

Doctor of Philosophy

STATISTICAL MODELLING AND ANALYSIS OF ACCIDENT DATA

by

David Thomas While

A dissertation submitted in 1991 in fulfilment
of the requirements for a Phd.

UNIVERSITY OF SOUTHAMPTON, 1991

FACULTY OF MATHEMATICAL STUDIES

STATISTICS

Doctor of Philosophy

STATISTICAL MODELLING AND ANALYSIS OF ACCIDENT DATA

By

David Thomas While

A dissertation submitted in fulfilment
of the requirements for a Phd.

ACKNOWLEDGEMENTS

I would like to thank Dr. S. M. Lewis at the University of Southampton for her encouragement, help and guidance.

CONTENTS		Page no.
ACKNOWLEDGEMENTS		ii
CONTENTS		iii
ABSTRACT		vi
CHAPTER 1. INTRODUCTION		1
1.1	The issue of child accidents	1
1.2	A history of accident research	5
1.2.1	The role of proneness	15
1.2.2	The methodology of two accident studies	26
1.2.3	Child accidents	37
1.3	The data set available for the thesis	43
1.3.1	The British Birth Cohort Survey of 1975	44
1.3.2	The reliability of the data and results	47
1.3.3	Ethnic differences in the data	56
1.3.4	Simple reported accident trends in the data	59
1.3.5	Simple reported accident associations in the male population	76
1.3.6	Simple reported accident associations in the female population	85
1.3.7	A simple comparison of the reported accident associations between gender	90
1.4	The aims of this thesis	96
CHAPTER 2. PROBABILITY MODELS FOR THE REPORTED ACCIDENT DISTRIBUTION		99
2.1	Introduction	99
2.2	Fitting the traditional accident distributions - an investigative tool	104
2.3	Extensions of probability models to include misreporting parameters	110
2.3.1	Generalised Poisson distributions	112
2.3.2	Generalised Negative Binomial distributions	119
2.4	Alternative probability distributions and extensions	121
2.4.1	The Neyman Type A distribution	122
2.4.2	The Polya-Eggenberger distribution	123
2.4.3	The Consul/Jain distribution	125
2.4.4	Compounding the Poisson distribution with the Binomial distribution	127
2.4.5	Generalisation of the Consul/Jain distribution	128
2.5	Truncating the traditional distributions	130
2.6	Regression analysis	132
2.7	Discussion	139

CONTENTS		Page no.
CHAPTER 3. ANALYSIS OF THE BRITISH BIRTH DATA SET, PART I; VARIANTS ON THE POISSON AND NEGATIVE BINOMIAL PROBABILITY DISTRIBUTIONS		142
3.1	Introduction	142
3.2	Fitting the traditional distributions to the full reported accident distribution	144
3.3	Fitting the Poisson accident distribution	158
3.4	Fitting the Negative Binomial accident distribution	162
3.5	Fitting the traditional modelling distributions to the serious reported accident distribution	170
3.6	Anomalous results and their investigation	175
3.7	Assessing the traditional probability distributions for the male and female serious accident subpopulations	179
3.8	Fitting the generalised traditional distributions to the full reported accident distribution	182
3.9	Comparison of the generalised distributions	192
3.10	Discussion	199
CHAPTER 4. ANALYSIS OF THE BRITISH BIRTH DATA SET, PART II; ALTERNATIVE PROBABILITY DISTRIBUTIONS		201
4.1	Introduction	201
4.2	Fitting alternative probability distributions to the full reported accident distribution	203
4.3	Fitting the truncated Poisson and Negative Binomial probability distributions to the truncated reported accident distribution	207
4.4	Looking for evidence of proneness	218
4.5	Discussion	223
CHAPTER 5. ANALYSIS OF THE BRITISH BIRTH DATA SET, PART III; NON-SUPERFICIAL REPORTED ACCIDENTS		225
5.1	Introduction	225
5.2	Modelling different accident outcomes	227
5.3	Modelling the non-superficial reported accident distribution using traditional models	233
5.3.1	Summary of overall results	234
5.4	Modelling the superficial reported accident distribution using traditional models	239
5.4.1	Summary of overall results	240
5.5	Discussion	245

CONTENTS		Page no.
CHAPTER 6. ANALYSIS OF THE BRITISH BIRTH DATA SET, PART IV; REGRESSION ANALYSES		248
6.1	Introduction	248
6.2	GLIM regression analysis	250
6.2.1	GLIM modelling of different reported accident outcomes	253
6.2.2	GLIM modelling of all reported accidents for the males	256
6.2.3	GLIM modelling of non-superficial reported accidents for the males	260
6.2.4	GLIM modelling of superficial reported accidents for the males	263
6.2.5	GLIM modelling of all reported accidents for the females	265
6.2.6	GLIM modelling of non-superficial reported accidents for the females	267
6.2.7	GLIM modelling of superficial reported accidents for the females	270
6.2.8	GLIM modelling of all reported accidents using a Negative Binomial error structure	272
6.3	Logistic regression analysis	275
6.3.1	Logistic modelling of different reported accident outcomes	278
6.3.2	Logistic modelling of all reported accidents for the males	281
6.3.3	Logistic modelling of non-superficial reported accidents for the males	287
6.3.4	Logistic modelling of superficial reported accidents for the males	288
6.3.5	Logistic modelling of all reported accidents for the females	290
6.3.6	Logistic modelling of non-superficial reported accidents for the females	295
6.3.7	Logistic modelling of superficial reported accidents for the females	297
6.4	A comparison of the regression models	300
6.5	Discussion	305
CHAPTER 7. DISCUSSION		306
7.1	Conclusions	306
7.2	Further work	310
REFERENCES		312

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MATHEMATICS

STATISTICS

Doctor of Philosophy

STATISTICAL MODELLING AND ANALYSIS OF ACCIDENT DATA

By David Thomas While

This project investigates different methods of modelling reported accident data in order to identify and quantify the effect of variables influencing accident rates. Strong evidence of differential reporting across the population is identified.

The full reported accident distribution is modelled directly using the traditional accident modelling probability distributions, namely the Poisson and Negative Binomial distributions. The male and female populations are split into non-overlapping sub-populations with respect to a single variable at a time and the reported accident distribution from each sub-population is modelled directly using the traditional accident modelling probability distributions.

It is found that the traditional models fail to adequately fit the data. A variety of alternative models is investigated and generalised models are devised which take into account accident misreporting. It is shown that the best description of the data is given by a model formed from generalising a Negative Binomial model to include the misreporting of two accidents as one accident. An alternative approach to avoid drawing inaccurate conclusions from the study is to focus the analysis on a large, more reliable subset of the data. This is achieved through the definition of a non-superficial accident distribution.

Regression models are investigated using GLIM and logistic regression modelling techniques. Different degrees of accident seriousness are investigated which result in the identification of different significant influencing factors. This provides further evidence that sections of the reported accident data are so unreliable that they lead to misleading conclusions.

CHAPTER 1

INTRODUCTION

In this Chapter we give a brief review of the literature available on previous accident research and discuss relevant accident studies. We examine problems inherent in an accident analysis and complete a preliminary, exploratory analysis of our data set in a search for deficiencies and inadequacies which may affect our later analysis. Accident trends and associations are highlighted and the aims of the thesis are also stated.

1.1 THE ISSUE OF CHILD ACCIDENTS

In recent years there have been major advances in many countries in the areas of

- i) medicines and drugs, for example, antibiotics, immunizations, etc.,
- ii) medical care such as pre-natal care, post-natal-care,
- iii) medical technology and machinery generally, for example, life-support machines, incubation chambers.

As a result, the most important cause of childhood morbidity, disability, and death is now childhood injury. Hence, as living and health care standards improve and child death rates decrease so attention is drawn towards the prevention of childhood accidents which cause injury.

A longitudinal study could certainly provide a strong data-base for indentifying variables associated with childhood accidents and the 1975 Follow-up Survey of the British Birth Cohort of 1970 has been made available to us for this purpose.

The United Kingdom has a history of initiating large surveys. The first report of the British Perinatal Mortality Survey was based on a national sample of 17,000 births which occurred in one week of 1958 and an additional sample of 7,000 perinatal deaths. This report was published in 1963 and had a great impact on obstetric and paediatric practice resulting in many improvements in the organisation of the maternity services.

