - way analysis of variance, which is an extension of the one way analysis of variance model described in chapter 7

Again, an analysis of variance table is constructed by the decomposition of variation The table is used to test different hypotheses concerning the population

The model assumptions are similar to those in the one-way model, that is normality and homogeneity of the variance We also assume that both variables (parental interest and tenure) are considered as fixed.

The output of the *SPSS-X* program displayed in job 9 2 includes the relevant analysis of variance table, adjusted and unadjusted mean reading test scores as well as histogram to check the appropriateness of the assumption of normality

SPSS-X PROGRAM

```
TITLE "JOB 9.2"

FILE HANDLE NCDS

GET FILE=NCDS

RECODE TENURE11 (3, 4 = 3) (5 = 4) (6 = -1)

FREQUENCIES VAR= READ11/ FORMAT= NOTABLE/ HISTOGRAM

ANOVA READ11 BY TENURE11 (1, 4) PARINT7 (1, 7)

STATISTICS 1, 3

FINISH
```

9.4 ANALYSIS OF COVARIANCE

Example 9.3

<u>Variables considered</u>

HT23	Height at age 23 in cm
PAHT	Father's height in inches
CURRSOC	Social class of current job at age 23
SEX	Sex of child

In chapter 7 example 7 1, we analysed differences in male height among different social classes. We concluded that it is rather unlikely that men in the four social class categories have on average, the same height In chapter 8, example 8 2, we analysed the relationship between fathers' and sons' height and concluded that some 20% of the variation in male height can be explained by the father's height

We will now consider again the question of differences in male height between different social classes but this time after adjusting for the father's height

The statistical procedure which can be used to answer such a question is analysis of covariance. Generally, regression procedures are used to remove variation in the dependent variable (height) due to the covariate (father's height) and an analysis of variance is then performed on the "correlated" scores The dependent variable is HT23, the covariate is PAHT and the factor is CURRSOC

We assume that the interaction between father's height and social class is zero A test of this assumption can be performed using the SPSS-X MANOVA procedure

The output of the SPSS-X program displayed in job 9.3 includes the following·

- An analysis of variance table for a model which includes the social class category and the father's height (as a covariate).

- A multiple classification analysis (MCA) which presents for each social class category both the unadjusted and the adjusted deviations from the the grand mean. (The adjustment is done for the covariate).

```
SPSS-X PROGRAM

TITLE "JOB 9.3"

FILE HANDLE NCDS

GET FILE=NCDS

SELECT IF (SEX EQ 1)

SELECT IF (HT23 LT 300)

RECODE CURRSOC (1, 2 = 1) (3 = 2) (4 = 3) (5, 6, 7 = 4)

ANOVA HT23 BY CURRSOC (1, 4) WITH PAHT/

STATISTICS 1

FINISH
```

APPPENDIX A

A.1 VARIABLE LIST

A list of variable names and labels.

VARIABLE LABELS	SEX	Sex of child
	NATIONO	Nation at birth
	BIRTHWT	Child's weight at birth in grams
	MASMOKE	Whether mother smoked in pregnancy
	MAHT	Mother's height in inches
	PAHT	Father's height in inches
	PASCO	Father's social class at child's birth
	READ7	Southgate reading score at age 7
	DRAW7	Draw-a-man score at age 7
	PARINT7	Parental interest in child's education at age 7
	HT11	Child's height at age 11 in cm
	HT11	Child's weight at age 11 in kgs
	READ11	Reading comprehension test score at age 11
	MATH11	Mathematics test score at age 11
	TENURE11	Housing tenure at age 11
	NKIDS11	No of children under 21 in family at age 11
	CROWD16	Number of persons per room at age 16
	READ16	Reading comprehension test score at age 16
	READT16	Transformed reading comprehension test score at age 16
	MATH16	Mathematics test score at age 16
	ATTEN16	Child's school attendance at age 16
	LIKES16	I do not like school at age 16
	SCHLT16	School type at age 16
	HT23	Height at 23 in cm
	LEFTED	Age completed FT education (month no)
	HIGHQUAL	Highest educational qualification at age 23
	WORKHRS	Current job hours of work at age 23
	NEARNPW	Net earnings per week (from main job) at age 23
	GEARNPW	Gross earnings per week (from main job) at age 23
	CURRSOC	Social class of current (or last) job at age 23
	UNEMTIME	Total months ever unemployed
	VOTED	Party voted for in 1979 general election

The data set also includes an additional variable the ID number (referred to also as CASEID) which is used as an identifier.

A.2 A DETAILED VARIABLE LIST

Variable names, codes and brief descriptions

SEX	Sex of child (1) Male (2) Female (-1) No answer
NATIONO	Nation at birth (1) England (2) Wales (3) Scotland (-1) No answer
BIRTHWT	Child's weight at birth in grams Range 1106 - 5330 (-1) No answer
MASMOKE	Whether mother smoked in pregnancy (1) Nonsmoker (2) 1 to 4 daily (3) 5 to 9 daily (4) 10 to 14 daily (5) 15 to 19 daily (6) 20 to 24 daily (7) 25 to 29 daily (8) 30 or more daily (9) Varies (-1) No answer
MAHT	Mother's height in inches (self reported) Range 54 - 72 (-1) No answer
PAHT	Father's height in inches (reported by mother) Range 50 - 76 (-1) No answer
PASCO	Father's social class at child's birth (1) Professional (2) Intermediate (3) Skilled non-manual (4) Skilled manual (5) Semi-skilled (6) Unskilled (-1) No answer
READ7	Southgate reading test score at age 7 Range 0 - 30 (-1) No answer

DRAW7
Draw-a-man score at age 7
Range 0 - 47
(-1) No answer
(-2) Not educational

PARINT7
Parental interest in child's education at age 7
(1) Both over concerned
(2) Both very interested
(3) Both little interested
(4) 1 over concerned 1 very interested
(5) 1 over concerned 1 not very interested
(6) 1 very interested 1 not very interested
(7) Either little interested
(8) Other combinations
(-1) No answer

HT11
Child's height at age 11 in cm
Range 112 - 170
(-1) No answer

WT11
Child's weight at age 11 in kgs
Range 22 7 - 77 1
(-1) No answer

READ11
Reading comprehension test score at age 11
Range 0 - 34
(-1) No answer

MATH11
Mathematics test score at age 11
Range 0 - 40
(-1) No answer

TENURE11
Housing tenure at age 11
(1) Owner occupier
(2) Council tenant
(3) Private rented - unfurnished
(4) Private rented - furnished
(5) Tied to occupation
(6) Other
(-1) No answer

NKIDS11
Number of children under 21 in the family, at age 11
Actual number (Range 1 - 8)
(9) Nine or more children
(-1) No answer

CROWD16
Number of persons per room at age 16
(1) Up to 1
(2) Over 1 and up to 1.5
(3) Over 1.5 and up to 2
(4) Over 2
(-1) No answer

READ16 Reading comprehension test score at age 16
Range 3 - 35
(-1) No answer

READT16 Transformed reading comprehension test score at age 16
Range -2 57 - +2 49
(-99) No answer

MATH16 Mathematics test score at age 16
Range 0 - 31
(-1) No answer

ATTEN16 Child's school attendance at age 16
Percentage of attendance in Autumn term 1973
(-1) No answer
(-2) No answer

LIKES16 I do not like school - at age 16
(1) Very true
(2) Partly true
(3) Cannot say
(4) Partly untrue
(5) Not true at all
(-1) No answer

SCHLT16 School type at age 16
(1) Comprehensive
(2) Grammar
(3) Secondary modern
(4) Independent and direct grant
(5) Other
(-1) No answer

HT23 Height at age 23 in cm
Range 137 16 - 226 06
(-1) No answer

LEFTED Age completed FT education
Age in months the respondent left full-time continuous education
Range 182 - 307
(-1) No answer

HIGHQUAL Highest educational qualification at age 23
(1) Higher degree
(2) Degree
(3) Teacher
(4) Higher technical
(5) Nurse
(6) At least 2 A levels
(7) 1 A level or ONC, TEC
(8) 5 O levels or crafts
(9) O levels and some other formal qualification (e g. RSA)
(10) O levels only
(11) No O level but some other formal qualification (e.g. RSA)
(12) Apprenticeship
(13) Foreign
(14) Other
(15) None
(-1) No answer

WORKHRS Current job hours of work at age 23
Range 6 - 99
(-2) Varies
(-1) No answer

NEARNPW Net earnings per week from main job at age 23 in pounds
Range 4 00 to 400.00
(-1) No answer

GEARNPW Gross earnings per week from main job at age 23 in pounds
Range 4.00 - 500.00
(-1) No answer

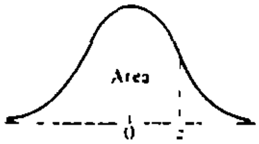
CURRSOC Social class of current or last job at age 23
(1) Professional
(2) Intermediate
(3) Skilled non-manual
(4) Skilled manual
(5) Semi-skilled non-manual
(6) Semi-skilled manual
(7) Unskilled
(-1) No answer

UNEMTIME Total months ever unemployed up to age 23
Range 0 - 86
(-1) No answer

VOTED	Party voted for in the 1979 general election	
	(1)	Conservative
	(2)	Labour
	(3)	Liberal
	(4)	Social Democrat
	(5)	Welsh National
	(6)	Scots National
	(7)	National Front
	(8)	Communist
	(9)	SWP
	(10)	WRP
	(11)	Other
	(97)	Refused
	(98)	Don't know (Don't remember)
	(99)	Did not vote
	(-1)	No answer

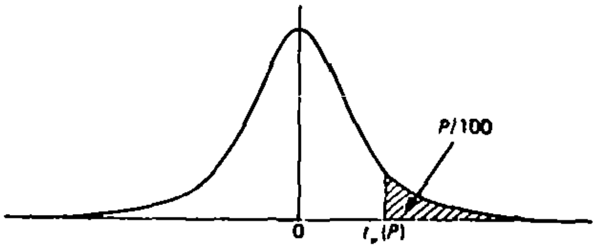
APPENDIX B
STATISTICAL TABLES

Table B.1: Areas under the Standard Normal Curve



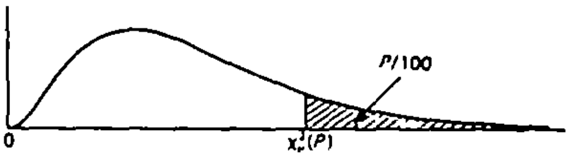
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0017	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0352	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0722	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table B.2: Areas under the t-Distribution



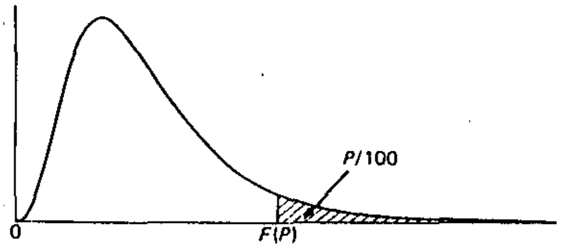
$\bar{P}$	40	30	25	20	15	10	5	2 5	1	0 5	0 1	0 05
ν = 1	0 3249	0 7265	1 0000	1 3764	1 963	3 078	6 314	12 71	31 82	63 66	318 3	636 6
2	2887	6172	0 8165	1 0607	386	1 886	2 920	4 303	6 965	9 925	22 33	31 60
3	2767	5844	7649	0 9785	250	638	353	3 182	4 541	5 841	10 21	12 92
4	2707	5686	7407	9410	190	533	132	2 776	3 747	4 604	7 173	8 610
5	0 2672	0 5594	0 7267	0 9195	1 156	1 476	2 015	2 571	3 365	4 032	5 893	6 869
6	2648	5534	7176	9057	134	440	1 943	447	3 143	3 707	5 208	5 959
7	2632	5491	7111	8960	119	415	895	365	2 998	499	4 785	5 408
8	2619	5459	7064	8889	108	397	860	306	896	355	501	5 041
9	2610	5435	7027	8834	100	383	833	262	821	250	297	4 781
10	0 2602	0 5415	0 6998	0 8791	1 093	1 372	1 812	2 228	2 764	3 169	4 144	4 587
11	2596	5399	6974	8755	088	363	796	201	718	3 106	4 025	437
12	2590	5386	6955	8726	083	356	782	179	681	3 055	3 930	318
13	2586	5375	6938	8702	079	350	771	160	650	3 012	852	221
14	2582	5366	6924	8681	076	345	761	145	624	2 977	787	140
15	0 2579	0 5357	0 6912	0 8662	1 074	1 341	1 753	2 131	2 602	2 947	3 733	4 073
16	2576	5350	6901	8647	071	337	746	120	583	921	686	4 015
17	2573	5344	6892	8633	069	333	740	110	567	898	646	3 965
18	2571	5338	6884	8620	067	330	734	101	552	878	610	922
19	2569	5333	6876	8610	066	328	729	093	539	861	579	883
20	0 2567	0 5329	0 6870	0 8600	1 064	1 325	1 725	2 086	2 528	2 845	3 552	3 850
21	2566	5325	6864	8591	063	323	721	080	518	831	527	819
22	2564	5321	6858	8583	061	321	717	074	508	819	505	792
23	2563	5317	6853	8575	060	319	714	069	500	807	485	768
24	2562	5314	6848	8569	059	318	711	064	492	797	467	745
25	0 2561	0 5312	0 6844	0 8562	1 058	1 316	1 708	2 060	2 485	2 787	3 450	3 725
26	2560	5309	6840	8557	058	315	706	056	479	779	435	707
27	2559	5306	6837	8551	057	314	703	052	473	771	421	690
28	2558	5304	6834	8546	056	313	701	048	467	763	408	674
29	2557	5302	6830	8542	055	311	699	045	462	756	396	659
30	0 2556	0 5300	0 6828	0 8538	1 055	1 310	1 697	2 042	2 457	2 750	3 385	3 646
32	2555	5297	6822	8530	054	309	694	037	449	738	365	622
34	2553	5294	6818	8523	052	307	691	032	441	728	348	601
36	2552	5291	6814	8517	052	306	688	028	434	719	333	582
38	2551	5288	6810	8512	051	304	686	024	429	712	319	566
40	0 2550	0 5286	0 6807	0 8507	1 050	1 303	1 684	2 021	2 423	2 704	3 307	3 551
50	2547	5278	6794	8489	047	299	676	2 009	403	678	261	496
60	2545	5272	6786	8477	045	296	671	2 000	390	660	232	460
120	2539	5258	6765	8446	041	289	658	1 980	358	617	160	373
∞	0 2533	0 5244	0 6745	0 8416	1 036	1 282	1 645	1 960	2 326	2 576	3 090	3 291