The second report based on the same national sample appeared in 1969. It brought forward additional information and a detailed analysis of higher risk pregnancies. The combined contribution and impact of these two publications on the levels of maternal and neonatal care led to a fresh national survey of all deliveries. This took place during one week in 1970 and was called the British Birth Cohort study of 1970. It was carried out during the week beginning 0.01 hours on April 5th 1970 in England, Wales, Scotland, and Northern Ireland

and ended at 24.00 hours on the 11th of April 1970. This was the starting point of the longitudinal study and data was collected retrospectively in 1975, 1980 and 1986 on the same cohort.

Another early study was the Thousand Family Study in which 1,142 infants born in Newcastle in 1947 were followed through their first 15 years of life. The purpose of the study was to 'demonstrate that the chief determinants of performance - biological growth, injury and disease, social adaptation and educational attainment - are so interrelated that separation, even for identification and description brings a sense of artificiality'.

The accident analysis from the Newcastle study concluded that 'during the school years the frequency of accidents diminished with increasing age as the child became more aware of the dangers in the environment' and that 'the sensible child seems to have a diminished likelihood of accident while the less intelligent child who fails to learn by experience has a greater risk of multiple accident and is more likely to come from a home with a relatively ineffective mother.' There was also an attempt to identify accidents which 'were serious or could have been serious under slightly different circumstances'.

The ultimate goal in an accident analysis must be to devise methods of accident prevention but this can only be attempted after a thorough investigation into the accident distribution and the ways in which accidents and injuries

occur. Knowledge of important social and environmental variables is also essential if high-risk sub-populations are to be identified. Extensive information in both these areas has already been collected in the 1975 Follow-up Survey and we therefore have a unique opportunity to investigate the underlying features of childhood accidents.

Before proceeding further it is clear that we should properly define the dependent reported accident outcome variable used in the analysis of our data set. The Chambers Concise 20th Century dictionary defines an accident as 'an unforeseen or unexpected event; a chance; a mishap'. The W.H.O. advisory group's definition (1957) is 'an unpremeditated event resulting in recognisable damage', a definition which is still in use today. Problems in definition have certainly arisen when using the word 'accident'. Some authors are now preferring to use the term 'unintentional injury event' since some accidents do not occur at random, such as unequal accident liabilities. Indeed, other events classified as accidents are premeditated, for example, cases of child abuse and suicides.

In our analysis we are not primarily concerned with who is directly responsible for each reported accident or whether the resulting injury is intentional or unintentional. The data set used in this analysis is the 1975 Follow-up Survey of the British Birth Cohort of 1970 and respondents were asked specifically for details of all

the child's accidents which required 'medical advice or treatment'. This was further clarified by requesting the parent to 'please include accidents in the road, home and elsewhere, accidental ingestion of medicines/poisons, burns/scalds, fractures, eye injuries, near-drowning, bad cuts and other injuries, with or without unconsciousness, and non-accidental injuries'. We therefore rely on this definition for our accident outcome variable and stress that we are dealing with 'reported' accidents which are likely to be underreported. We should also note that the need for medical advice or treatment could vary between particular groups of parents.

We shall now discuss the general history of early accident research and comment on some previous accident studies.

1.2 A HISTORY OF ACCIDENT RESEARCH

In the following sections we introduce earlier accident studies and present examples of successful interventions which resulted in the reduction of child accidents.

Since World War I greater interest has been developed in the area of industrial accident analysis, especially during World War II when the need was recognised for efficient rather than cheap labour. A healthy workforce is needed to ensure maximum operating efficiency. In an

extension of this, public and private bodies began to accumulate data on the frequency and severity of accidents occurring in everyday life.

There are some early, classic papers which must be included in any review of accident research. Work by Newbold, Greenwood, Yule, Feller, Adelstein, Farmer, Chambers, Arbous and Kerrich are well known and we will briefly discuss their contributions here.

Greenwood and Yule were pioneers in the area of accident analysis with their paper entitled 'An enquiry into the nature of frequency distribution of multiple happenings representative with particular reference to the occurrence of multiple attacks of disease or of repeated events' (1920).

They introduced the practical, statistical problem described as follows. 'Of n households exposed to risk, m_0 returned 0 cases of disease, m_1 returned each a single case, m_2 each two cases, ..., m_n returned n cases. Might such a distribution have arisen from sampling a population, each member of which was subject to constant chance of infection throughout the period of exposure, or is the form of the distribution valid evidence that particular households were especially prone to take the disease in question?'. This situation was extended to apply to repeated accidents occurring in factories.

Greenwood and Yule developed probability distributions applicable to their problem and fitted these distributions to reported data. Their first models were models of

simple, pure chance and these were extended and became more intricate as the assumptions of accident independence were altered.

In their paper they developed two forms of the contagion model. The first was the model of 'true' contagion where an event influenced the occurrence of a future event, i.e. an event might either increase or decrease the probability of the next event, and this was referred to as their 'biased' model. The second was the model of 'apparent' contagion caused by heterogeneity in the population. Greenwood and Yule called this a model of 'unequal liabilities' as the population is considered non-homogenous in its ability to sustain accidents. This was later defined as a model which could describe 'accident proneness'.

In addition, mathematical models of the 'biased' accident distribution were also introduced by Greenwood and Yule. These rely on the assumption that the occurrence of future accidents depends on an individual's previous accident history. For example, suppose we assume that workers in a factory all start off with equal liability to having an accident. Once a worker has an accident his liability changes and he is then either (i) more likely to have an accident in the future, or (ii) less likely to have an accident in the future. If we assume (i) then the accident rate of the worker will increase over time and if we assume (ii) then his accident rate will decrease over time.

Researchers believe that modelling just for (i) or (ii) is not an accurate description of what actually happens and the most likely situation would be a mixture of both assumptions which would be very difficult to model mathematically.

The concept of 'accident proneness' was first introduced when it was suspected that small groups of workers were responsible for large numbers of accidents. The term 'accident proneness' was coined in 1926 by Farmer and Chambers who found that factory workers who did well in various psycho-physical tests tended to have fewer accidents than those who did badly. They believed that accident proneness referred to personal factors that affected an individual's accident rate and they stated that 'accident proneness implies possession of those qualities which have been found from independent research to lead to an undue number of accidents.'

Chambers later added in 1941 that accident proneness 'is relatively stable in the sense that persons with a larger number of accidents than their fellows in one observational period tend to have more accidents in subsequent periods and that persons tending to sustain a number of one type of accident tend also to sustain a number of other types'.

An updated definition of accident proneness was supplied by Haddon, Suchman and Klein in 1964 who attributed proneness to 'an enduring or stable personality characteristic that predisposes an individual towards

having accidents'. Hence, accident liability can be defined as the tendency to have an accident and includes all factors, whereas accident proneness includes only personal factors.

Newbold was closely associated with Greenwood and Yule and continued their work in a paper entitled 'Practical application of the statistics of repeated events, particularly to industrial accidents', (1927). She examined data containing serious accidents, minor accidents, serious and minor illnesses and clarified a number of reporting problems. She believed that we deal with the tendency to report accidents or illnesses rather than the actual accidents or illnesses that really occur. She also found associations between the number of accidents and illnesses.

In his paper entitled 'On a general class of 'contagious' distributions', (1943), Feller brought together early research and extended Neyman's work on contagious distributions. He suggested relationships with topics such as accident statistics, telephone traffic, fire damage, sickness and life insurance, risk theory and an engineering problem.

Feller also showed that the Greenwood and Yule model of apparent contagion (their model of unequal liabilities) could be obtained from Poya-Eggenberger's (1923) model of true contagion. Hence, in the case of a good fitting model we then have the problem of deciding which underlying assumptions might be applicable to the accident data being

modelled.

In 1951 Arbous and Kerrich gave a comprehensive review of the research in accident analysis in their paper entitled 'Accident statistics and the concept of accident proneness'. Their paper contains two sections. The first is a critical evaluation of previous and current work at that time and the second section gives the mathematical background.

They believed that work on accident proneness had hardly progressed from the days of Greenwood et al. and were extremely critical about the early vagueness surrounding the concept of accident proneness. Of the accident studies that they summarised they were especially critical of studies which relied too heavily on the proness concept and they encouraged other researchers to be as critical.

Arbous and Kerrich argued that the causes of accidents could be divided into two groups. The first group is 'impersonal' where the primary cause is environmental, i.e. daylight, equipment, machinery, etc., and the second group is 'personal' where the primary cause is the individual, i.e. health, age, experience, alcohol, fatigue, boredom, etc.. They believed that in any particular accident the division between these two categories is arbitrary and that the frequency in which they both operated would vary considerably.

In their paper they summarised contemporary research to confirm the general belief amongst previous

investigators that accident proneness did exist. They also discussed the belief that accident proneness was considered a fairly stable attribute in the population in that whatever the situation, person X is always more prone than person Y although the liabilities of both will have changed by the same amount owing to the different risks in the situation itself. Arbous and Kerrich however, did not agree that accident proneness was a stable personal attribute and argued against the possibility of modelling the concept in a univariate distribution. They asked the question that if there are two people X and Y, and that during a given time period X had two accidents and Y had five accidents, then does this prove that Y is more accident prone than X? They thought not, since accidents might occur through pure chance and some studies overlooked that fact. In different time periods X might even be more accident prone than Y.

As already mentioned, previous analyses of accident data have shown some confusion when defining the accident outcome variable. For example, are we looking at accidents or the injuries resulting from accidents? It is also recognised that accidents do not always result in injury. For example, a tyre blow-out whilst motoring at 70 m.p.h. on the motorway can result in a wide range of outcomes from fatal injury to no injury at all. Similarly, accidentally spilling a pan of luke warm water over a child does not produce the same injury as accidentally spilling a pan of boiling water over the child,

although both are accidents by definition. In addition, in this particular accident the person spilling the water was responsible for the accident and yet the child was the victim who suffered the injury.

Arbous and Kerrich (1951) split the population of accidents into four non-overlapping sub-populations which we shall now discuss. The first accident sub-population consists of errors, slips, close calls and near accidents. Data for this accident sub-population is impossible to collect accurately since these events occur every day, sometimes without an individual even being aware of an occurrence. For example, consider a child who trips and falls into the road. If a car collided with the child then the result could have been fatal, yet if no collision occurred then no injury or accident was recorded.

The second accident sub-population consists of those accidents which result in minor injuries. An accurate data set containing just minor injuries would have to be collected under tightly controlled conditions. Very precise injury definitions would be necessary since the definition of a minor injury varies from individual to individual and can range from something as trivial as a broken fingernail to an injury requiring medical advice/treatment.

In the third sub-population they included accidents which result in major injuries. Here we can be sure of a reasonably accurate data set. However, it is possible that a major injury will remove an individual from further

exposure until recovery. For example, a male who accidentally breaks a rib whilst playing football is immobile for some time as the rib heals. There are fewer problems in defining major injuries and we would expect the reported major accident sub-population to be very close to the 'true' major accident sub-population.

The fourth accident sub-population contains those accidents resulting in death. Here we can be certain of accurate data because of the finality of the accident outcome. For obvious reasons however, this sub-population would not include any accident repeaters.

In a summary of these four accident sub-populations we can only be sure of collecting accurate data on fatal or major accidents and the data collected is obviously in terms of the subsequent injury rather than the accident itself.

Disease has been considered as an accidental injury and this was discussed by Waller in 1987. He used the example of a jackhammer operator who after prolonged use of the jackhammer can suffer from deafness, white finger syndrome or thromboangiitis obliterans. This can be contrasted to the situation which occurs when a jackhammer operator accidentally punches a hole in his foot whilst using a jackhammer. This injury is immediately recognised as an accidental injury but the use of a jackhammer over a long period of time results in other injury outcomes which may not be recognised until later. The hole in the foot is an immediate, acute and very visible injury whilst

deafness and white finger syndrome are accidental injuries or diseases which appear much later.

Two terms often used in the context of an accident analysis are 'accident prevention' and 'injury control'. The concept of accident prevention is to prevent the circumstances that result in injury. For example, the child-proof tablet container which prevents the access of medicines to young children. The motivation of injury control is to lessen the impact of the injury. For example, the compulsory wearing of safety helmets by motor-cycle riders which is associated with the reduction of the severity of motor-cycle accident head injuries.

In protecting the population however, both concepts require action that is sometimes seen to restrict the freedom of the individual in some way. Examples include speed restrictions, driving age limits, building regulations and restrictions, firearms restrictions, sale of fireworks restrictions, compulsory wearing of safety belts and safety helmets, drinking age limits and drinking/driving laws.

The implications of child accident prevention were examined by Langley and Silva (1982) in a survey of mothers whose children had been accidentally injured. It was seen that many of the mothers had not taken any preventive action to stop a recurrence of the event which resulted in the injury. This was not because the event was difficult to control or that action was inappropriate or even that the event was a useful education for the

child. The results indicated that the mothers viewed each accident by strict definition as an 'accident' which could not be prevented.

This has implications for our analysis since accident repeaters may be incorrectly categorised. For example, children of mothers who do take preventive action might have lower accident rates than children of mothers who do not take preventive action. This means that a child's accident rate may be dependent on attributes of the mother's character.

Previous research has uncovered a number of problems with accident analysis. The concepts of proneness and prevention are obviously key factors in the number of accidents that occur. However, in many instances it is impossible to isolate individual factors which influence accident rates, especially since in different time periods an individual's proneness may change dramatically. We now examine the concept of accident proneness in more detail.

1.2.1 THE ROLE OF PRONESS

In this section we discuss the concept of 'accident proneness' and highlight the problems of identifying an accident-prone sub-population.

From its first introduction there has been confusion in the use of the term 'accident proneness'. The difficulty is in deciding whether it is a general variable

or is more specific and applies only to certain people in certain circumstances. It is generally agreed that accident proneness refers to personal rather than impersonal attributes. The main problems are in deciding whether or not the accident prone population is a stable part of the population and whether or not it can be identified.

It is quite feasible that the age and/or experience of an individual is directly related to their own level of accident proneness. For example, a person may have an accident which in slightly altered circumstances may have been fatal. In a similar, repeated situation this person has three alternative paths of action to take, bearing in mind their previous accident experience.

Firstly, he may be more cautious and wary in the repeated situation in an effort to avoid an accident. In this way he avoids the possibility of a serious and possibly fatal injury because he has learnt from his earlier experience. Hence, this can be considered as a positive learning effect.

Secondly, he may be more on edge or nervous as a result of his previous 'close call' and therefore put himself at a higher risk of having an accident purely because he is so anxious and jumpy. This can be considered as a negative learning effect.

Thirdly, he may have learnt or remembered nothing from the previous accident and act as if the previous 'close call' had never occurred. This can be considered as a

non-learning effect and each accident would be considered completely independent of the previous accident history.

Bearing in mind these three alternatives we consider the 'true life' situation of a saw-mill worker who returns to work after recovering from an accident in which he had a finger removed by the electric log-saw that he operates. When returning to work he may be more careful when operating the electric log-saw in fear of losing another finger, or alternatively, might act as if the accident never occurred, or could be so nervous when operating the machine that another finger is lost.

The assumption of independence between observations as required by many accident distribution models may not be a realistic one. Over/underdispersion relative to the binomial model was investigated by Rudolfer (1990) who developed a Markov chain model of extrabinomial variation. His alternative model adopted an explicit description of the underlying stochastic mechanism generating the data and (using maximum likelihood) his results compared favourably with other models.

In a paper on accidents in the railway industry, Adelstein (1952) presented an extensive analysis of 1,452 accidents of railway shunters. In the discussion of the paper J.O. Irwin commented that it was 'the most comprehensive treatment of a body of actual data by statistical methods which have become associated with accident proneness since Miss Newbold's classical work presented to the society in 1927'. Adelstein highlighted

the problems of accident classification and accident proneness and searched for some evidence of proneness by calculating the coefficients of correlation between accidents sustained by individuals during two consecutive time periods. This type of approach was recommended when searching for evidence of proneness by Kerrich and Arbous (1951), and Feller (1943).

In Adelstein's data set only 122 railway shunters had been shunters in excess of 11 years so he took these individuals as his sample. The 122 men had each been in service from zero to 25 years before the start of the time periods in the study. He took the first time period to be the first six years of service and the remaining five years the second period. Adelstein found that a significant correlation of 0.258 did exist between the number of accidents in the two time periods.

His next step was to split the 11 years into four consecutive time periods and see if correlation still existed between consecutive periods. The first period was for three years, the second period the next three years, the third period the next 30 months and the final period the final 30 months. The correlation coefficients for each pair of consecutive time periods were 0.109, 0.048 and 0.205 with significant correlation between the third and fourth periods only.

He concluded that the distribution for the 11 year sample was of the Negative Binomial type and that there was indeed some correlation between the first six year

period and the second five year period. He believed that the individual accident rates in the group differed a great deal but were so clouded by chance events that it took many years before a reliable measure of the differences could be sought. After many years of service the shunters have a very stable accident rate so the proneness should show up in the accident rates.