Table B 7: Areas under the X^2 - Distribution



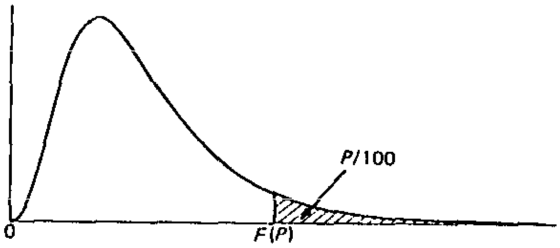
P	50	40	30	20	10	5	2 5	1	0 5	0 1	0 05
ν = 1	0 4549	0 7083	1 074	1 642	2 706	3 841	5 024	6 635	7 879	10 83	12 12
2	1 386	1 833	2 408	3 219	4 605	5 991	7 378	9 210	10 60	13 82	15 20
3	2 366	2 946	3 665	4 642	6 251	7 815	9 348	11 34	12 84	16 27	17 73
4	3 357	4 045	4 878	5 989	7 779	9 488	11 14	13 28	14 86	18 47	20 00
5	4 351	5 132	6 064	7 289	9 236	11 07	12 83	15 09	16 75	20 52	22 11
6	5 348	6 211	7 231	8 558	10 64	12 59	14 45	16 81	18 55	22 46	24 10
7	6 346	7 283	8 383	9 803	12 02	14 07	16 01	18 48	20 28	24 32	26 02
8	7 344	8 351	9 524	11 03	13 36	15 51	17 53	20 09	21 95	26 12	27 87
9	8 343	9 414	10 66	12 24	14 68	16 92	19 02	21 67	23 59	27 88	29 67
10	9 342	10 47	11 78	13 44	15 99	18 31	20 48	23 21	25 19	29 59	31 42
11	10 34	11 53	12 90	14 63	17 28	19 68	21 92	24 72	26 76	31 26	33 14
12	11 34	12 58	14 01	15 81	18 55	21 03	23 34	26 22	28 30	32 91	34 82
13	12 34	13 64	15 12	16 98	19 81	22 36	24 74	27 69	29 82	34 53	36 48
14	13 34	14 69	16 22	18 15	21 06	23 68	26 12	29 14	31 32	36 12	38 11
15	14 34	15 73	17 32	19 31	22 31	25 00	27 49	30 58	32 80	37 70	39 72
16	15 34	16 78	18 42	20 47	23 54	26 30	28 85	32 00	34 27	39 25	41 31
17	16 34	17 82	19 51	21 61	24 77	27 59	30 19	33 41	35 72	40 79	42 88
18	17 34	18 87	20 60	22 76	25 99	28 87	31 53	34 81	37 16	42 31	44 43
19	18 34	19 91	21 69	23 90	27 20	30 14	32 85	36 19	38 58	43 82	45 97
20	19 34	20 95	22 77	25 04	28 41	31 41	34 17	37 57	40 00	45 31	47 50
21	20 34	21 99	23 86	26 17	29 62	32 67	35 48	38 93	41 40	46 80	49 01
22	21 34	23 03	24 94	27 30	30 81	33 92	36 78	40 29	42 80	48 27	50 51
23	22 34	24 07	26 02	28 43	32 01	35 17	38 08	41 64	44 18	49 73	52 00
24	23 34	25 11	27 10	29 55	33 20	36 42	39 36	42 98	45 56	51 18	53 48
25	24 34	26 14	28 17	30 68	34 38	37 65	40 65	44 31	46 93	52 62	54 95
26	25 34	27 18	29 25	31 79	35 56	38 89	41 92	45 64	48 29	54 05	56 41
27	26 34	28 21	30 32	32 91	36 74	40 11	43 19	46 96	49 64	55 48	57 86
28	27 34	29 25	31 39	34 03	37 92	41 34	44 46	48 28	50 99	56 89	59 30
29	28 34	30 28	32 46	35 14	39 09	42 56	45 72	49 59	52 34	58 30	60 73
30	29 34	31 32	33 53	36 25	40 26	43 77	46 98	50 89	53 67	59 70	62 16
32	31 34	33 38	35 66	38 47	42 58	46 19	49 48	53 49	56 33	62 49	65 00
34	33 34	35 44	37 80	40 68	44 90	48 60	31 97	56 06	58 96	65 25	67 80
36	35 34	37 50	39 92	42 88	47 21	51 00	54 44	58 62	61 58	67 99	70 59
38	37 34	39 56	42 05	45 08	49 51	53 38	56 90	61 16	64 18	70 70	73 35
40	39 34	41 62	44 16	47 27	51 81	55 76	59 34	63 69	66 77	73 40	76 09
50	49 33	51 89	54 72	58 16	63 17	67 50	71 42	76 15	79 49	86 66	89 56
60	59 33	62 13	65 23	68 97	74 40	79 08	83 30	88 38	91 95	99 61	102 7
70	69 33	72 36	75 69	79 71	85 53	90 53	95 02	100 4	104 2	112 3	115 6
80	79 33	82 57	86 12	90 41	96 58	101 9	106 6	112 3	116 3	124 8	128 3
90	89 33	92 76	96 52	101 1	107 6	113 1	118 1	124 1	128 3	137 2	140 8
100	99 33	102 9	106 9	111 7	118 5	124 3	129 6	135 8	140 2	149 4	153 2

Table B.4(1): 5 per cent points of the F-Distribution



$\nu_1 =$	1	2	3	4	5	6	7	8	10	12	24	∞
$\nu_2 =$ 1	161·4	199·5	215·7	224·6	230·2	234·0	236·8	238·9	241·9	243·9	249·1	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·35	19·37	19·40	19·41	19·45	19·50
3	10·13	9·552	9·277	9·117	9·013	8·941	8·887	8·845	8·786	8·745	8·639	8·526
4	7·709	6·944	6·591	6·388	6·256	6·163	6·094	6·041	5·964	5·912	5·774	5·628
5	6·608	5·786	5·409	5·192	5·050	4·950	4·876	4·818	4·735	4·678	4·527	4·365
6	5·987	5·143	4·757	4·534	4·387	4·284	4·207	4·147	4·060	4·000	3·841	3·669
7	5·591	4·737	4·347	4·120	3·972	3·866	3·787	3·726	3·637	3·575	3·410	3·230
8	5·318	4·459	4·066	3·838	3·687	3·581	3·500	3·438	3·347	3·284	3·115	2·928
9	5·117	4·256	3·863	3·633	3·482	3·374	3·293	3·230	3·137	3·073	2·900	2·707
10	4·965	4·103	3·708	3·478	3·326	3·217	3·135	3·072	2·978	2·913	2·737	2·538
11	4·844	3·982	3·587	3·357	3·204	3·095	3·012	2·948	2·854	2·788	2·609	2·404
12	4·747	3·885	3·490	3·259	3·106	2·996	2·913	2·849	2·753	2·687	2·505	2·296
13	4·667	3·806	3·411	3·179	3·025	2·915	2·832	2·767	2·671	2·604	2·420	2·206
14	4·600	3·739	3·344	3·112	2·958	2·848	2·764	2·699	2·602	2·534	2·349	2·131
15	4·543	3·682	3·287	3·056	2·901	2·790	2·707	2·641	2·544	2·475	2·288	2·066
16	4·494	3·634	3·239	3·007	2·852	2·741	2·657	2·591	2·494	2·425	2·235	2·010
17	4·451	3·592	3·197	2·965	2·810	2·699	2·614	2·548	2·450	2·381	2·190	1·960
18	4·414	3·555	3·160	2·928	2·773	2·661	2·577	2·510	2·412	2·342	2·150	1·917
19	4·381	3·522	3·127	2·895	2·740	2·628	2·544	2·477	2·378	2·308	2·114	1·878
20	4·351	3·493	3·098	2·866	2·711	2·599	2·514	2·447	2·348	2·278	2·082	1·843
21	4·325	3·467	3·072	2·840	2·685	2·573	2·488	2·420	2·321	2·250	2·054	1·812
22	4·301	3·443	3·049	2·817	2·661	2·549	2·464	2·397	2·297	2·226	2·028	1·783
23	4·279	3·422	3·028	2·796	2·640	2·528	2·442	2·375	2·275	2·204	2·005	1·757
24	4·260	3·403	3·009	2·776	2·621	2·508	2·423	2·355	2·255	2·183	1·984	1·733
25	4·242	3·385	2·991	2·759	2·603	2·490	2·405	2·337	2·236	2·165	1·964	1·711
26	4·225	3·369	2·975	2·743	2·587	2·474	2·388	2·321	2·220	2·148	1·946	1·691
27	4·210	3·354	2·960	2·728	2·572	2·459	2·373	2·305	2·204	2·132	1·930	1·672
28	4·196	3·340	2·947	2·714	2·558	2·445	2·359	2·291	2·190	2·118	1·915	1·654
29	4·183	3·328	2·934	2·701	2·545	2·432	2·346	2·278	2·177	2·104	1·901	1·638
30	4·171	3·316	2·922	2·690	2·534	2·421	2·334	2·266	2·165	2·092	1·887	1·622
32	4·149	3·295	2·901	2·668	2·512	2·399	2·313	2·244	2·142	2·070	1·864	1·594
34	4·130	3·276	2·883	2·650	2·494	2·380	2·294	2·225	2·123	2·050	1·843	1·569
36	4·113	3·259	2·866	2·634	2·477	2·364	2·277	2·209	2·106	2·033	1·824	1·547
38	4·098	3·245	2·852	2·619	2·463	2·349	2·262	2·194	2·091	2·017	1·808	1·527
40	4·085	3·232	2·839	2·606	2·449	2·336	2·249	2·180	2·077	2·003	1·793	1·509
60	4·001	3·150	2·758	2·525	2·368	2·254	2·167	2·097	1·993	1·917	1·700	1·389
120	3·920	3·072	2·680	2·447	2·290	2·175	2·087	2·016	1·910	1·834	1·608	1·254
∞	3·841	2·996	2·605	2·372	2·214	2·099	2·010	1·938	1·831	1·752	1·517	1·000