A similar approach was used for the problem of sickness. Adelstein took minor sicknesses as those lasting for less than ten days and classified major sicknesses as those lasting for ten days or more. Using the same 122 men as before in the 11 year analysis of correlation with two consecutive time periods, he found a correlation of 0.661 for minor sicknesses, 0.507 for major sickness and 0.527 for major sickness against minor sickness, all of which are significant. He also rightly believed that he was measuring in part the tendency to report sick rather than the tendency to be sick.

Adelstein's data were quite extensive and he was able to use the same techniques to search for correlation in accidents resulting in damage to property and also in home injuries. In both cases he found good evidence of correlation although he did not rely too heavily on the data for home injuries since he suspected they were not accurate.

Adelstein found no evidence of correlation between industrial and home injuries, or between minor and major injuries (defined according to the length of time taken

off work for the injury), and from this he concluded that accident proneness was not a stable attribute. He also searched for evidence that men tended to repeat the same types of accidents but was unsuccessful in this search.

Adelstein also used another method of searching for proneness and this was to identify men with high accident rates and then remove them from the next year's analysis. However, he found that by removing these supposedly prone men he did not dramatically alter the next year's accident rates.

In the shunters case the problem of identifying proneness seems practically impossible since prone shunters may well discover the dangers to them at an early stage, hence their proneness does not manifest itself in the long service analysis. It is also feasible that a disabling accident may put a shunter out of the data set and he too will be prevented from allowing his proneness to show in any further analysis. Adelstein does not take into account that the fear of disciplinary action may also have had some effect on accident rates as those with high accident rates may be disciplined or alternatively transferred to other duties. Through his own experience a shunter may become more or less accident prone but other factors would also undoubtably contribute to the proneness.

Adelstein had difficulty in interpreting the effects of age and experience on the accident rates of the shunters. He decided that for his particular data set both

age and experience played an important initial role in the first years of a shunter's career but then rapidly lost their effects. For the railway shunters he found that the accident rate for all age groups tended to stabilise after two years of service.

The question of major importance here is whether the proneness factor was significant or not? Adelstein thought not, and believed that proneness was not a general factor but operated in a specific type of event in a specific environment and that it may alter in an individual in accordance with his changes of activity. He believed that accident proneness existed but was not a stable attribute in the population.

Using the methodology of Adelstein as a guide, Mohr and Clemmer (1988) searched for evidence of proneness in a study of workers in heavy industry. They had access to the records of a longitudinal study of offshore petroleum workers in the Gulf of Mexico and examined the number of injuries in consecutive time periods. The aim of the study was to answer the following three questions.

- 1) Was the proportion of workers with excess injuries in successive time periods greater than could be expected by chance?
- 2) What proportion of injuries could be attributed to these workers?
- 3) What effect would their removal and replacement at the end of one time period have on the number of injuries occurring during the next time period?

Their analysis was based on medically attended injuries over a two year period and the data were

controlled for job hazards, exposure, age and changes in job and location. From a data base of 1,023 workers, 610 had sufficient length of service to be included in the final analysis and, over the two year period, 411 medically attended injuries were recorded.

Mohr and Clemmer (1988) concluded that accident repeaters were not a stable part of the workforce and that the removal of workers with excessive injuries in a given time period would not appreciatively decrease the number of injuries in succeeding time periods. They added that there was a 'break in' period for new workers during which time there was a high risk of injury. They also concluded that replacing accident repeaters was simply not an economic policy after taking into account the costs of administration, retraining, location, etc..

Another approach to the problem of modelling accidents was suggested by Cresswell and Froggatt (1963) in their book on the causation of bus driver accidents. They derived a 'spells' model as a three parameter convolution of a Poisson distribution and a Neyman Type A distribution. This was not a model of proneness and they made the following assumptions.

- 1) Every driver is liable to 'spells' - time periods during which his performance is sub-standard so that he is liable to incur accidents. The number of spells in a given time period is assumed to be a Poisson variable with parameter L . This implies that the number of spells suffered by a driver in any one period is independent of the number he suffers in any other period, and also that the length of a spell is short relative to the period of time under consideration.

- 2) All drivers are equally liable to the occurrence of a spell.
- 3) The probability of an accident occurring within a spell is constant and not dependent on the particular driver. Accidents within spells are therefore assumed to have a Poisson distribution with a constant parameter.
- 4) No accidents can occur outside spells.

The results, however, suggested that there were severe limitations on the use of the distribution to model reported accident data.

The concept of proneness is not restricted to accident data and in the paper entitled 'the distribution of Doctor-patient contacts in the National Health Service', Ashford and Hunt (1974) developed a trivariate Poisson/Gamma/modified log distribution to model actual illness data.

The basic concept of the model is that an individual is subject to episodes of ill-health which start at defined points in time and generate varying numbers of different types of contact with the National Health Service. The trivariate Poisson/Gamma/modified log distribution was obtained by assuming that episodes of illness for each individual in a population of patients are events in a Poisson process, the rate parameter varying within the population according to some specified distribution, in their case a Gamma distribution. The number of N.H.S. contacts of a given type generated from the illness episodes are then independent random variables following modified log distributions. Hence, from the

first contact with a G.P. a patient could be referred to a hospital for further tests, for example, x-rays or surgery, or no further action may be necessary.

The data they used came from the population of Exeter and comprised 70,000 registered patients of the G.P.'s taking part during one calendar year. The data were split according to age and sex and good models were obtained for the males. However, the results for the females were not as encouraging, especially for very young or very old females.

In response to the paper it was generally agreed that more should be known about the heterogeneous nature of the data. It was observed that there were a large number of 'no-contact' patients and it was also thought that some patients made routine visits to their G.P., say during pregnancy or for the purpose of illness prevention. For example, every spring for hay-fever medication, or every autumn for flu immunization.

Data from the National Health Services were also used by Gillings (1974) in his paper entitled 'Some further results for bivariate generalisations for the Neyman Type A Distribution'. He investigated the joint usage of two types of Health Services simultaneously, using bivariate generalisations of the Neyman Type A distribution. However, his models did not provide an adequate fit to the data.

If we accept that accident proneness does exist there are some basic but important questions that must be asked

about the accident-prone individual especially when searching for accident proneness in workers in high-risk occupations. It would seem there are two sub-populations of accident-prone people; those who recognise that they are accident-prone and those who do not. It seems unlikely that those aware of their proneness would actively seek dangerous employment. However, as mentioned earlier there are three courses of action to take after experiencing an accident, namely, positive, negative and non-learning and this process can also be applied in a wider sense with respect to employment.

We must ask whether an accident prone person who recognises their own proneness would search for an appropriate occupation. One must consider whether an accident-prone person would actually apply for, or indeed stay, in a high risk occupation once they discover that they are indeed at high risk. A further problem arises when considering those who have not yet recognised their proneness and we must ask if they can be identified among the other accident repeaters or would this group be mixed with other groups.

There are also other problems of classification. Using Arbous and Kerrich's breakdown of the accident population we shall consider the accident history in a particular time period of two people, X and Y. Suppose person X has a series of accidents in the time period and, in chronological order they are as follows; a minor accident, a minor accident, a near miss, a minor accident,

a major accident, a near miss, and a minor accident. Now suppose that during the same time period person Y has a series of accidents as follows; a major accident, a minor accident, a major accident, a major accident, and a near miss. During the same time period X has only one major accident yet six lesser accidents, and Y has three major accidents and two lesser accidents.

The question we must therefore ask is whether we can confidently identify which of these two people is the more accident-prone. The general consensus of opinion is that accident proneness does exist, i.e. personal attributes are at work in influencing an individual's accident rate, but that the accident prone population is not a stable part of the population and as such cannot be properly identified.

We now give a critical discussion of two accident studies. The methodology of the analysis of the child accident data adapts some of the methodology used in these two studies.

1.2.2 THE METHODOLOGY OF TWO ACCIDENT STUDIES

In this section we discuss the methodology of two particular accident studies, both of which identify high risk sub-populations by predicting the distribution of future events based on past accident information. The second study also highlights high risk roundabouts where

accidents are most likely to occur in an attempt to influence the design of future roundabouts in order to reduce accident rates.

Our first detailed example of an accident study which attempted to identify high-risk accident groups is a study which used motor vehicle accident data. Weber (1971) predicted the distribution of future accidents based on the number of previous reported accidents and explanatory variables. He obtained a sample of 148,000 drivers in California from the 1964 California Driver Record Study and fitted a Negative Binomial probability distribution to the reported accident frequencies for 1963. The Negative Binomial model provided a good fit and he explored further by splitting the 148,000 individuals into 2,880 groups according to sex, marital status, age, area of residence, conviction history and reported accident history. He fitted a separate Poisson distribution to each of the 193 groups that contained 100 or more individuals and in 167 (87%) of these groups the fitted Poisson model was acceptable at the five percent significance level. He therefore concluded that the six criteria did an adequate job of classifying the individuals into Poisson groups.

Weber assumed that a driver's reported accident record was a Poisson process with a parameter L which is a linear function of k criteria variables and used an iterative weighted least-squares procedure to solve the system of maximum likelihood equations for estimating L . He obtained the following model for male drivers.

$$\hat{L} = 0.00274 + 0.00909 \times X1 + 0.0532 \times X2 + 0.0223 \times X3 \\ + 0.02160 \times X4 + 0.0169 \times X5,$$

where,

X1 = the natural logarithm of the traffic density index of the county in which the driver resided.

X2 = 5 / (age - 13).

X3 = the number of countable convictions incurred during the years 1961-1962.

X4 = the number of accident involvements incurred during the years 1961-1962.

X5 = the number of non-countable convictions incurred during the years 1961-1962.

Examining each variable in turn we see that the effect of age is such that as age increases so the estimate of L decreases. Hence, the younger the driver the higher the estimate of L. The inverse age relationship seen in the variable X2 is more apparent when we assume that the Californian driving age in 1963 was 18 (a point not clear in the paper). If we set the variables X3, X4 and X5 all to zero, then an increase in either of X3, X4 or X5 will have the effect of increasing the estimate of L. Countable/non-countable convictions were not defined by Weber but it is assumed they are driving offences which may/may not count towards endorsements on the driver's license. Table 1 shows some estimates of L and the standard deviation for particular male drivers.

We can identify the highest risk group as being young drivers who reside in a county with a high traffic density index and have a history of accident involvement and countable/non-countable convictions. The lowest risk

group consists of older drivers who reside in a county with a low traffic density index and have no previous history of accidents or countable/non-countable convictions.

TABLE 1. Estimates of L and the standard deviation for individual male drivers.

Traffic density	Age	Accident involvement	Convictions		L	var L
			non-count.	count.		
10	60	0	0	0	0.0293	0.0023
50	60	0	1	1	0.0832	0.0027
150	60	1	1	1	0.1147	0.0034
10	40	0	2	0	0.0781	0.0032
50	40	1	0	0	0.0698	0.0027
150	40	0	0	0	0.0581	0.0011
10	20	1	1	0	0.1056	0.0045
50	20	0	0	0	0.0763	0.0035
150	20	2	3	1	0.2133	0.0059

Although Weber had a means of estimating L for an individual driver he was restricted in the accuracy of his estimate by the variables in the data set. It is not clear whether other variables were available to him. It would seem intuitive that the more time a driver spends at the wheel the more he increases his accident risk exposure. For example, a motorist who drives 1,000 miles per year would have less exposure than a motorist who drives 10,000 miles a year. Hence, the risk of having an accident would be directly related to the number of hours spent at the wheel - a variable not included in the model.

Weber took into account age but not experience. Age and experience have been found to be closely linked in other accident studies such as Adelstein's (1952) study,

and clearly it is difficult to separate these two variables. In assessing driving skills it is likely that experience is an important variable and one way of measuring this would be to take driving experience as the time from passing the driving test to the time of the survey. Clearly this would also be linked with age and the number of hours at the wheel but it is very difficult to isolate or distinguish between these types of influences. Nevertheless, an investigation would have made the study more useful.

Our second example is more recent and uses accident data obtained at four-arm roundabouts in the U.K.. Maycock and Hall (1984) investigated accidents which occurred at particular types of roundabouts and identified influential design variables. Their aim was to derive relationships between roundabout accident frequency, traffic flow and geometry which could be used to optimise designs and for predicting mean accident frequencies (personal-injury accidents per year) for use in junction appraisal.

The data set comprised personal-injury accidents which had occurred within 20 metres of particular target roundabouts selected from a larger group of roundabouts. The listing of roundabouts were compiled from several sources; official listings of all reported injury accidents at junctions on Class A roads for 1974-1976, information from local authorities, inspection of 1:50,000 Ordnance Survey maps and Report LR 774 from the Transport and Road Research Laboratory (1977). The final selection

of target roundabouts was thought to be representative of four-arm roundabouts throughout England and Wales.

In the first stage of the analysis the roundabouts were separated into groups according to the size of the roundabout (small, conventional and dual-carriageway), and speed-limit (30-40 mph. and 50-70 mph.). As expected, different accident rates were observed for the different types of roundabouts. The comparisons of accident frequencies suggested that smaller roundabouts (30-40 mph.) generated more accidents on average while the conventional (30-40 mph.) were generating rather fewer. This can be seen in Table 2.

TABLE 2. Accident frequencies by roundabout category with the standard error of the mean values (calculated from the site to site variation) in brackets.

Roundabout category	Number of sites	Accident frequency (per junction year)
Small:		
30-40 mph.	25	4.38 (0.48)
50-70 mph.	11	2.83 (0.65)
Total	36	3.89 (0.40)
Conventional:		
30-40 mph.	11	2.36 (0.20)
50-70 mph.	11	3.10 (1.01)
Dual-carriageway:		
30-40 mph.	14	3.37 (0.57)
50-70 mph.	12	2.88 (0.67)
Total	48	2.94 (0.33)
All categories	84	3.31 (0.26)

With the use of total vehicle inflow as a measure of exposure to risk they brought down the conventional (30-40 mph.) roundabout accident frequency to be similar to the dual-carriageway roundabouts accident frequency, although the small (30-40 mph.) roundabout accident frequency remained notably higher than that of the conventional designs. This can be seen in Table 3.

TABLE 3. Accident rates taking into account traffic flows with the standard errors of the mean values calculated from the site to site variation in brackets.

Roundabout category	Number of sites	Accident frequency (per 100 million vehicles)
Small:		
30-40 mph.	25	37.1 (3.1)
50-70 mph.	11	28.7 (3.3)
Total	36	34.8 (2.5)
Conventional:		
30-40 mph.	11	21.2 (2.0)
50-70 mph.	11	28.7 (6.3)
Dual-carriageway:		
30-40 mph.	14	22.5 (2.5)
50-70 mph.	12	22.4 (3.1)
Total	48	23.5 (1.8)
All categories	84	27.5 (1.6)

Maycock and Hall then investigated accident types and road-user involvement. Accident types were: entering-circulating accidents, approaching accidents, single vehicle accidents, other accidents and pedestrian accidents.

Road user categories were: pedal cycle accidents,

motor cycle accidents, motor car accidents, LGV accidents, HGV accidents, bus/coach accidents and pedestrian accidents.

The categories were not mutually exclusive since a collision between a car and a motor-cycle would be attributed to both categories. The main feature of their findings in this area was the very high accident involvement rates of two-wheeled vehicles which were 10-15 times that of cars.

Their regression analysis relating the accident frequency (the average number of accidents per year) to the range of explanatory variables was in two parts. Their aim was to determine a relationship which could be used to predict site-specific mean accident frequencies from such variables. Multiple regression techniques were used and the explanatory variables included traffic flow and a range of geometric variables.

The first part of the analysis consisted of relating accidents to roundabouts as a whole and the second part consisted of analysing arm-specific accidents with respect to the type of accident as previously defined.

The first model investigated was a simple linear model of the form:

$$A = B_0 + B_1 \times V,$$

where A = average accident frequency (accidents per year) at individual roundabouts,

V = annual average 24-hour vehicle inflows (in thousands of vehicles),

and B₀, B₁ are coefficients to be estimated.

Significant relationships were found in all groups where full data was available. They progressed to the next model which was constrained to pass through the origin. This was of the form.

$$A = k \times (Q)^a,$$

where A = accidents per year as before, Q = the flow function, an algebraic combination of the various elements of flow within the junction (all 24-hour averages expressed in thousands of vehicles), and k, a are parameters to be estimated.

Three basic flow functions(Q) were tried and were defined as follows:

Total inflow: the sum of the four entering flows.

Cross product: the product of the total entering flows on one pair of opposite arms with the total entering flows on the other pair of opposite arms.

Entering-circulating: the product of the entering and circulating flows at each entry summed over the four arms.