Table B.4(2): 1 per cent points of the F-Distribution



$\nu_1 =$	1	2	3	4	5	6	7	8	10	12	24	∞
$\nu_2 =$ 1	4052	4999	5403	5625	5764	5859	5928	5981	6056	6106	6235	6366
2	98·50	99·00	99·17	99·25	99·30	99·33	99·36	99·37	99·40	99·42	99·46	99·50
3	34·12	30·82	29·46	28·71	28·24	27·91	27·67	27·49	27·23	27·05	26·60	26·13
4	21·20	18·00	16·69	15·98	15·52	15·21	14·98	14·80	14·55	14·37	13·93	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·46	10·29	10·05	9·888	9·466	9·020
6	13·75	10·92	9·780	9·148	8·746	8·466	8·260	8·102	7·874	7·718	7·313	6·880
7	12·25	9·547	8·451	7·847	7·460	7·191	6·993	6·840	6·620	6·469	6·074	5·650
8	11·26	8·649	7·591	7·006	6·632	6·371	6·178	6·029	5·814	5·667	5·279	4·859
9	10·56	8·022	6·992	6·422	6·057	5·802	5·613	5·467	5·257	5·111	4·729	4·311
10	10·04	7·559	6·552	5·994	5·636	5·386	5·200	5·057	4·849	4·706	4·327	3·909
11	9·646	7·206	6·217	5·668	5·316	5·069	4·886	4·744	4·539	4·397	4·021	3·602
12	9·330	6·927	5·953	5·412	5·064	4·821	4·640	4·499	4·296	4·155	3·780	3·361
13	9·074	6·701	5·739	5·205	4·862	4·620	4·441	4·302	4·100	3·960	3·587	3·165
14	8·862	6·515	5·564	5·035	4·695	4·456	4·278	4·140	3·939	3·800	3·427	3·004
15	8·683	6·359	5·417	4·893	4·556	4·318	4·142	4·004	3·805	3·666	3·294	2·868
16	8·531	6·226	5·292	4·773	4·437	4·202	4·026	3·890	3·691	3·553	3·181	2·753
17	8·400	6·112	5·185	4·669	4·336	4·102	3·927	3·791	3·593	3·455	3·084	2·653
18	8·285	6·013	5·092	4·579	4·248	4·015	3·841	3·705	3·508	3·371	2·999	2·566
19	8·185	5·926	5·010	4·500	4·171	3·939	3·765	3·631	3·434	3·297	2·925	2·489
20	8·096	5·849	4·938	4·431	4·103	3·871	3·699	3·564	3·368	3·231	2·859	2·421
21	8·017	5·780	4·874	4·369	4·042	3·812	3·640	3·506	3·310	3·173	2·801	2·360
22	7·945	5·719	4·817	4·313	3·988	3·758	3·587	3·453	3·258	3·121	2·749	2·305
23	7·881	5·664	4·765	4·264	3·939	3·710	3·539	3·406	3·211	3·074	2·702	2·256
24	7·823	5·614	4·718	4·218	3·895	3·667	3·496	3·363	3·168	3·032	2·659	2·211
25	7·770	5·568	4·675	4·177	3·855	3·627	3·457	3·324	3·129	2·993	2·620	2·169
26	7·721	5·526	4·637	4·140	3·818	3·591	3·421	3·288	3·094	2·958	2·585	2·131
27	7·677	5·488	4·601	4·106	3·785	3·558	3·388	3·256	3·062	2·926	2·552	2·097
28	7·636	5·453	4·568	4·074	3·754	3·528	3·358	3·226	3·032	2·896	2·522	2·064
29	7·598	5·420	4·538	4·045	3·725	3·499	3·330	3·198	3·005	2·868	2·495	2·034
30	7·562	5·390	4·510	4·018	3·699	3·473	3·304	3·173	2·979	2·843	2·469	2·006
32	7·499	5·336	4·459	3·969	3·652	3·427	3·258	3·127	2·934	2·798	2·423	1·956
34	7·444	5·289	4·416	3·927	3·611	3·386	3·218	3·087	2·894	2·758	2·383	1·911
36	7·396	5·248	4·377	3·890	3·574	3·351	3·183	3·052	2·859	2·723	2·347	1·872
38	7·353	5·211	4·343	3·858	3·542	3·319	3·152	3·021	2·828	2·692	2·316	1·837
40	7·314	5·179	4·313	3·828	3·514	3·291	3·124	2·993	2·801	2·665	2·288	1·805
60	7·077	4·977	4·126	3·649	3·339	3·119	2·953	2·823	2·632	2·496	2·115	1·601
120	6·851	4·787	3·949	3·480	3·174	2·956	2·792	2·663	2·472	2·336	1·950	1·381
∞	6·635	4·605	3·782	3·319	3·017	2·802	2·639	2·511	2·321	2·185	1·791	1·000

REFERENCES

The reference list includes 3 sections.

Section A - NCDS. Out of the large number of publications related to the NCDS this list includes only a small number of general publications.

Section B - Introductory Statistics. The list includes several books dealing with introductory statistics.

Section C - SPSS-X.

A. NCDS

Davie R. Butler N. and Goldstein H. (1972). From Birth to Seven: the second Report of the NCDS. London, Longman.

Fogelman K. (ed) (1983). Growing up in Great Britain. London, Macmillan.

NCDS4 Research Team (1987). The Fourth Follow-Up of the National Child Development Study: an account of the methodology and summary of the early findings. London City University, (Social Statistics Research Unit, Working Paper No. 20)

Shepherd P. (1988). National Child Development Study - Teaching Data Sets. Introduction to the Background to the Study and the Methods of Collection. ESRC Data Archive.

B. INTRODUCTORY STATISTICS

Armore S.J. (1966). Introduction to Statistical Analysis and Inference for Psychology and Education. New York, Wiley & Sons.

Blalock H.M. (1972). Social Statistics. New York, McGraw Hill.

Chatfield C. (1983). Statistics for Technology. London, Chapman and Hall.

Hoel P.J. and Jessen R.J. (1982). Basic Statistics for Business and Economics. New York, Wiley & Sons.

Thomas J.J. (1973). An Introduction to Statistical Analysis for Economist. London, Weidenfeld and Nicolson.

Walpole R.E. (1982). Introduction to Statistics. New York, Macmillan.

Wonnacott T.H. and Wonnacott R.J. (1984). Introductory Statistics for Business and Economics. New York, Wiley and Sons.

C. SPSS-X

Norusis M.J. (1983). SPSS-X. Introductory Statistics Guide. New York, McGraw-Hill.

SPSS-Inc. (1986). SPSS-X User's Guide. New York, McGraw-Hill.

NATIONAL CHILD DEVELOPMENT STUDY

SOCIOLOGY TEACHING SET

CLASS AND GENDER STRATIFICATION IN YOUTH

Data sets and documentation for analysing data from

The National Child Development Study

Containing:

2000 Members of the NCDS 1958 Cohort
32 Variables
SPSS-X System file

Prepared by

Gill Jones

Thomas Coram Research Unit
University of London Institute of Education
41 Brunswick Square
London WC1N 1AZ

In conjunction with:

The NCDS User Support Group
Social Statistics Research Unit
City University
Northampton Square
London EC1V 0HB

1987

Class and Gender Stratification in Youth

CONTENTS **page**

Note:
The material contained in this data set should be used for teaching purposes only. It should not be used for research purposes or as the basis for publications.

Acknowledgements:
Much of the work on this teaching data set was undertaken while the author was a postgraduate sociology student at the University of Surrey. This NCDS teaching data set has been prepared on a similar model to that of the GHS Teaching Data Set, Exploring British Society, prepared by Sara Arber, Angela Dale, Nigel Gilbert and Lyn Rajan, Stratification and Employment Group, Department of Sociology, University of Surrey. I am particularly indebted to Sara Arber and Nigel Gilbert for their invaluable comments and advice.

1 INTRODUCTION TO THE TEACHING DATA SET

The NCDS Sociology Teaching Data Set has been developed in order to give undergraduate sociology students a practical introduction to sociology. The teaching set has been organised around the theme of stratification in society, and has been designed in order to allow students to explore for themselves some of the elements of structural inequality.

The Teaching Data Set is a subset of the full National Child Development Study, and contains variables from the original perinatal study, and later sweeps when the cohort was aged 7, 11, 16 and 23 years (see General Introduction for more details).

Why use a teaching data set?

a) To teach students about stratification in society. The teaching data set can be used as an adjunct to theoretical teaching. It shows that apparently abstract concepts, such as stratification and inequality, are empirically researchable.

b) To train students to look at methodology and data as potential contributors to sociological theory. The NCDS Teaching data set in particular shows how access to longitudinal data allows longitudinal analysis of class and the development of the concept of class trajectories.

c) To provide practical training in quantitative social research. Through the NCDS Teaching data set, students can learn basic computing skills, and some simple data analytic techniques.

d) The NCDS Teaching data set introduces students to the idea of research through the secondary analysis of existing data sets. This is a growing area of social research. Specifically, students are introduced to the NCDS as an available source of longitudinal data.

How to use it

The teaching data set assumes no prior knowledge of computing or quantitative research methods on the part of the student. The exercises are derived with the undergraduate sociology student in mind. The data sets are supplied in the form of SPSS-X system files, and the exercises have been written for use with SPSS-X, but could be adapted to other statistical packages, such as P-STAT or SAS. These exercises are only intended to be illustrative; students may wish to undertake different analyses or use different techniques.

The Codebook supplied with the exercises introduces the data set to the student. Each variable is briefly described and frequencies given. Also included are some explanatory notes on using SPSS-X.

Guide to sections

The exercises have been divided into seven sections, each dealing with a different aspect of social inequality, as follows:

Occupational Class
Gender
Education
Inter-generational Mobility
Intra-generational Mobility
Housing
Class Identification

On each topic exercises are proposed which begin to address some of the theoretical and methodological problems. The aim is to encourage students to question some of the standard ways in which class and inequality are conceived, and to re-assess the use of basic indicators of inequality.

Occupational class is examined first. The difficulties of using this measure as an indicator of inequality among young people are shown. Gender inequalities both between men and women and among women are addressed in the next section, and some examples of inequalities and the reasons they exist are indicated through the exercises. The section on education shows differential access to educational opportunity and the extent to which inequalities relating to gender and class of origin can be overcome through educational achievement.

In the next two sections inter-generational and intra-generational class mobility are explored, and class transmission is revealed as a complicated process. Class careers of families over three generations, and of individuals over sixteen years are examined in consecutive sections.

Finally, the exercises end with an examination of some of the social outcomes of stratification. First, housing tenure is examined, both in terms of housing trends between 1958 and 1974, and in terms of the housing "careers" of the cohort and their parents. Then, class identification is briefly addressed through analysis of the voting behaviour of the cohort and their membership of trade unions.

The importance of longitudinal data is shown through the exercises concerned with social mobility and with housing trends. The exercises suggest that cross-sectional data may only provide a partial picture and the processes underlying current behaviour should still be recognised. The social scientist who lacks access to longitudinal data can thus still take account of the social processes which are continually taking place and which affect the cross-sectional picture of society which is more generally studied.

2 STRATIFICATION IN SOCIETY

The concept of social stratification is a central theme in sociology. It is used to describe the structure of inequality, and to define the position of the individual or group within this structure. In industrial societies, inequalities of access to wealth and power may be associated with a number factors: social class, gender and race are three. In non-industrial societies, power and prestige may be more associated with age. In general, it might be said that society is stratified along different dimensions, which inter-relate. There are problems when we attempt to indentify the dimensions of stratification, and when we try to determine the relationship between them.

It is impossible to provide more than a brief overview of some of the issues in the few pages available. We shall consider only some of the issues concerning the concept and measurement of social class, and gender inequalities.

Most studies of social inequality are based on social class. The measurement of social class is a preoccupation of many sociologists but should always be seen as leading towards a more meaningful understanding of the social structure rather than towards a rigid categorisation. The meaning of social class in its widest context should constantly be informing the analysis of class in the narrower confines of the occupational structure.

In the United States, high levels of social mobility led liberal sociologists like Blau and Duncan (1967) to conclude that social stratification was far from rigid, as Marx had suggested, and that the mobility in American society would eventually lead to universalism, or social equality. Social class is thus treated as a question of achievement rather than ascription. Marxist and Weberian explanations of stratification have generally been rejected in favour of a schema which produces a hierarchical ordering of socio-economic groups along a continuum, with no implication of a class struggle, or even division between manual and non-manual workers ("non-egalitarian classnessness", according to Ossowski, 1963).

Most European analyses of stratification tend to emphasise the rigidity rather than fluidity of the class structure. Post Marxist analyses of class stress the relationship between labour and capital, but many also emphasise the juxtaposition of manual and non-manual occupational classes. Some Marxian class analysis therefore sees manual and non-manual workers in opposition to one another, rather than as part of the same hierarchy. E.O. Wright (1976) has extended the definition of the capitalist class to include managers, who have power over the working class though they do not own capital, while manual and other non-manual workers form the proletariat. He has been criticised for paying undue attention to the class affiliation of intermediate non-manual workers, by Goldthorpe (1980) who points out that this is a highly mobile and fluctuating group.

Most British approaches to class analysis produce a roughly hierarchical ordering of social class according to socio-economic criteria, on the basis (following Weber) that the categories reflect life chances. There is some agreement among sociologists about the relatively stable upper and lower ends of the occupational hierarchy (professions and senior managers at the top, and unskilled manual workers at the bottom), but disagreement over the treatment of the more mobile intermediate classes. Hall and Jones (1950) perceive junior non-manual workers to be in a lower position than foremen but on the same level as skilled manual workers. Goldthorpe (1980) on the other hand, defines four distinct intermediate classes (clerical workers, small shopkeepers, foremen and skilled manual workers) movement between which would not constitute upward mobility. Broadly, as Heath (1981) points out, class schemata vary according to the use to which they are to be applied, as well as to theoretical considerations.

In the relative national prosperity on the 1950s and 1960s, the improved situation of the working class led many people to believe that class divisions were decreasing, as the working class became members of the bourgeoisie. The occupational structure was changing, and more people were entering white collar work; the 1948 Education Act led to greater educational opportunities for many. The traditional working class appeared to be in decline. The Oxford Mobility Study of the "Affluent Worker" tested the embourgeoisement theory, and found little evidence that the manual workers in their study shared the values and attitudes of the middle class (Goldthorpe et al., 1967).

Social Mobility

The study of social mobility attempts to determine the degree of openness of the class structure. If the class structure is open, then rates of social mobility will be high, while if the class structure is relatively closed, classes will be more stable. There are two main elements to social mobility, both of which must be examined if conclusions about the class structure, and prospects for change, are to be drawn: intergenerational social mobility, the extent to which social class is transmitted from one generation to another, and intragenerational social mobility, which describes class careers during an individual's life course.

What are the routes to upward social mobility? Heath (1981) and Goldthorpe (1980) have suggested three main routes to the higher classes in society: inheritance of privilege, education, and promotion from the shop floor. Direct inheritance of privilege occurs rarely in the present day, and class of origin is more likely to affect social class destination in indirect ways, such as through education, health and housing. The educational route is perhaps the chief means through which the working class may gain upward mobility. Achievement through work careers can help those of both classes of origin to gain social mobility intra-generationally. Access to the various routes to upward mobility varies, though.