With the exception of the conventional roundabouts (30-40 mph.) all categories yielded highly significant relationships (one per cent or better) relating total accidents to flow for all flow functions. There was no significant difference in the exponent (a) between any of the roundabout categories so a common value was fitted and the variation in the data was reflected in the different k values. This can be seen in Table 4. The most appropriate flow function was considered to be cross product flow.

TABLE 4. Parameters in the relationship $A=kxQ^a$ for 1974-1979 accident frequencies by roundabout category (Q is the cross-product flow function).

Roundabout category	Number of sites	k	a
Small:			0.68
30-40 mph.	25	0.101	
50-70 mph.	11	0.081	
Total	36	0.095	
Conventional:			
30-40 mph.	11	0.057	
50-70 mph.	11	0.080	
Dual-carriageway:			
30-40 mph.	14	0.057	
50-70 mph.	12	0.061	
Total	48	0.062	

The second part of their analysis focussed on arm-specific accident analysis. Since it was expected that different types of accidents might be correlated with different combinations of flow, arm-specific accidents were disaggregated into the five accident type categories defined earlier. For each subset equations of the following form;

$$A = k \times (Q)^a, \quad \text{or} \quad A = k \times (Q1)^a \times (Q2)^b,$$

were fitted where,

A = accidents per year per arm,
Q, Q1, Q2 = flow functions and
k, a, b are parameters to be determined.

In order to take into account the geometric variables the model was extended to,

$$A = k \times (Q)^a \times [\exp (b_1 \times G_1 + b_2 \times G_2 + \dots)]$$

where,

A = accidents per year per arm,
 Q = relevant flow function,
 G_i = geometric variables (i=1,2,3,...), and
 k, a, b_i (i=1,2,3,...), are parameters to be estimated.

Influential geometric variables were identified from this analysis based on a criterion of plausibility, design usefulness and statistical significance.

Maycock and Hall made use of the statistical packages GENSTAT and GLIM for their analyses. The majority of the regression models were fitted using a Poisson error structure but they also considered the Negative Binomial error. In terms of the roundabout accidents they assumed that there were two components to be considered: firstly the within roundabout-site error (regarded as Poisson), and secondly an unspecified between site error distribution. By assuming that the probability density function for the distribution of mean accident frequencies between sites is of the Gamma form, they were able to produce a Negative Binomial type distribution for the combined distribution.

In summary, both these studies attempted to predict future accidents using a regression model based on previous reported accident data. Weber identified high risk drivers whilst Maycock and Hall identified high risk vehicles, i.e. motor-bikes, and also design faults associated with high risk roundabouts.

We now present examples identifying high risk groups in the child population.

1.2.3 CHILD ACCIDENTS

In this section we introduce examples of high risk child accident groups. Bijur et al (1983) produced a brief history of early 'proneness' research in child accident analysis in their paper titled 'Child behaviour and accidental injury in 11,966 preschool children'. They traced research back to the 1930's when the associations of childhood accidents and behaviour were first investigated by psychoanalysts.

Until the 1960's most research was focussed on the identification of a single accident proneness trait but since then investigators have broadened their search and now explore a variety of behavioural and environmental variables. Previous research indicated that multiple accidents were associated with increasing levels of behavioural problems.

In their own research Bijur et al (1983) had access to the 1975 Follow-up Survey of the British Birth Cohort of 1970 and searched for associations between behavioural variables and different types of accidents. They found associations between aggressive behaviour and the following three accident outcome variables; hospitalised accidents, non-hospitalised accidents, and accidental

ingestions. They also found an association between overactive behaviour and non-hospitalised accidents.

In line with previous research they believed that accidents were associated with increasing levels of behavioural problems and that 'the identification of high risk groups can be used to direct education and to provide a safer physical environment to families in which there are children with high levels of high risk behaviours and which share other risk factors'. This suggests that underlying accident trends can be uncovered and that interventions can be made and directed to those groups in most need.

There are many instances where the identification of a high risk sub-population has resulted in successful preventive action and Rivera (1982) presented a number of these in a review of research into the epidemiology of childhood injuries. When considering specific types of accidents, ingestions say, we need only recall how the introduction of the childproof tablet container successfully reduced the number of young children admitted to hospitals with medicine ingestions (Clarke and Walton, 1979). In this particular instance the high-risk accident group was identified as young children and preventive action was formulated to target that group.

Preventive action is not always successful however, and the first step in the campaign to reduce childhood ingestions took the form of a media campaign. The aim of the campaign was to re-educate both parents and children

of the dangers involved when leaving medicines in easily accessible places. This particular campaign was not successful and prompted the introduction of the child-proof tablet container which was more successful.

Another example of preventive action being successful is the introduction of flame resistant clothing which reduced the number of severe burns in young children (McLoughlin, Clarke, Stahl et al, 1977, and Blum and Ames, 1977). Children are often unaware of the dangers of being too close to a fire, especially when wearing clothing that is not flame resistant. Hence inflammable materials were impregnated with flame resistant chemicals designed to reduce the risk of fire. Unfortunately it was also found that some chemicals used in the production were carcinogenic and some types of clothing had to be withdrawn.

Consider also the introduction of thermostatically controlled water heaters in the home which were very important in reducing the severity of scalds in young children (Fieldman, Schaller, Feldman et al, 1978). A child's skin is much more tender and sensitive than an adults skin, hence a child who turns on the hot water tap thinking it is the cold water tap could easily scald themselves if the water is too hot.

In terms of both accident prevention and injury control defined earlier, the introduction of the child car-safety harness succeeded in reducing both the severity and the number of injuries in motor vehicle accidents

(Williams and Zandor, 1977). It is now the law that any child under five years old in the front seats of a car must be restrained in a car-safety harness or chair. This is a sensible precaution since a motorist is unable to devote full attention or supervision to an unrestrained child in his car. In a dangerous situation the driver could well be distracted by the movements and antics of the child and if the driver were to brake sharply then the child could be severely injured.

Great value is placed upon aggressive competitiveness in sport and a child's risk exposure to accidents and injuries is certainly increased when involved with body contact sports (Waller and Klein, 1973). This is especially the case where aggressive and violent behaviour together with the encouragement of risk-taking has become paramount in educating children in the 'who dares win' attitude to sport.

The impact of television is yet to be fully understood in relation to child accidents but it is inevitable that some people, children especially, will try to emulate their favourite television characters (Danen, O'Connor and Briggs, 1976, and Levin, McLain and Blair, 1976). For example, some children may feel tempted to copy the antics of particular cartoon characters believing mistakenly that they will not be injured. By doing so they increase their own risk of having an accident by putting themselves into situations where accidents are likely to occur. Cartoon characters are often involved in death-defying, violent

and unreal situations which would result in fatal consequences if they were human and not 'make-believe'.

In the U.K. television stations show programmes which are not considered suitable for children at times when they are expected to be asleep. In addition, it should be recognised that television and film censorship is also imposed on the adult population with respect to excessive and unnecessary violence, sex-scenes, etc..

The accuracy of the perception of risk should also be taken into account since some accidents occur because of a mis-match between the levels of skill demanded by a particular situation and those actually possessed by a child. For example, a parent is quite likely to present a child with an oversized bicycle so that the child can grow into it, just like a pair of shoes or a coat. The fact that the bicycle is too big for the child and will lead to a lack of control on the child's part is not recognised by the parent. In this way the parents have underestimated the danger to their child.

This particular concept can also be extended to the whole population. For example, on Britain's motorways there are many average drivers who willingly and intentionally take dangerous and sometimes fatal risks because they lack the perception to realise the danger to themselves and other road users.

With child accidents specifically in mind we can also try to reduce accidental injuries by supplying more supervision. An example of this is the case of the

lifeguard being present at a swimming pool or beach (Metropol Life Insurance Company, 1978). The life-guard's purpose is to prevent dangerous activity and also be available as a source of immediate basic medical aid.

Unsupervised children also have great trouble when crossing the road (Sandels, 1970). Read et al (1963) showed that intervention in the form of crossing guards was a very successful form of prevention since no crossing accidents occurred in their study when school safety patrols were on duty.

Many child accidents can occur in the playground area where falls, cuts and abrasions are often caused by faulty or inadequate equipment, or more simply the type of flooring in the play area (Rutherford, 1979). The traditional concrete or hard-surface playing area can cause serious injury to young children who fall upon it. These types of surfaces are out of date and modern playgrounds are now usually surfaced with woodchips or some type of matting to reduce the severity of injuries. Similarly, modern playground equipment is also designed to minimize injury.

Young children should be taught proper behaviour as quickly as possible and it is the responsibility of the parents to educate a young child to the dangers surrounding them in the home. For example, in the child's home the kitchen should be considered as a very dangerous place for a child. An injury can result from many of the items in a kitchen such as burners on the stove, teapots,

saucepans and kettles containing boiling water, knives, forks and other utensils, bottles, glasses and other crockery, washing up liquid, bleach and cleaning solvents, etc.. A child should also be discouraged from swallowing anything other than food or drink as children have been known to swallow needles, buttons, hairgrips, marbles and even small toys.

Clearly the identification of high risk sub-populations would be a good first step in understanding the nature of accidents and more knowledge of the underlying accident distribution is needed.

We now discuss the data set available for our analysis.

1.3 THE DATA SET AVAILABLE FOR THE THESIS

In the following sections we describe the data used in the analysis and summarise the main features. In this preliminary, exploratory analysis the reliability of the data is assessed and reported accident frequency trends and associations are identified. We also search for any inadequacies or deficiencies in the data which may affect the later analysis.

1.3.1 THE BRITISH BIRTH COHORT SURVEY OF 1975

The British Birth Cohort Survey of 1970 was the first point of a longitudinal study for which the analysis is still proceeding. During the first week of April 1970, all births in the U.K. were recorded and this was the starting point of the study. Retrospective follow-up surveys were conducted in 1975, 1980 and 1986, and each retrospective follow up survey was refined and improved upon where necessary. The British Birth Cohort data set used in this project is the 1975 British Birth Cohort retrospective follow-up study, hence all the children in the study were five years old when the data was collected. Data for each child was collected retrospectively from the following sources.

- A) A Developmental History Schedule completed by the Health Visitor. This was an assessment of the child's physical development based on all the child's health records available at that time, i.e. Child Health Clinic records, Health Visitor records, Health Centre records, G.P. records and Child Welfare records.
- B) A Maternal Self-completion Questionnaire completed by the child's mother or mother figure. This questionnaire contained questions related to the mother's health and social attitudes. It also contained questions concerning the child's behaviour, personality traits and habits. It should be noted that the mother's assessment of her child's personality and her own health may be biased.
- C) A Home Interview Questionnaire completed by the Health Visitor whilst interviewing the family at the child's home. The Health Visitor was also required to make assessments of the household and neighbourhood. It should be noted that the Health Visitor's assessment of the household and neighbourhood may be biased.

D) A Child Test Booklet comprised of intelligence tests completed by the child. This was completed in the child's home with the Health Visitor supervising.

The data set made available for our analysis consists of 13,135 surviving children and 369 variables. In particular, extensive detail is available for each of the reported accidents that each child experienced during the first five years of life.

It is expected that the definitions of an accident and the seriousness of an accident will vary between groups of parents and this could result in accident misreporting. Since the data was collected retrospectively at the end of the five year period it also seems likely there would be an underreporting of minor accidents which were simply forgotten. In addition, it is also suspected that accident recall over the five year period would be more accurate over the later part of the period than for the earlier part.

The accident information used in our analysis was obtained in the Home Interview Questionnaire which asked for details of all the child's accidents upto the age of five years old which required 'medical advice or treatment'. As mentioned in Chapter 1.1 this was further clarified by requesting the parent to 'please include accidents in the road, home and elsewhere, accidental ingestion of medicines/poisons, burns/scalds, fractures, eye injuries, near-drowning, bad cuts and other injuries, with or without unconsciousness, and non-accidental injuries'. For each reported accident the following

information was obtained.

- 1) The age in months of the child when the accident occurred.
- 2) Where the accident occurred;
 - i) in the child's home,
 - ii) in someone else's home,
 - iii) on the road, e.g. a road traffic accident or an accident on the road which did not involve a vehicle,
 - iv) in a play area, such as a playground or playgroup,
 - v) in a place which not included in any of the above categories, e.g. a shop or is unknown.
- 3) A description of the injury incurred;
 - i) a bone fracture,
 - ii) a dislocation or a sprain,
 - iii) lacerations which required stitches,
 - iv) a superficial injury such as bruising, a minor cut, graze, etc.,
 - v) an ingestion such as swallowing bleach, objects, etc., or a foreign body injury such as glass in an eye, an object stuck in an ear, etc.,
 - vi) a burn or a scald,
 - vii) an injury which is not included in any of the above categories or is unidentified.
- 4) Where the injury was treated;
 - i) a hospital visit which resulted in a stay of one night or more,
 - ii) a casualty department,
 - iii) at the General Practitioners,
 - iv) at a place not included in any of the above categories or is unknown.
- 5) The treatment that was given;
 - i) the administration of a bandage, plaster or splint,
 - ii) an x-ray,
 - iii) observation,
 - iv) a stomach pump,
 - v) some medicine,
 - vi) stitches,
 - vii) surgery,
 - viii) not specified or no treatment was given.

- 6) The part of the child's body that was injured;
- i) the head,
 - ii) the arms,
 - iii) the legs,
 - iv) the torso,
 - v) a combination of the above categories or the body part injured is unknown.

It should also be noted that during the coding of the data the reported accidents which resulted in both a bone fracture and a laceration requiring stitches (see 3 above) were always coded as a bone fracture, i.e. the most serious injury was recorded.

The remaining variables available to us included such information as the age of each parent in 1970, the social class of the family based on the father's occupation, the number of household moves during 1970-1975, the parents' ethnic group, the home environment, variables relating to the mother's health, characteristics of the child's personality, and so on.

We now discuss the reliability of the data.

1.3.2 THE RELIABILITY OF THE DATA AND RESULTS

In this section we assess the reliability of the data that is available to us. In particular we discuss the dependent accident outcome variable since we suspect that certain types of accidents are recalled more readily than others. For example, an accident which resulted in a broken arm encased in plaster for six weeks is more likely

to be remembered in retrospect than an accident which resulted in a sprained ankle.

It often occurs in accident studies that the data are based on the retrospective reporting of accidents rather than on the accidents themselves and it has been well established in earlier literature that the word 'accident' means different things to different people. Accidents resulting in personal injury are more easily recalled than those resulting in damage to property, objects or other people, hence the need to ascertain the reliability of the data set.

Taking hospitalised accidents as the dependent accident outcome variable, i.e. hospital admissions as a result of an accident, it must be noted that the perception of the need for medical attention varies with variables such as social class, mothers qualification's, the size of the family, the mother's age, etc.. This was shown by Stewart-Brown, Golding, Bijur and Butler (1983) who postulated that there is an underreporting of minor injuries and also of injuries occurring in families where stressful life events are common-place. Accidents which are considered serious enough to warrant a hospital admission are more likely to be remembered than accidents which resulted in minor injury. The timing of an accident is also important and the time of an accident occurring in the last year is more likely to be recalled accurately than an accident which occurred four years ago.

A specific example of a recall problem is shown by

While (1985) in his Msc. dissertation on a small sample of the British Birth Cohort retrospective follow-up study of 1975. He found that a rounding method was used to record the time of a child's accident. This resulted in a 'heaping' of reported accident times at six monthly intervals, i.e. at six months, 12 months, 18 months, 24 months, etc..

More generally, we can express the probability, $g(t)$, of an accident being reported in a particular month t ($0 < t < 60$) in terms of the probability, $f(t)$, of it occurring in month t as follows:

$$g(t) = \sum_{k=0}^{t-1} f(t-k) \times a(t-k, t) + \sum_{k=1}^{60-t} f(t+k) \times a(t, t+k),$$

where $a(i, j)$ = the conditional probability that an accident occurring in month i is reported in month j .

Hence, for any particular t , both $f(t)$ and $a(i, j)$ are unknown. This phenomenon obviously raises major problems in any analysis which relies on the accident times. In particular in Section 4.3 we search for evidence of accident proneness by calculating the coefficients of correlation between the number of accidents occurring during two consecutive time periods. The unreliability of the reported accident times becomes apparent in this analysis and casts doubt on any conclusions based on the misreported accident times.

If we were to conclude that a rounding procedure is not used, this would imply that children are at

particularly high risk of having accidents at six monthly intervals and at low risk for the rest of the year. One way of overcoming the problem of inaccurately reported accident times would be to refer directly to medical records, or, in the future, to use a diary type of survey. A continuous diary-style survey procedure in which accidents were recorded on a weekly or monthly basis may help to reduce this particular event recall problem.