Women and Class Theory

It is only in recent years that any serious attempt has been made to incorporate women into class theory directly. Hitherto, social class was only seen as ascribed to a woman through her father's or husband's occupational class, and the woman's own occupational class was not considered to be of sociological interest. Women have thus been excluded from most studies of social mobility and social class on the grounds that they are economically dependent on their husbands even if they work (there has been little consideration of the possibility that husbands may be equally dependent upon them) and that gender differences tend to underline class differences rather than cut across them (Goldthorpe, 1980; Westergaard and Resler, 1975).

Erik Olin Wright's (1978) position was not dissimilar, though his comments are restricted to housewives and do not include women in the paid labour force. He considered that housewives held a "contradictory class location" since their class interests were essentially those of their husband, the worker and "that the sexual division of labour does not create a division of fundamental class interests between husbands and their housewives".

The suggestion that the class fates of women are determined by their menfolk has only recently been more fully debated. There has been criticism of intellectual sexism among sociologists (see Oakley, 1974; Delphy, 1981), which is likely to influence both orientation to research and findings. In contrast, others have shown that a woman's own occupational class crucially affects fertility and voting behaviour (Heath and Britten, 1984). Studies of the domestic division of labour have highlighted the inequalities within the home (J. Pahl,1983).

Whether gender inequalities cross-cut social class inequalities or occur within the class structure is an issue widely debated in the present day. Much of the debate concerns the division of labour in the home, and the recognition of housework as unpaid work. If housewives are not to be ascribed a social class according to their husband's occupational class, then it becomes necessary, among women who are not in paid work, to assign them to an occupational class on the basis of their work in the home.

Our concern here, however, is with the position of women in the labour market, since most women in the NCDS are in paid employment at 23 years rather than full-time housewives. It is not the division of labour in the home which we shall be examining, though expectations of marriage and having children may lead women to have lower expectations of work careers than men.

Suggested General Reading

Beechey, V. and Whitelegg, E. (eds) (1986) Women in Britain Today, Open University Press.

Dex, S. (1985) The Sexual Division of Work, Wheatsheaf Books.

Giddens, A. and MacKenzie, G. (eds) (1982) Social Class and The Division of Labour, Cambridge University Press.

Haralambos, M. (ed.) (1985) Sociology: New Directions Causeway Press.

O'Donnell, M. (1983) New Introductory Reader in Sociology, Harrap.

Worsley, P. (ed.)(1987) The New Introducing Sociology, Penguin Books

3 OCCUPATIONAL CLASS

There have been a number of schema of occupational class. The first originated in 1911 with the Registrar General's Classification which attempts to group people according to their levels of occupational skill. The most common current classification is by OPCS socio-economic groups, which aims to bring together people with jobs of similar social and economic status (OPCS, 1980).

Occupational class in the NCDS 1981 Sweep is based on the OPCS socio-economic group of the cohort's current (1981) or last job. Those who have never been in paid employment therefore have no socio-economic group or occupational class assigned to them. This is an important point. In fact, among the NCDS cohort, few have never held a job, since unemployment among school leavers was relatively low in the mid-seventies when they left education, in contrast to the situation currently. The study of a cohort of school leavers in the mid 1980s, would have to consider alternative indicators of social class, including the possibility of assigning all those who have never been employed to a separate class or consider them as marginal to the class structure of society. This is not necessary for the current NCDS data.

There are considerable problems with the relevance of these occupational class schema for women. They were developed for men and reflect the male occupational structure. Women tend to be concentrated in junior non-manual and semi-skilled manual work, and under-represented in some other groups.

Table 3.1 Occupational Class Schema

Occupational Class	Socio-Economic Group	Examples of Occupations
1 Higher Professional	3, 4	Accountants,lawyers, medical practitioners
2 Intermediate Non-manual & Lower Professionals	1, 2, 5*, 13	Managers,self-employed, teachers, nurses
3 Junior Non-Manual	5*, 6*	Office supervisors, typists, clerks
4 Skilled Manual	8,9 12,14	Foremen, drivers, craftsmen, skilled production workers
5 Semi-Skilled workers, Manual packers	6*, 7, 10, 15	Personal service shop assistants,
6 Unskilled Manual	11	Labourers, cleaners

* Shop assistants have been grouped with semi-skilled manual workers, and lower professionals with intermediate non-manual workers in order to create a class schema which is more meaningful for the study of women and young people.

Research Questions

How well can young workers be categorised according to their occupational class? While occupational class may provide a good indicator of the position of older adults in the social structure, for example in terms of their life chances, their social status, is this the case for young people? (Jones, 1986b)

One way to begin to examine this question is to look at the extent of heterogeneity within occupational classes. Who occupies the different occupational classes? If occupational class is the most important indicator of life chances for the NCDS cohort, then classes might be expected to be fairly homogeneous in other respects. The greater the heterogeneity of class composition, the more likely it is that other factors are present which may influence life chances.

The following exercise examines this. The analysis will show whether the young working class and middle class are homogeneous groupings. Analysis will be by class of origin, current occupational class and gender and will show the extent of heterogeneity in each group. To what extent is occupational class likely to be a reliable measure of social class in youth? Might other indicators such as class of origin and gender also be needed for an analysis of inequality?

Variables Needed:

SEX	Sex of Respondent
CLASS	Occupational Class at 23 years
PACLASS	Father's Occupational Class 1974 or 1969

Suggested Analysis:

Exercise 3.1: Occupational Class distributions by gender

Are classes equally occupied by men and women, or are some classes dominated by one sex? The following crosstabulation of occupational class by sex will show the extent of gender segregation in the occupational structure.

```
     CROSSTABS CLASS BY SEX
     OPTIONS 4
```

Can you see and account for any gender differences in class composition? They are explored more fully in later exercises.

Exercise 3.2: Heterogeneity of Class of Origin

To what extent does current occupational class at 23 years reflect class of origin? Does inter-generational stability occur equally within the manual and non-manual classes? Is there homogeneity of class of origin within current occupational classes?

(Note: In two-way tables, we can examine the way in which the dependent variable varies according to the categories of the independent variable. In three-way tables, where a control variable is used, we examine variation in the dependent variable according to the categories of the independent variable, by eliminating the effects of a third variable which may be influencing the relationship with the first two).

Crosstabulate class by father's class by sex. Then do a further analysis, crosstabulating father's class by respondent's occupational class by sex. The first tables will show class destinations and the second tables will show class origins. Since the occupational class structure varies by age and gender, some variation between father's class and son's or daughter's class is to be expected. Obtain standardised residuals from a model of no association for each table. This procedure is a means of controlling for the different marginal distributions (for respondent's class and father's class). Residuals of more than 2 or less than -2 are considered significant.

```
a)   CROSSTABS CLASS BY PACLASS BY SEX
     OPTIONS 4 16
```

These tables show class destinations. Do people have equal chances of entering each occupational class, regardless of their class of origin?

```
b)   CROSSTABS PACLASS BY CLASS BY SEX
     OPTIONS 4 16
```

These table show class origins of current occupants of occupational classes. Does class membership occur among people of the same class of origin or is there no association between current occupational class and class of origin?

Which classes are the most homogeneous and which the most heterogeneous in terms of class of origin?

Exercise 3.3

Does upward mobility equal downward mobility in the above tables? Measure upward or downward mobility as movement by one class or more from class of origin, in the class destination tables. The best way to do this is to draw the leading diagonal (those in the same class as their fathers), sum the cell counts and work out the percentage of the total table count. Then sum the cell counts for the triangle top right of the diagonal (these are the upwardly mobile) and work out what percentage they are of the total table count. Finally do the same for the lower left triangle (the downwardly mobile). Then you can compare upward and downward mobility across the tables. Does upward mobility equal downward mobility? If not, can you account for this?

These introductory exercises have shown that occupational class is only one dimension of stratification in society. In the following exercises, we shall examine other class indicators and look more closely at social mobility.

4. GENDER

Gender inequality is a major dimension of stratification in society. Gender inequalities in the labour market can be seen in the gender segregation of occupations, and the lower prestige and pay of many occupations typically held by women. Women's position in the labour market is, however, crucially affected by the sexual division of labour in the home. Woman are seen as responsible for child care and housework, and these gender-related domestic roles affect their labour market participation. Women with children, particularly young ones, are thus more likely to be out of the paid labour force, or in part-time work.

Research Questions

Does gender inequality occur within class or do class divisions occur within the gender structure of society? What is the extent of inequality between women? Is there more inequality between working class women and middle class women, for example, than there is between men and women within the middle class or within the working class?

Variables Needed:

SEX	Sex of respondent
CLASS	Occupational Class at 23 years
PACLASS	Father's Occupational Class 1974 or 1969
EARNINGS	Respondent's Hourly Income
EMPSTAT	Employment Status
WORKHRS	Hours of Work
MARITAL	Marital Status
CHILDREN	Number of Children
EDLAGE	Age in years left full-time continuous education

Suggested Analysis:

Exercise 4.1:

Exercise 3.1 has already shown that there are gender differences in the occupational class structure, suggesting that women have less access than men to jobs in the higher occupational classes. Perhaps this is because men obtain more educational qualifications than women. Is there equality of occupational opportunity among men and women with the same educational levels? Or does gender inequality over-ride educational ability? The following analysis will show whether, when educational level is controlled for, men or women are more likely to be in the higher occupational classes (1 and 2).

By controlling for educational level, we can analyse class by gender within educational categories, and thus see whether, even among those of similar education, occupational class is still affected by gender.

a) Decide where to dichotomise EDLAGE and recode accordingly.
For example:

```
RECODE EDLAGE (15 THRU 17=1)(18 THRU HI=2)
```

and add your own value labels. For example:

```
VALUE LABELS EDLAGE 1 'under 18' 2 '18 or over'
```

b)
```
CROSSTABS CLASS BY SEX BY EDLAGE
OPTIONS 4
```

The above analysis will also show, however, that among the less educated, women are in a higher occupational class than men. While more men are in occupational Class 4 than any other occupational class, women are most often in Occupational Class 3. Does this mean that lack of education is less of a handicap for women than for men? This will be examined next.

Exercise 4.2: Relative Earnings of Men and Women

Should the occupational class structure be seen as a hierarchy, or are the intermediate classes, 3 and 4, far less clearly ordered? Exercise 2.1 showed that women are mainly in Class 3 while men are mainly in Class 4. Does this mean that women are advantaged in comparison with men?

This exercise examines the characteristics of Classes 3 and 4. Earnings can also be compared within the other occupational classes. Only those in paid employment are selected.

First recode EARNINGS into two values. Then crosstabulate by class and sex.

```
SELECT IF (EMPSTAT EQ 2)
RECODE EARNINGS (     THRU     =1)(     THRU     =2)
VALUE LABELS EARNINGS 1 '  ' 2 '  '
CROSSTABS EARNINGS BY CLASS BY SEX
OPTIONS 4
```

Is the manual/non-manual division as meaningful for women as it is for men? When incomes are compared, are male manual workers seen to earn more than female non-manual workers? Who earns most, a man in Class 4 or a woman in Class 3?
Exercise 4.3: Intra-gender inequalities

One of the main reasons why women are disadvantaged in the labour market, is because their working careers are interrupted by child-rearing. When they return to the labour market after having children, they are often in a lower occupational class than they were in before. The following exercise explores some of the reasons for this, and examines some within-gender inequalities which result from childbirth.

Compare single, childless women with married women and women with children to show the effect of marriage and childbirth on women's occupational class.

```
SELECT IF ((SEX EQ 2) AND (EMPSTAT EQ 2))
CROSSTABS CLASS BY MARITAL, CHILDREN
OPTIONS 4
```

Is it marriage or having children which most affects women's class chances?

Exercise 4.4:

Repeat the above exercises for men:

```
SELECT IF (SEX EQ 1)
```

then as before. Do marriage and children have the same effect on men's occupational status?

Exercise 4.5: The effect of gender roles in the home

Gender inequalities in class and income may result from women's restricted participation in the labour force, due to their additional roles in the home. The working hours of women, especially with young children, may therefore account for many of the gender inequalities in the labour market. To what extent are gender inequalities in the labour market the outcome of differences in working hours, and gender roles in the home. Do marriage and children affect labour market outcomes for men in the same way as they do for women?

Among those with children and currently employed, what is the association between hours in paid work and current occupational class? What are the gender differences when hours of work are controlled for?

a)

```
SELECT IF ((CHILDREN EQ 1) AND (EMPSTAT EQ 2))
RECODE WORKHRS (1 THRU 29=1)(30 THRU HI=2)
VALUE LABELS WORKHRS 1 'PART-TIME' 2 'FULL-TIME'
CROSSTABS CLASS BY SEX BY WORKHRS
OPTIONS 4
```

If current class may be affected by working hours (and ultimately by gender roles in the home) will there still be gender differences in class distribution and earnings if respondents with children are excluded from the analysis?

b)

```
SELECT IF ((CHILDREN EQ 0) AND (EMPSTAT EQ 2))
RECODE EARNINGS (    THRU  =1)(    THRU   =2)/
       WORKHRS (1 THRU 29=1)(30 THRU HI=2)/
VALUE LABELS EARNINGS 1 '  ' 2 '   '/
       WORKHRS 1 'PART-TIME' 2 'FULL-TIME'/
CROSSTABS EARNINGS BY SEX BY WORKHRS
OPTIONS 4
```

What do the results suggest? Is there gender inequality in class and earnings even between childless women and men? To what extent is gender inequality in the labour market the direct result of the division of labour, and gender roles, in the home?

5. EDUCATION

The education system provides the major possibility of enhancing a person's life chances. The association between education and production is close enough for it to have been described as a "class-allocatory device" (Bernstein, 1975). Success in the educational sphere will often be followed by success in the occupational sphere, sometimes involving upward class mobility. Conversely, lack of educational success may lead to low prestige work and downward social mobility. In general, though, because educational opportunities vary by social class, the education system is more likely to help perpetuate social inequalities, as "an integral element in the reproduction of the prevailing class structure of society" (Bowles and Gintis, 1976).

Research Questions

The exercises will examine educational achievement as one of the major means of achieving high occupational status, of class stability among the middle class, and of achieving upward inter-generational mobility from the working class.

If education is important to social class outcomes, do people have equal access to education, or can inequality of access be seen? The analysis will show whether there is equal access to educational opportunity. The exercises will show the effect of class of origin on years in education, by gender.

Variables Needed

SEX	Sex of respondent
CLASS	Occupational Class at 23 years
EDLAGE	Age in years left continuous full-time education
PACLASS	Father's Occupational Class in 1974 or 1969
MEDLAGE	Age mother left full-time education
PEDLAGE	Age father left full-time education
HIGHQUAL	Highest educational qualification at 23 years

Suggested Analysis

Exercise 5.1:

The exercise examines the association between class of origin and age of leaving full-time continuous education. Do all children have equal opportunities in education or does class of origin affect educational level?

```
RECODE PACLASS (1,2,3=1)(4,5,6=2)/
       EDLAGE (15 THRU 17=1)(18 THRU HI=2)/
VALUE LABELS PACLASS 1 'NON-MANUAL' 2 'MANUAL'/
       EDLAGE 1 'UNDER 18' 2 '18 OR OVER'/
CROSSTABS EDLAGE BY PACLASS
OPTIONS 4
```

Exercise 5.2:

Inter-generational comparisons can show the respondent's age at leaving education in relation to that of his/her parents, to show family orientations towards educational achievement. Staying on at school after the minimum school leaving age may be associated with the parents' own educational backgrounds, as well as with their occupational class.

Compare the age at which the respondent left full-time education with that of the father and mother. Is there variation within class of origin, according to the family's history of education?

First recode EDLAGE into two categories. Then examine the relationship between the respondent's educational level and that of the parent, thus:

```
CROSSTABS EDLAGE BY PEDLAGE/
          EDLAGE BY MEDLAGE/
OPTIONS 4
```

What gender differences occur? For example, does a father's education affect a son's education more than that of a daughter, while a mother's education may seem to affect that of her daughter? Do the crosstabulation again, first for sons and then for daughters, by selecting for SEX.

Exercise 5.3:

What is the association between educational achievement and occupational class? Cross-tabulate Current Occupational Class by Highest Educational Qualification controlling for gender.

First, recode HIGHQUAL into three categories 1) No qualifications, 2) Up to 5 'O' Levels, 3) Higher qualifications. Look at the original category labels in the CODEBOOK when deciding how to do this.

```
RECODE HIGHQUAL (     =1)(     =2)(     =3)/
VALUE LABELS HIGHQUAL 1 'NO QUAL' 2 '5 O LEVELS OR
          LESS'
             3 'HIGHER QUALS'/
CROSSTABS CLASS BY HIGHQUAL BY SEX
OPTIONS 4
```

Exercise 5.4:

To what extent can educational qualifications overcome the effect of class of origin on current occupational class? Do the following exercise first for men and then for women. First recode HIGHQUAL as before, then

```
SELECT IF (SEX EQ 1)
RECODE PACLASS (1,2,3=1)(4,5,6=2)/
CROSSTABS CLASS BY PACLASS BY HIGHQUAL
OPTIONS 4
```

Then repeat, substituting

```
SELECT IF (SEX EQ 2)
```

Exercise 5.5:

Does educational achievement help women to enter the higher occupational classes to the same extent as it helps men? Or are women in lower occupational classes than men even when educational level is controlled for?

First, recode HIGHQUAL as before. Then recode CLASS into three groups. This will allow simpler gender comparisons.

```
RECODE CLASS (1,2=1)(3,4=2)(5,6=3)
CROSSTABS CLASS BY SEX BY HIGHQUAL
OPTIONS 4
```

6. INTER-GENERATIONAL MOBILITY

The study of inter-generational mobility can indicate the fluidity of the class structure and its openness to new members from each generation. The study is made difficult, however, because of the problems associated with comparing, for example, the occupational class of, a son with that of his father. The point at which the measurement of class is chosen for each may reflect their different ages, different points in their work careers, and the changing occupational class structure. The question arises whether the researcher is measuring social mobility, occupational class mobility or the historical changes in the class structure. As far as the latter is concerned, the principal change has been in the decrease in manual workers, with the decline in manufacturing industry over the last twenty years, and the increase in non-manual work in the service industries. The last few decades have also seen an increase in the numbers of women returning to the labour market, often on a part-time basis, after having children. The increased participation of women in the labour market has changed the overall occupational structure. For all these reasons, the increase in the numbers of non-manual workers on a structural level should not be confused with upward inter-generational mobility within families.

This section will show the extent to which class is "inherited", generally indirectly through access to or lack of advantages in education and other class related resources.

Research Questions

The following exercises examine INTER-generational social mobility: occupational class by father's class by gender, using a revised class schema to allow gender comparisons. As was seen earlier, it is not appropriate to dichotomise occupational classes into non-manual and manual in the case of women. An expedient which allows recoding into fewer categories, and which can be applied to both men and women, is to group into higher, intermediate and lower classes.

The exercises show how the class structure is maintained, so that children tend to be in the same occupational class as their parents. The NCDS data suggests, though, that the transmission of class from parent to child is a complicated process. These exercises and the next section, on INTRA-generational mobility, show the extent to which both are often needed to ensure the stability of the middle class.

The exercises can also tell us something about the nature of the class structure. They will indicate its fluidity, by showing the extent of upward and downward mobility. In addition, though, they will also help us test hypotheses about whether or not occupational classes should be seen as a continuous scale ranging from low to high classes. Do the upwardly mobile move through the intermediate classes to the higher classes? Or do you find that manual and non-manual classes each are formed of separate hierarchies?

Variables Needed

SEX	Sex of respondent
CLASS	Occupational Class at 23 years
CLASS1	Occupational Class in first job
PACLASS	Father's Occupational Class in 1974 or 1969
MACLS58	Mother's Occupational Class in 1958
MACLS74	Mother's Occupational Class in 1974
PGFCLASS	Paternal Grandfather's Occupational Class
MGFCLASS	Maternal Grandfather's Occupational Class

Suggested Analysis

Exercise 6.1:

Do women "inherit" their father's class to the same extent as men? Are there actual gender differences in inter-generational mobility or are apparent differences the result of the difficulty of comparing women's occupational class with that of their fathers? This exercise compares the inter-generational mobility of men and women at 23 years, by cross-tabulating Occupational Class by father's class by gender. Selection of Option 16 will obtain standardised residuals (see Exercise 3.2).

```
CROSSTABS  CLASS BY PACLASS BY SEX
OPTIONS 4 16
```

A model of no association would show that all have equal chances of entering the higher occuptionsl classes. Analysis of the standardised residuals will show how well the "no association model" fits.

Exercise 6.2: Mobility across three generations

Inter-generational mobility across three generations can also be examined. The class route from grandfathers (both paternal and maternal) to parents to the respondent can be examined, though it will be necessary to take into account the changes in the occupational structure that have taken place over time. In other words, time trend effects will confuse the picture of inter-generational mobility.

How much inter-generational class stability and mobility is there across three generations? Which classes are the most stable? It will be best to restrict this analysis to men, though you could repeat the exercise for women later.

```
SELECT IF (SEX EQ 1)
RECODE CLASS PACLASS PGFCLASS (1,2,3=1)(4,5,6=2)
VALUE LABELS CLASS PACLASS PGFCLASS 1 ' NON-MANUAL'
             2 'MANUAL'/
CROSSTABS CLASS BY PACLASS BY PGFCLASS
OPTIONS 4 16
```

Since these exercises will be difficult to interpret from conventional tables, a more graphic representation of the data is suggested. "Progress charts" on the following lines can be drawn from the information contained in the cross-tabulations.

Table 6.2: Basis for Progress Chart

GRANDFATHERS CLASS %	FATHER'S CLASS %	RESPONDENT'S CLASS %	
		-	
	-		
NON-MANUAL			
NON-		-	
MANUAL			
-			
		-	
	-		MANUAL
		-	
-----	-----	-----	
100	100	100	
-----	-----	-----	

The movement from class of origin of the children of manual worker fathers can then be depicted in the same way.

The extent of social mobility between generations will be clearly seen if the findings from the following exercises are formulated in this way. The format can also be adapted for the next section on intra-generational mobility.

Exercise 6.3:

Is mother's occupational class an important factor in inter-generational class stability? Crosstabulation of occupational class by mother's class at start of pregnancy (MACLS58) and in 1974 (MACLS74) will show the difficulty of measuring occupational class for women, particularly since the nature of women's work and participation in the labour market has changed over time.

What are the difficulties about doing an analysis similar to that in Exercise 6.1 among women, comparing their own occupational class with that of their mothers? Many mothers will not have been employed at the start of their pregnancy, or in 1974, and the analysis will therefore show many missing values for mother's occupational class. One way of dealing with these is to incorporate them into the table - if you wish to do this, include OPTION 7 in your SPSS-X program.

```
SELECT IF (SEX EQ 2)
RECODE MACLS74  MACLS58 (1,2=1)(3,4=2)(5,6=3)/
VALUE LABELS MACLS74 MACLS58 1 'HIGHER'
                         2   'INTERMEDIATE'
             3 'LOWER'/
CROSSTABS CLASS by MACLS74/ CLASS BY MACLS58/
OPTIONS 4 7
```

Consider the difficulties of comparing a daughter without children, with their mother who has by definition had children. What is an analysis likely to show - try it and see.

Consider also the changing nature of women's employment over the last few decades. Is a comparison of the daughter's occupational class with that of her mother's occupational class in 1958 valid, or will historical trends make results difficult to interpret?

7. INTRA-GENERATIONAL MOBILITY

Intra-generational mobility refers to the extent to which individiduals can improve their class positions during their life time, through work careers. An examination of the structure of the class system therefore needs to see how open the structure is to this kind of class movement. Again, the changing occupational class structure over time needs to be taken into account when doing these exercises, so that intragenerational mobility is not confused with structural change.

These exercises examine the work-route to upward mobility of the working class, and the "counter-mobility" of the middle class. Counter-mobility (Goldthorpe, 1980) is the phenomenon whereby many of the less educationally successful sons and daughters of middle class parents manage to regain their class of origin through work routes, after initial downward mobility (inter-generationally) on entry into the labour market. Unless class of origin is controlled for, the counter-mobile are therefore likely to be confused with the upwardly mobile working class, and the openness of the class structure might thus be over-estimated. Counter mobility is not so much upward mobility (reflecting openness) as class reproduction (Jones, 1987b).

Measurement of the extent to which people are intra-generationally mobile gives an indication of the chances of improving one's class position through work as well as, or instead of, through education.

Research Questions

To what extent can occupational class change through work-life or intra-generational mobility? Is social mobility achievable through work routes as well as educational ones, especially for those with low educational achievements? In other words are work routes to upward mobility equally available to all those who are not educationally successful, regardless of class of origin?

Variables Needed:

SEX	Sex of respondent
CHILDREN	Number of respondent's children
YCLASS	"Youth Class" - a longitudinal class measure (see notes)
CLASS	Occupational Class at 23 years
CLASS1	Occupational Class in first job

PACLASS	Father's Occup Class in 1974 or 1969
PACLS74	Father's Occup Class in 1974
PACLS69	Father's Occup Class in 1969
PACLS65	Father's Occup Class in 1965
PACLS58	Father's Occup Class in 1958
MACLS74	Mothers occupational class in 1974
MACLS58	Mother's occupational class at pregnancy in 1958
SAMEMAPA	With own parents in 1965, 1969 and 1974
FETRAP	Further education, Training or Apprenticeship

Suggested Analysis

Exercise 7.1:

This first exercise examines the intra-generational mobility among the cohort members between their first job and their current job. Since having children is likely to affect the mobility of women, as Exercise 4.3 showed, it is advisable to restrict the analysis to respondents without children.

Simple analysis of respondent's class in first job and current class will reveal the extent to which young people are downwardly mobile on entry into the labour market, but make up ground through occupational careers.

```
SELECT IF (CHILDREN EQ 0)
CROSSTABS CLASS BY CLASS1 BY SEX
OPTIONS 4
```

Do men and women have equal rates of upward mobility through work?
Examine the gender differences in upward mobility within different classes. Examine also intra-gender differences: who are the upwardly mobile through work?

Exercise 7.2:

Is there a link between inter-generational mobility and intra-generational mobility. The second may sometimes be needed in addition to the first as a means of class continuation. The relationship between inter-generational and intra-generational mobility can be identified in the "counter-mobility" of some of the middle class.

```
SELECT IF (SEX EQ 1)
RECODE CLASS CLASS1 PACLASS (1,2,3=1)(4,5,6=2)
VALUE LABELS CLASS CLASS1 PACLASS 1 'NON-MANUAL'
             2 'MANUAL'
CROSSTABS CLASS BY CLASS1 BY PACLASS
OPTIONS 4
```

Then do the same for women:

```
SELECT IF (SEX EQ 2)
```

From the data in the resulting tables, draw charts (one for men and one for women) showing the trajectory from class of origin to current occupational class (this can be a variation of the table illustrated for Exercise 6.2 with columns for Father's Class, Class in First Job, and Class in Current Job). You can measure downward mobility from father's class to first occupational class and then see the extent to which later intra-generational mobility can make up for any initial downward inter-generational mobility.

Since the exercise, for the sake of simplicity involves dichotomised occupational class for women along manual/non-manual lines, results for women will not be as clear as those for men. In particular, women will appear to be more often upwardly mobile inter-generationally. You may wish to consider different ways of measuring and dichotomising women's occupational class. For example, income could be taken into consideration as a means of dividing women in non-manual work, so that poorly paid non-manual workers are classed with women in manual work.

Exercise 7.3:

The following analysis will make use of a longitudinal class variable, which categorises young people according to their class trajectories from class of origin, through their age at leaving continuous full-time education, to their occupational class in their first job, and finally their current educational class. Those who have been socially mobile both inter-generationally and intra-generationally can thus be identified (Jones, 1987b).

The "Youth Class" variable was derived from a class trajectory chart similar to the one you may have just completed. It classifies people according to the following categories:

1 **Stable Middle Class**
comprising those of middle class backgrounds who appear to move directly into non-manual work (reflecting the effect of class origins and education);

2 **Education-Mobile Working Class**
those of working class backgrounds who achieve upward mobility into non-manual work through full-time education;

3 **Counter-Mobile Middle Class**
those of middle class backgrounds who enter manual work (or, in the case of women, low-grade non-manual work), and later retrieve their class positions through work mobility, or a combination of work and education routes;

4 **Work-Route Working Class**
those from working class families who achieve upward mobility, through work rather than education routes;

5 **Downwardly-Mobile Middle Class**
middle class early education-leavers who enter manual work, some of whom will become counter-mobile in time, while some will remain downwardly mobile;

6 **Stable Working Class**
those from working class backgrounds who are early school leavers, move into manual work and are unlikely to be upwardly mobile.

The work route to upward mobility of the working class, or "counter-mobility" of the middle class could be identified in the last exercise. How is mobility through work route achieved? Some of the work-related routes to upward intra-generational mobility (further education, apprenticeship, and training since leaving school) can now be examined with the variable FETRAP. Can they be seen as routes to upward mobility, and if so, is access to them shared? Who does what?

```
CROSSTABS FETRAP BY YCLASS BY SEX
OPTIONS 4
```

Exercise 7.4:

The NCDS contains information about parents as well as about the cohort themselves, so it is possible to extend the analysis of intra-generational mobility by examining how the class career of the respondents' fathers have developed over time. The occupational class mobility of the respondent's father over a period of 16 years could be examined, using his occupational class in 1958, 1965, 1969 and 1974. Intra-generational mobility can thus be examined at four points in the father's life course.

The following exercise examines intra-generational mobility of fathers over 8 years, between 1965 and 1974. Fathers are likely to be at their most occupationally stable during the family building years in their lives, and this is why most studies of social mobility study men between 25 and 50 years of age. The analysis will show the extent to which upward mobility through work careers occurred for the fathers of the NCDS cohort.

It is important first to ensure that the father figure has not changed over time, and that for each respondent, the same father is being referred to in the data. The variable SAMEMAPA is therefore needed here, so that we can be sure we are reconstructing the class careers of the same father throughout.

```
SELECT IF (SAMEMAPA EQ 1)
RECODE PACLS74 PACLS69 PACLS65 (1,2,3=1)(4,5,6=2)
VALUE LABELS PACLS74 PACLS69 PACLS65 1 'NON-MANUAL'
             2 'MANUAL'
CROSSTABS PACLS74 BY PACLS69 BY PACLS65
OPTIONS 4
```

8. **HOUSING**

Housing tenure is sometimes used in sociology research as an indicator of social class. It has also been considered as a dimension of stratification in itself (the notion of "housing class"). Both approaches are problematic. In the following exercises, we shall be looking at housing as an outcome of social stratification rather than as a dimension of it, and will see the effect of a further variable, marital status, on housing outcomes.

Just as the labour market and the industrial structure have affected occupational class distributions over time, so the changing housing market has affected housing tenure. The post-war period saw an increase in public sector rented housing, particularly in the 1950s and 1960s. Overlapping with this trend has been a decrease in the amount of private rented housing available. The three tier structure of home-ownership, private rented tenancy and council tenancy has largely given way to a two tier structure which omits the private rented sector. In recent years, government policies of encouraging the public to buy property and councils to sell their housing stock, is leading to a housing market which provides ever more limited choice.

Research Questions

The notion of housing class can be examined, as well as the relationship between housing tenure and occupational class. The exercises will provide arguments for and against using tenure instead of occupational class as a measure of inequality.

The idea of "housing careers", equivalent to "class careers" can be assessed, but the difficulty of identifying a housing career in the context of overall changes in the housing market should be considered.

Perhaps the most important issue here is the extent to which tenure has become less associated with class over time, as more of the working class become home owners.

Variables Needed:

HOH	Relation to Head of Household
TENURE	Current tenure
TENURE1	First tenure on first leaving home
PTEN74	Parents' tenure in 1974
PTEN69	Parents' tenure in 1969
PTEN65	Parents' tenurein1965
MARITAL	Marital Status
CLASS	Occupational Class at 23 years
PACLS65	Fathers' occupational class in 1965
PACLS69	Fathers' occupational class in 1965
PACLS74	Fathers' occupational class in 1974

Suggested Analysis

Exercise 8.1:

This exercise examines the housing careers of the family of origin and will show both the housing careers of the parents, and housing trends. The results will show the difficulty of separating age and housing trend effects.

First, recode the tenure variables into the following categories:
1) Owned 2) Rented.

```
RECODE PTEN65 PTEN69 PTEN74 (   =1)(   =2)
VALUE LABELS PTEN65 PTEN69 PTEN74
             1 'OWNED' 2 'RENTED'
CROSSTABS PTEN74 BY PTEN69 BY PTEN65
OPTIONS 4
```

The easiest way to examine the housing careers of parents from 1965 to 1974 will be to draw career paths in the same way as class trajectories. How have housing trends changed over the years?

Exercise 8.2:

Try an analysis by PACLASS too, to see to what extent homeownership has increased among the working class over time. You can either re-do the above analysis controlling for PACLASS, or relate tenure to occupational class at the time, as follows:

First, you may wish to recode tenure into three categories, homeownership, private rented and public rented, so that you will see not only the increase in ownership, but also the decrease in the private rented sector over time.

```
RECODE PTEN65 PTEN69 PTEN74 (   =1)(   =2)(   =3)/
       PACLS65 PACLS69 PACLS74 (1,2,3=1)(4,5,6=2)/
VALUE LABELS PTEN65 PTEN69 PTEN74
       1 'OWNED' 2 'PRIVATE RENT' 3 'PUBLIC RENT'/
       PACLS65 PACLS69 PACLS74
       1 'NON-MANUAL' 2 'MANUAL'/
CROSSTABS PTEN65 BY PACLS65/
          PTEN69 BY PACLS69/
          PTEN74 BY PACLS74/
OPTIONS 4
STATISTICS 1 2
```

This exercise will show the changing relationship between tenure and occupational class over time.

Exercise 8.3:

Gender and class differences in tenure are reduced in young adulthood because of the patterns whereby the working class and women of all classes are likely to marry earlier than middle class men (Jones, 1987). In young adulthood, women are more likely than men to be home owners, and similarly the class differentials are small.

Among the NCDS cohort, some have moved away from their parents and are householders in their own right. Their tenure is more likely to be associated with their current occupational class than in the case of young people living in their parent's home. Analyse their tenure in their homes at 23 years.

```
      RECODE HOH (2,3=1)
      SELECT IF (HOH EQ 1)
      CROSSTABS TENURE BY SEX/
                TENURE by CLASS/
      OPTIONS 4
```

Exercise 8.4:

To what extent, then, is homeownership in young adulthood associated with marital status rather than occupational class? The exercise shows differences according to the marital circumstances of members of the NCDS cohort. The importance of marital circumstances as an independent variable in housing analysis, especially among this age group, is brought to light, while the use of tenure as an indicator of inequality in youth is questioned.

```
      RECODE HOH (2,3=1)
      SELECT IF (HOH EQ 1)
      RECODE TENURE (    =1)(    =2)/
             CLASS (1,2,3=1)(4,5,6=2)/
      VALUE LABELS TENURE 1 'OWNED' 2 'RENTED'/
                   CLASS 1 'NON-MANUAL' 2 'MANUAL'
      CROSSTABS TENURE BY CLASS BY MARITAL BY SEX
```

Exercise 8.3:

Does housing tenure follow inter-generational patterns in the same way as occupational class? The following exerice is restricted to those in the cohort who have been married, and compares their marital housing circumstances with the parental homes they have left.

```
        SELECT IF (MARITAL EQ 2)
        RECODE TENURE PTEN74 (    =1)(    =2)/
        VALUE LABELS TENURE PTEN74 1 'OWNED' 2 'RENTED'/
        CROSSTABS TENURE BY PTEN74
        OPTIONS 4
```

Exercise 8.4:

Finally, in this analysis of housing circumstances and trends, we will examine housing "careers" to date of the cohort, among those who are householders, rather than living with parents or other relatives.

```
        RECODE HOH (3,2=1)
        SELECT IF (HOH EQ 1)
        RECODE TENURE (    =1)(    =2)/
        VALUE LABELS TENURE 1 'OWNED' 2 'RENTED'/
        CROSSTABS TENURE BY TENURE1
        OPTIONS 4
```

Is there any evidence of a progression to "better" accommodation in the early housing careers of the NCDS cohort?

9 CLASS IDENTIFICATION

The subjective area of class identification cannot easily be examined with survey data. Nevertheless it is important to see to what extent social relationships may be associated with class. The NCDS allows examination of two outcomes of class, which give some indication at least of class identification. These are voting behaviour and trade union membership.

Research Questions

The question, then, is to what extent does occupational class affect voting and trade union membership? Can the notion of class trajectory, operationalised in the variable YCLASS, be seen as a more effective indicator of class identification? What are the gender differences?

Variables Needed

SEX	Sex of Respondent
CLASS	Occupational Class at 23 years
YCLASS	Respondent's Class Trajectory
VOTING	How voted in 1979 General Election
TRADEUN	Trade Union or Staff Association member

Suggested Analyses

Exercise 9.1: Voting Behaviour

To what extent is Labour party voting associated with occupational class? Is the voting behaviour of women as closely associated with their occupational class as that of men?

Two indicators of class are compared here; current occupational class, and a longitudinal measure of class describing the class trajectory. Since the General Election took place in 1979 and the Current Occupational Class is measured in 1981, then a longitudinal measure of class may reflect with greater accuracy the respondent's class position in 1979. You may wish to recode variables at the start of the exercise.

```
      CROSSTABS VOTING BY CLASS BY SEX/
                VOTING BY YCLASS BY SEX/
      OPTIONS 4
```

Exercise 9.2: Trade Union Membership

The NCDS Respondents were asked in 1981 whether they had ever been members of a Trade Union or Staff Association. To what extent is membership likely to be associated with class, measured either at 23 years or longitudinally. What are the gender differences?

Again, you may wish to recode variables at the start of the exercise.

```
         CROSSTABS TRADEUN BY CLASS BY SEX/
                   TRADEUN BY YCLASS BY SEX/
      OPTIONS 4
```

10 REFERENCES

Bernstein, B.(1977) Class, Codes and Control, Vol 3, Towards a Theory of Educational Transmission, 2nd ed., Routledge and Kegan Paul, London.

Blau, P.M. and Duncan, O.D. (1967) The American Occupational Structure, Wiley, New York.

Bowles, S. and Gintis, H.(1975) Schooling in Capitalist America, Basic Books, New York.

Delphy, C. (1981) "Women in Stratification Studies",in H. Roberts (ed) Doing Feminist Research, Routledge, London.

Dex, S.(1984) Women's Work Histories, Research Paper No 46, Department of Employment, HMSO, London.

Goldthorpe, J.(1984) "Women and Class Analysis: In Defence of the Conventional View", Sociology, Vol 17, No 4.

Goldthorpe, J.(1980) Social Mobility and Class Structure in Modern Britain, Clarendon Press, Oxford.

Goldthorpe, J. et al. (1969) The Affluent Worker in the Class Structure, Cambridge University Press, Cambridge.

Heath, A.(1981) Social Mobility, Fontana Paperbacks.

Heath, A. and Britten, N.(1984) "Women's Jobs do Make a Difference", Sociology, Vol 18, No 4,,pp 475-490.

Jones, G.E. (1986) Youth in the Social Structure: Transitions to Adulthood and their Stratification by Class and Gender, Unpublished PhD Dissertation, University of Surrey.

Jones, G.E. (1987a) "Leaving the Parental Home: an Analysis of Early Housing Careers", Journal of Social Policy, Vol 16, No 1, pp 49-74.

Jones, G.E. (1987b) "Young Workers in the Class Structure", Work, Employment and Society, 1,4.

Norusis, M.J. (1983) SPSS-X Introductory Statistics Guide, McGraw-Hill, New York.

Oakley, A. (1974) The Sociology of Housework, Martin Robertson, Bath.

Office of Population Censuses and Surveys (1980) Classification of Occupations 1980, OPCS, HMSO, London.

Ossowski, S. (1963) Class Structure in the Social Consciousness, London.

Pahl, J. (1983) "The Allocation of Money and the Structuring of Inequality within Marriage", Sociological Review, 31,2.

Westergaard, J. and Resler (1975) Class in Capitalist Society: a Study of Contemporary Britain, Heinemann, London.

Wright, E.O. (1976) "Class Boundaries in Advanced Capitalist Societies", New Left Review, No.98, pp3-41.

Wright, E. Olin (1978) Class, Crisis and the State, 2nd edition, Verso.

11 A NOTE ON SPSS-X

The standard data set is supplied as an SPSS-X Version 2.2 system file with labels attached to each variable.

Each exercise contains clear SPSS-X instructions. In addition, students will need to know how to access the system file on their institution's computer system. This will probably involve adding FILE HANDLE and GET FILE commands to the instructions given.

If different or more complex analyses are required, it is recommended that the student reads the SPSS-X Manual. The exercises suggested in the text use only the simplest computing techniques. These are briefly explained below in relation to the SPSS-X instructions needed for Exercise

Example SPSS-S File

```
SELECT IF (CHILDREN EQ 0)
RECODE EARNINGS (    THRU  =1)(    THRU    =2)/
       WORKHRS (1 THRU 29=1)(30 THRU HI=2)/
VALUE LABELS EARNINGS 1 '    ' 2 '    '/
       WORKHRS 1 'PART-TIME' 2 'FULL-TIME'/
CROSSTABS EARNINGS BY SEX BY WORKHRS
OPTIONS 4
STATISTICS 1 2
```

SELECT IF - This optional instruction allows you to select a subset of cases. In the example, only respondents without children were selected. "EQ" (equal to) is a logical operator. Other logical operators, such as "NE" (not equal to) can be used in the same way.

RECODE - Variables can be recoded into fewer categories than are held in the full data, for clearer analysis. Any number of variables can be recoded as above, but where a string of variables are recoded differently, a backslash "/" is needed in the SPSS-X instructions. If two or more variables are to be recoded in the same way (in some Exercises, CLASS and PACLASS for example) the variables to be recoded can be listed together, with a space between each variable, and followed with the recode specification:

e.g. RECODE CLASS PACLASS (1,2,3=1)(4,5,6=2)

Continuations to the RECODE line can start anywhere except in Column 1 on the screen.

VALUE LABELS - These are optional, and for human benefit only. The data set already contains value labels for each variable. However, you may like to revise the value labels of variables you have recoded (EARNINGS and WORKHRS in the example), so that tables can be more easily read.

CROSSTABS - The CROSSTABS instruction is followed by the Dependent Variables, then the Independent Variable(s). Where several crosstabulations are needed, each is followed by a backslash "/", so that SPSS-X can identify the end of the instruction. In the example, EARNINGS is the Dependent Variable.

OPTIONS - The SPSS-X Manual gives all the Options available. Only three are used here: Option 4 causes the column percentages to be produced in tables; Option 16 produces Standardised Residuals from a model of no association; Option 7 causes Missing Values to be included in the Tabulation.

STATISTICS - Many statistics are available. Try adding

```
      STATISTICS ALL
```

to the end of one of the exercises. The most useful statistics, which you may wish to obtain for several of the exercises are numbers 1 CHI-SQUARED and 2 PHI for 2 by 2 Tables and CRAMER'S V for larger tables. See Norusis (1983) for further information.

12 ALPHABETICAL LIST OF VARIABLES

Variable names and Labels are given with a brief explanation and their position in the Codebook.

CHILDREN	(30)	Whether Has Children Whether respondent has children at 23 years
CLASS	(3)	Occupational Class at 23 years Occupational Class of respondent in current or last job at 23 years
CLASS1	(2)	Occupational Class in first job Occupational Class of respondent in first job. This could be the same as the current job where the respondent has only had one job
EARNINGS	(13)	Respondent's Hourly Income Gross hourly income in current or last job
EDLAGE	(16)	Age in years left F-T cont education Age respondent left full-time continuous education.
EMPSTAT	(15)	Employment Status
FETRAP	(20)	Further education, training or apprenticeship Whether respondent received post-school further education, training or apprenticeship
HIGHQUAL	(19)	Highest educ qualification at 23 Highest educational qualification of respondent by 23 years
HOH	(26)	Relationship to HOH Relationship of respondent to head of household
MACLS58	(9)	Mother's occupational class in 1958 Mother's occupational class at start of pregnancy in 1958
MACLS74	(10)	Mother's Occupational Class in 1974
MARITAL	(29)	Marital Status Marital status of respondent at 23 years
MEDLAGE	(18)	Age mother left full-time education

MGFCLASS (11) MGF's Occupational Class
Maternal Grandfather's occupational class when mother left school

PACLASS (4) Father's Occupational Class 1974 or 1969
Father's occupational class in 1974, or earlier if 1974 data missing

PACLS58 (5) Fathers' occupational class in 1958

PACLS65 (6) Fathers' occupational class in 1965

PACLS69 (7) Fathers' occupational class in 1965

PACLS74 (8) Fathers' occupational class in 1974

PEDLAGE (17) Age father left full-time education

PGFCLASS (12) PGF's Occupational Class
Paternal Grandfather's Occupational Class when father left school

PTEN65 (23) Parents'tenure in 1965

PTEN69 (24) Parents' tenure in 1969

PTEN74 (25) Parents' tenure in 1974

SAMEMAPA (28) With own parents in 1965, 1969 and 1974
Whether respondent has been with the same parents at each Sweep of the NCDS

SEX (27) Sex of Respondent

TENURE (21) Current tenure
Current tenure of respondent. Where the respondent is living with family or friends, this variable refers to the tenure of the Head of Household

TENURE1 (22) First tenure on first leaving home
First tenure of respondent after leaving the parental home

TRADEUN (32) Trade Union or Staff Association member
Is the respondent a member of a trade union or staff association?

VOTING	(31)	How voted in 1979 General Election
WORKHRS	(14)	Hours of Work Hours of work in current job
YCLASS	(1)	"Youth Class" - a longitudinal class measure. See notes in text.

TENURE1 FIRST TENURE ON FIRST LEAVING HOME

```
                                                        VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

OWNER OCCUPIED                1.00        359     18.0     24.7     24.7
COUNCIL RENTED                2.00        111      5.6      7.6     32.4
HOUSING CHARITY               3.00         27      1.4      1.9     34.2
PRIVATE RENTED                4.00        211     10.6     14.5     48.8
SHARE WITH KIN                5.00        134      6.7      9.2     58.0
SHARE WITH OTHERS             6.00        163      8.1     11.2     69.2
OTHER                         7.00        447     22.4     30.8    100.0
                             -9.00        548     27.4  MISSING
                                      -------  -------  -------
                             TOTAL       2000    100.0    100.0

VALID CASES     1452        MISSING CASES    548
```

PTEN65 PARENTS' TENURE IN 65

```
                                                        VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

OWNER OCCUPIED                1.00        753     37.7     44.0     44.0
COUNCIL RENTED                2.00        688     34.4     40.2     84.2
PRIVATE RENTED                3.00        184      9.2     10.8     95.0
OTHER                         4.00         86      4.3      5.0    100.0
                             -9.00        289     14.5  MISSING
                                      -------  -------  -------
                             TOTAL       2000    100.0    100.0

VALID CASES     1711        MISSING CASES    289
```

PTEN69 PARENTS' TENURE IN 69

```
                                                        VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

OWNER OCCUPIED                1.00        824     41.2     48.5     48.5
COUNCIL RENTED                2.00        701     35.1     41.3     89.8
PRIVATE RENTED                3.00        103      5.2      6.1     95.8
OTHER                         4.00         71      3.6      4.2    100.0
                             -9.00        301     15.1  MISSING
                                      -------  -------  -------
                             TOTAL       2000    100.0    100.0

VALID CASES     1699        MISSING CASES    301
```

HIGHQUAL HIGHEST EDUCATIONAL QUAL AT 23

```
                                                            VALID      CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

 DEGREE                       1.00        225     11.3      11.3      11.3
 OTHER HIGHER QUALIFI         2.00        185      9.3       9.3      20.5
 5 O LEVELS TO 2 A L.         3.00        582     29.1      29.1      49.6
 LESS THAN 5 O LEVELS         4.00        478     23.9      23.9      73.5
 NO QUALIFICATIONS            5.00        530     26.5      26.5     100.0
                                      -------  -------   -------
                             TOTAL       2000    100.0     100.0

 VALID CASES     2000       MISSING CASES       0
```

FETRAP FURTHER EDUCATION OR TRAINING

```
                                                            VALID      CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

 NO EDUC-TRAING-APP            0.0        579     29.0      29.0      29.0
 JUST APPRENTICE              1.00        282     14.1      14.1      43.1
 APPRENTICE + TRAING          2.00         62      3.1       3.1      46.2
 APPRENTICE + EDUC            3.00         59      3.0       3.0      49.1
 APPRENT + TRNG + EDU         4.00         15       .8        .8      49.8
 JUST TRAINING                5.00        313     15.7      15.7      65.5
 TRAINING + EDUC              6.00        202     10.1      10.1      75.6
 JUST EDUCATION               7.00        488     24.4      24.4     100.0
                                      -------  -------   -------
                             TOTAL       2000    100.0     100.0

 VALID CASES     2000       MISSING CASES       0
```

TENURE CURRENT TENURE

```
                                                            VALID      CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

 OWNER OCCUPIED               1.00        602     30.1      31.1      31.1
 COUNCIL RENTED               2.00        244     12.2      12.6      43.7
 HOUSING CHARITY              3.00         43      2.2       2.2      45.9
 PRIVATE RENTED               4.00        192      9.6       9.9      55.8
 SHARE WITH KIN               5.00        730     36.5      37.7      93.5
 SHARE WITH OTHERS            6.00         42      2.1       2.2      95.7
 OTHER                        7.00         84      4.2       4.3     100.0
                             -9.00         63      3.2   MISSING
                                      -------  -------   -------
                             TOTAL       2000    100.0     100.0

 VALID CASES     1937       MISSING CASES      63
```

EDLAGE AGE LEFT FT CONT EDUC

```
                                                   VALID     CUM
  VALUE LABEL           VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

                        15.00         28      1.4      1.4      1.4
                        16.00       1222     61.1     61.2     62.6
                        17.00        203     10.2     10.2     72.8
                        18.00        245     12.3     12.3     85.0
                        19.00         47      2.4      2.4     87.4
                        20.00         22      1.1      1.1     88.5
                        21.00        110      5.5      5.5     94.0
                        22.00         83      4.2      4.2     98.1
                        23.00         37      1.9      1.9    100.0
                        -9.00          3       .2  MISSING
                                 -------  -------  -------
                        TOTAL       2000    100.0    100.0

VALID CASES     1997       MISSING CASES      3
```

PEDLAGE AGE FATHER LEFT FT EDUCATION

```
                                                   VALID     CUM
  VALUE LABEL           VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

15 YEARS & UNDER         1.00        842     42.1     58.5     58.5
OVER 15 YEARS            2.00        598     29.9     41.5    100.0
                        -9.00        560     28.0  MISSING
                                 -------  -------  -------
                        TOTAL       2000    100.0    100.0

VALID CASES     1440       MISSING CASES    560
```

MEDLAGE AGE MOTHER LEFT FT EDUCATION

```
                                                   VALID     CUM
  VALUE LABEL           VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

15 YEARS & UNDER         1.00        718     35.9     49.2     49.2
OVER 15 YEARS            2.00        740     37.0     50.8    100.0
                        -9.00        542     27.1  MISSING
                                 -------  -------  -------
                        TOTAL       2000    100.0    100.0

VALID CASES     1458       MISSING CASES    542
```

EARNINGS HOURLY INCOME <IN PENCE>

```
                                                         VALID      CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

VARIES                        -1.00         45      2.3      3.2      3.2
0 TO 119                       1.00        128      6.4      9.2     12.4
120 TO 159                     2.00        407     20.4     29.2     41.6
160 TO 209                     3.00        457     22.9     32.8     74.4
210 TO 259                     4.00        245     12.3     17.6     92.0
260 TO 339                     5.00         87      4.4      6.2     98.2
340 TO 499                     6.00         19      1.0      1.4     99.6
500 OR MORE                    7.00          6       .3       .4    100.0
                              -9.00        606     30.3  MISSING
                                       -------  -------  -------
                              TOTAL       2000    100.0    100.0

VALID CASES     1394      MISSING CASES    606
```

WORKHRS WORKING HOURS PER WEEK

```
                                                         VALID      CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

VARIES                        -1.00         65      3.3      4.4      4.4
1 TO 29                        1.00         72      3.6      4.9      9.3
30 TO 39                       2.00        648     32.4     43.8     53.1
40 TO 49                       3.00        521     26.1     35.2     88.3
50 OR MORE                     4.00        173      8.7     11.7    100.0
                              -9.00        521     26.1  MISSING
                                       -------  -------  -------
                              TOTAL       2000    100.0    100.0

VALID CASES     1479      MISSING CASES    521
```

EMPSTAT EMPLOYMENT STATUS

```
                                                         VALID      CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

FULL TIME EDUCATION            1.00         59      3.0      3.0      3.0
EMPLOYED                       2.00       1476     73.8     73.9     76.9
UNEMPLOYED                     3.00        178      8.9      8.9     85.8
ECONOM INACTIVE                4.00        284     14.2     14.2    100.0
                              -9.00          3       .2  MISSING
                                       -------  -------  -------
                              TOTAL       2000    100.0    100.0

VALID CASES     1997      MISSING CASES      3
```

MACLS74 MOTHER'S OCC CLASS IN 74

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT  PERCENT

HIGHER PROF                    1.00          9       .5        .9       .9
MAN-EMP &LOWPROF               2.00        161      8.1      16.6     17.6
JUNIOR NON-MANUAL              3.00        335     16.8      34.6     52.2
SKILLED MANUAL                 4.00        176      8.8      18.2     70.4
SEMI-SKILLED                   5.00        287     14.4      29.6    100.0
                              -9.00       1032     51.6   MISSING
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES      968        MISSING CASES   1032
```

MGFCLASS MGF'S OCCUPATIONAL CLASS

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT  PERCENT

HIGHER PROF                    1.00         53      2.7       3.5      3.5
MAN-EMP &LOWPROF               2.00        267     13.4      17.5     21.0
JUNIOR NON-MANUAL              3.00         48      2.4       3.2     24.2
SKILLED MANUAL                 4.00        640     32.0      42.0     66.2
SEMI-SKILLED                   5.00        291     14.6      19.1     85.3
UNSKILLED                      6.00        224     11.2      14.7    100.0
                              -9.00        477     23.9   MISSING
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES     1523        MISSING CASES    477
```

PGFCLASS PGF'S OCCUPATIONAL CLASS

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT  PERCENT

HIGHER PROF                    1.00         28      1.4       2.0      2.0
MAN-EMP &LOWPROF               2.00        218     10.9      15.6     17.6
JUNIOR NON-MANUAL              3.00        139      7.0       9.9     27.5
SKILLED MANUAL                 4.00        560     28.0      40.0     67.5
SEMI-SKILLED                   5.00        327     16.4      23.4     90.9
UNSKILLED                      6.00        127      6.3       9.1    100.0
                              -9.00        601     30.1   MISSING
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES     1399        MISSING CASES    601
```

PACLS69 FATHER'S OCC CLASS IN 69

```
                                                            VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT    PERCENT   PERCENT

HIGHER PROF                   1.00        83      4.2        5.0       5.0
MAN-EMP &LOWPROF              2.00       377     18.9       22.9      27.9
JUNIOR NON-MANUAL             3.00       142      7.1        8.6      36.6
SKILLED MANUAL                4.00       712     35.6       43.3      79.8
SEMI-SKILLED                  5.00       250     12.5       15.2      95.0
UNSKILLED                     6.00        82      4.1        5.0     100.0
                             -9.00       354     17.7    MISSING
                                     -------  -------    -------
                             TOTAL      2000    100.0      100.0

VALID CASES     1646       MISSING CASES    354
```

PACLS74 FATHER'S OCC CLASS IN 74

```
                                                            VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT    PERCENT   PERCENT

HIGHER PROF                   1.00        74      3.7        5.4       5.4
MAN-EMP &LOWPROF              2.00       328     16.4       24.1      29.5
JUNIOR NON-MANUAL             3.00        90      4.5        6.6      36.1
SKILLED MANUAL                4.00       613     30.7       45.0      81.2
SEMI-SKILLED                  5.00       185      9.3       13.6      94.8
UNSKILLED                     6.00        71      3.6        5.2     100.0
                             -9.00       639     32.0    MISSING
                                     -------  -------    -------
                             TOTAL      2000    100.0      100.0

VALID CASES     1361       MISSING CASES    639
```

MACLS58 MOTHER'S OCC CLASS IN 58

```
                                                            VALID     CUM
  VALUE LABEL                VALUE  FREQUENCY  PERCENT    PERCENT   PERCENT

MAN-EMP &LOWPROF              2.00        72      3.6       10.1      10.1
JUNIOR NON-MANUAL             3.00       262     13.1       36.8      46.9
SKILLED MANUAL                4.00        40      2.0        5.6      52.5
SEMI-SKILLED                  5.00       299     15.0       42.0      94.5
UNSKILLED                     6.00        39      2.0        5.5     100.0
                             -9.00      1288     64.4    MISSING
                                     -------  -------    -------
                             TOTAL      2000    100.0      100.0

VALID CASES      712       MISSING CASES   1288
```

PACLASS FATHER'S OCCUPATIONAL CLASS

```
                                                        VALID       CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

 HIGHER PROF                   1.00        104      5.2      5.6      5.6
 MAN-EMP &LOWPROF              2.00        432     21.6     23.1     28.6
 JUNIOR NON-MANUAL             3.00        135      6.8      7.2     35.9
 SKILLED MANUAL                4.00        838     41.9     44.8     80.7
 SEMI-SKILLED                  5.00        268     13.4     14.3     95.0
 UNSKILLED                     6.00         94      4.7      5.0    100.0
                              -9.00        129      6.5  MISSING
                                     -------  -------  -------
                              TOTAL       2000    100.0    100.0

 VALID CASES     1871       MISSING CASES    129
```

PACLS58 FATHER'S OCC CLASS IN 58

```
                                                        VALID       CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

 HIGHER PROF                   1.00         73      3.7      4.2      4.2
 MAN-EMP &LOWPROF              2.00        253     12.6     14.5     18.7
 JUNIOR NON-MANUAL             3.00         89      4.5      5.1     23.8
 SKILLED MANUAL                4.00        869     43.5     49.9     73.7
 SEMI-SKILLED                  5.00        302     15.1     17.3     91.0
 UNSKILLED                     6.00        156      7.8      9.0    100.0
                              -9.00        258     12.9  MISSING
                                     -------  -------  -------
                              TOTAL       2000    100.0    100.0

 VALID CASES     1742       MISSING CASES    258
```

PACLS65 FATHER'S OCC CLASS IN 65

```
                                                      VALID      CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

 HIGHER PROF                   1.00         89      4.5      5.3      5.3
 MAN-EMP &LOWPROF              2.00        269     13.5     16.0     21.3
 JUNIOR NON-MANUAL             3.00        195      9.8     11.6     33.0
 SKILLED MANUAL                4.00        748     37.4     44.6     77.6
 SEMI-SKILLED                  5.00        275     13.8     16.4     94.0
 UNSKILLED                     6.00        101      5.1      6.0    100.0
                              -9.00        323     16.2  MISSING
                                     -------  -------  -------
                              TOTAL       2000    100.0    100.0

 VALID CASES     1677       MISSING CASES    323
```

YCLASS YOUTH CLASS

```
                                                          VALID     CUM
 VALUE LABEL              VALUE  FREQUENCY   PERCENT    PERCENT   PERCENT

STABLE MIDDLE              1.00        305      15.2       17.4      17.4
EDUC-MOBILE W-C            2.00        356      17.8       20.3      37.7
COUNTER-MOBILE             3.00        107       5.4        6.1      43.8
WORK-MOBILE W-C            4.00        115       5.8        6.6      50.4
DOWNWARD M-C               5.00        199      10.0       11.4      61.8
STABLE WORKING             6.00        670      33.5       38.2     100.0
                          -9.00        248      12.4    MISSING
                                   -------   -------    -------
                          TOTAL       2000     100.0      100.0

VALID CASES     1752      MISSING CASES    248
```

CLASS1 OCCUPATIONAL CLASS IN 1ST JOB

```
                                                          VALID     CUM
 VALUE LABEL              VALUE  FREQUENCY   PERCENT    PERCENT   PERCENT

HIGHER PROF                1.00         59       3.0        3.1       3.1
MAN-EMP& PROF              2.00        197       9.9       10.4      13.5
JUNIOR NON-MAN             3.00        565      28.3       29.7      43.2
SKILLED                    4.00        432      21.6       22.7      65.9
SEMI-SKILLED               5.00        560      28.0       29.5      95.4
UNSKILLED                  6.00         87       4.4        4.6     100.0
                          -9.00        100       5.0    MISSING
                                   -------   -------    -------
                          TOTAL       2000     100.0      100.0

VALID CASES     1900      MISSING CASES    100
```

CLASS OCCUPATIONAL CLASS AT 23

```
                                                          VALID     CUM
 VALUE LABEL              VALUE  FREQUENCY   PERCENT    PERCENT   PERCENT

HIGHER PROF                1.00         76       3.8        4.0       4.0
MAN-EMP &LOWPROF           2.00        332      16.6       17.6      21.6
JUNIOR NON-MANUAL          3.00        542      27.1       28.7      50.4
SKILLED MANUAL             4.00        423      21.2       22.4      72.8
SEMI-SKILLED               5.00        440      22.0       23.3      96.1
UNSKILLED                  6.00         73       3.7        3.9     100.0
                          -9.00        114       5.7    MISSING
                                   -------   -------    -------
                          TOTAL       2000     100.0      100.0

VALID CASES     1886      MISSING CASES    114
```

PTEN74 PARENTS' TENURE IN 74

```
                                                       VALID      CUM
  VALUE LABEL              VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

OWNER OCCUPIED              1.00       793     39.7     53.6     53.6
COUNCIL RENTED              2.00       571     28.6     38.6     92.2
PRIVATE RENTED              3.00        69      3.5      4.7     96.8
OTHER                       4.00        47      2.4      3.2    100.0
                           -9.00       520     26.0  MISSING
                                   -------  -------  -------
                           TOTAL      2000    100.0    100.0

VALID CASES     1480      MISSING CASES    520
```

HOH RELATIONSHIP TO HOH

```
                                                       VALID      CUM
  VALUE LABEL              VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

HEAD OF HOUSEHOLD           1.00       253     12.6     12.7     12.7
PARTNER OF HOH              2.00       135      6.8      6.8     19.4
JOINT HEAD                  3.00       760     38.0     38.0     57.5
PARENT IS HOH               4.00       687     34.4     34.4     91.8
OTHER RELATIVE IS HO        5.00        42      2.1      2.1     93.9
OTHER                       6.00       121      6.1      6.1    100.0
                           -9.00         2       .1  MISSING
                                   -------  -------  -------
                           TOTAL      2000    100.0    100.0

VALID CASES     1998      MISSING CASES      2
```

SEX SEX OF RESPONDENT

```
                                                       VALID      CUM
  VALUE LABEL              VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

MALE                        1.00       997     49.8     49.8     49.8
FEMALE                      2.00      1003     50.2     50.2    100.0
                                   -------  -------  -------
                           TOTAL      2000    100.0    100.0

VALID CASES     2000      MISSING CASES      0
```

SAMEMAPA WITH OWN PARENTS 58-74

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

YES                            1.00       1006     50.3      85.6      85.6
NO                             2.00        169      8.5      14.4     100.0
                              -9.00        825     41.3   MISSING
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES     1175       MISSING CASES   825
```

MARITAL MARITAL STATUS

```
                                                                VALID     CUM
     VALUE LABEL                     VALUE  FREQUENCY  PERCENT  PERCENT   PERCENT

SINGLE                                1.00        914     45.7     45.7      45.7
LIVING AS MARRIED                     2.00       1032     51.6     51.6      97.3
DIVORCED,WIDOWED,SEP                  3.00         53      2.7      2.7     100.0
                                     -9.00          1       .1  MISSING
                                              -------  -------  -------
                                     TOTAL       2000    100.0    100.0

VALID CASES     1999       MISSING CASES      1
```

CHILDREN WHETHER HAS CHILDREN

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

NO CHILDREN                     0.0       1497     74.9      74.9      74.9
HAS CHILDREN                   1.00        503     25.2      25.2     100.0
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES     2000       MISSING CASES      0
```

VOTING HOW VOTED IN 1979 GEN ELECTION

```
                                                            VALID     CUM
  VALUE LABEL                 VALUE  FREQUENCY  PERCENT   PERCENT   PERCENT

DID NOT VOTE                   1.00        660     33.0      33.7      33.7
CONSERVATIVE                   2.00        508     25.4      25.9      59.7
LABOUR                         3.00        571     28.6      29.2      88.8
LIBERAL                        4.00        172      8.6       8.8      97.6
OTHER PARTY                    5.00         47      2.4       2.4     100.0
                              -9.00         42      2.1   MISSING
                                       -------  -------   -------
                              TOTAL       2000    100.0     100.0

VALID CASES     1958       MISSING CASES     42
```

TRADEUN TRADE UNION OR STAFF ASSOC MEMBER

```
                                                         VALID     CUM
  VALUE LABEL              VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

YES                         1.00        694     34.7     60.6     60.6
NO                          2.00        448     22.4     39.1     99.7
DON"T   KNOW                8.00          4       .2       .3    100.0
                           -9.00        854     42.7  MISSING
                                    -------  -------  -------
                           TOTAL       2000    100.0    100.0

VALID CASES     1146     MISSING CASES    854
```

WTFACTOR

```
                                                         VALID     CUM
  VALUE LABEL              VALUE  FREQUENCY  PERCENT  PERCENT  PERCENT

                            1.00       2000    100.0    100.0    100.0
                                    -------  -------  -------
                           TOTAL       2000    100.0    100.0

VALID CASES     2000     MISSING CASES      0
```