A similar type of age recall problem was investigated by Caldwell (1966) in a study of age misstatement among young children in Ghana. Using the 1960 Ghana Population Census Caldwell attempted to enumerate the population of Ghana with respect to age in single years and it became evident when analysing the age of the population that there was an 'excess' of three year olds and a 'shortage' of one year old children. An identical phenomenon had been seen in age data collected from other countries such as Malasia and Pakistan, and Caldwell was able to investigate the phenomenon using the 1960 Population Census of Ghana as a data base.

To find out whether there was any age misreporting Caldwell took a sample of 1,000 children aged 0-99 months old at the time of the survey and traced them in the guise of the Census. Only 608 of these children were used in his following analysis since there were problems involving deaths, incorrect addresses and other incorrect information on the Census.

It was found that 65% of the Census reported ages of

these 608 children were correct with no misreporting differences across gender. It was also seen that most of the misreported ages were moved to the next adjoining age group. Nine percent of the ages were misreported below the correct age and 26% were misreported above the correct age. Of the 215 cases of misreporting, 62% were misreported to the next higher age group, 21% to the next lower age group and 17% further afield.

The children's ages in years and months were also sought since it had earlier been suspected that a seasonal effect might be present and that errors of 12 months would be common. It was found however that errors of 12 months were no more common than errors of 11 months or 13 months, say. It was also seen that reported age differences of more than one year were spread evenly above and below the correct age.

An interesting phenomenon was observed when examining the ages of children under one year old since these ages were reported with the greatest accuracy and then the reporting accuracy declined as the child got older. In fact it was seen that as age increased the reporting discrepancies in whole years became greater. It became clear that the accuracy of the reported ages under one year was due to remembering the child's progress through the different childhood development stages, such as teething, weaning, movement, etc..

Caldwell concluded that age misreporting occurred partly because of digital preference but mostly because of

the tendency to approximate age to the next nearest birthday, in particular age advancement.

Two types of recall error were identified by Sudman and Bradburn (1973) who examined 500 studies in an attempt to identify the effects of time and memory factors on responses in surveys. The first error type was due to memory affects which lead to an underreporting of events; this was when a person simply forgot that certain events had taken place. The second type of error involved telescoping affects where events were remembered but placed incorrectly on a time scale. This type of error could lead to both overreporting and underreporting affects, i.e. forward telescoping and/or backward telescoping of events into a particular time period. An example using both types of error is where a person tries to recall all the cheques written in the past four weeks. He may accurately recall six of the eight checks written during the four week period and then telescope forward a cheque written five weeks previously.

Sudman and Bradman (1973) determined that both of these error components played a major role in event recall but that it also depended on the subject of study, i.e. medical, financial, voting or household expenditure. One important result they identified was that event overreporting did not occur in medical studies even when the reference period was small. This is of some importance to us since we are looking specifically at accidents resulting in medical advice or treatment.

A different manifestation of this type of phenomenon can be seen in a quality control problem. For example, consider the situation where an inspector on a production line searches for a defective product and misses some articles which are defective. The observed distribution would not be the 'true' distribution but would be a distribution with reporting errors.

The problem of recall was also examined by Sikkel (1985) who highlighted problems of event recall. Sikkel concluded that memory affect may depend on the elapsed time since the event took place and the number of events that had taken place. For example, events that occurred ten years ago may be forgotten and multiple events may be underreported. Hence, the longer the reference period, the stronger the memory effects.

Sikkel (1985) investigated questions of the type 'How many times did you undertake action X during t time units immediately preceeding the interview and when did these actions take place?' With access to the 1981 Dutch Health Survey he examined the number of G.P. contacts, specialist contacts and dentist contacts separately. He assumed that the number of each type of contact followed a Poisson distribution and could be modelled after taking into account memory and time effects. He then defined a three month reference period which was partitioned into t subintervals with a maximum of six contacts occurring during the reference period.

Several models were investigated which included both

time and memory effects. The best model for the number of G.P. contacts was defined such that memory effects increased with each week and after the first reported contact there was an additional effect. From the models of the number of specialist contacts Sikkel found that the number of contacts did not seem to play a role and he assumed that individuals remembered these events more accurately. Sikkel found that the models for the number of dentist contacts did poorly and provided three reasons for their failure. He split the population into three sub-populations, those who often go to the dentist, those who go every six months and those who never go. He concluded that the population was too heterogeneous for his models to produce adequate fits.

As a result of recall problems we suspect that different sections of the population have different recall abilities and that different conclusions would therefore be drawn from using different types of accident data. This is confirmed by Stewart-Brown et al (1985) who showed that different conclusions could be drawn when different accident outcome variables were used. They had access to the follow up 1975 British Birth Cohort Survey and considered 'all accidents' and 'hospitalised accidents' as two separate measures of accident rates for the children.

Using the type of neighbourhood variable (namely, poor urban area, average area, well-to-do area and rural area), they found that children were at high risk of having accidents if they lived in an average

neighbourhood, and that children living in poor urban areas were less likely to experience an accident than those in an average urban area. However, children from poor urban areas experienced admissions to hospital much more often than children from better urban areas and slightly more often than those from average urban areas. This contradiction could be due to the different definitions of an accident between the parents or perhaps accident recall difficulty, both of which produce a reporting bias.

Stewart-Brown et al (1985) also found that the most important risk variables for 'all accidents' were high social class, poor maternal physical and mental health, marital discord, child aggression and child overactivity. These variables are not necessarily independent of each other and interpretation needs careful consideration. For example, a low social class may mean a high risk of hospital admission but low social class is not independent of the child's neighbourhood, parents' income level, features of accommodation (such as overcrowding, no central heating, floor coverings), quality of toys and furnishings or other articles in the home (for example, dangerous toys, fragile furniture, faulty electrical appliances) and other variables. Similarly, poor maternal health may be due to a variety of factors such as the type of accommodation, social class or marital discord. Any kind of analysis must be aware of this confounding and a great deal of care must be taken in interpreting the

trends.

We now discuss the results of the preliminary, exploratory analysis of the data set available for our analysis.

1.3.3 ETHNIC DIFFERENCES IN THE DATA

The first step of the reported accident distribution modelling was to split the data by a single variable at a time into non-overlapping sub-populations. Probability distributions were then fitted to each of the reported accident distributions obtained from the sub-populations. Further details of the model fitting are given in Chapter 3 but at this stage it should be sufficient to say that the early analysis resulted in the decision to analyse the male and female populations separately and to use only those children of U.K. ethnic origin. Table 5 shows the reported accident mean and variance for each of the ethnic groups by gender and suggests that accident misreporting existed in the smaller ethnic groups.

It can be seen from Table 5 that 92% of the children in our data set came from the U.K. ethnic origin group. For three of the six male ethnic sub-populations the reported accident mean was greater than the variance and this was also seen in two of the six female ethnic sub-populations. The male ethnic group with the highest reported accident mean was the non-U.K. European male

group with a mean of 0.7460, and the lowest 0.3093, was seen in the Indian/Pakistani males. For the females the highest reported accident mean was 0.7500 for females from 'other' ethnic origin, and the lowest, 0.1600 for Indian/Pakistani females. The results certainly suggest a difference in the reported accident means between the different ethnic origin groups.

TABLE 5. The reported accident mean and variance for each ethnic group by gender obtained from the full data set of 13,135 children.

Sex	Ethnic group	Sample size	Mean	Variance
M A L E	European (U.K.)	6310	0.6957	0.8029
	European (other)	63	0.7460	0.6764
	West Indian	77	0.4156	0.3513
	Indian/Pakistani	97	0.3093	0.3825
	Other	33	0.5152	0.4451
	Unknown	228	0.6847	0.7418
	TOTAL	6808	0.6860	0.7893
F E M A L E	European (U.K.)	5830	0.5208	0.5903
	European (other)	65	0.2923	0.3038
	West Indian	75	0.1867	0.1539
	Indian/Pakistani	100	0.1600	0.1560
	Other	40	0.7500	0.9610
	Unknown	217	0.4700	0.5373
	TOTAL	6327	0.5085	0.5797

Further justification for using only those children of U.K. origin can be seen when the full data set of 13,135 children is split into sub-populations according to two variables only; ethnic origin (U.K. or non U.K.) and neighbourhood (poor urban area, average area, well-to-do area or rural area).

The results show that after splitting the full population of 13,135 children with respect to neighbourhood only, children from an average neighbourhood had the highest reported accident mean (0.6414) followed by poor (0.5892), well-to-do (0.5694) then rural (0.5256) neighbourhoods. Taking into account the ethnic origin of the child as well we examine the results for the U.K. origin children. The highest reported accident mean group was for children from a poor neighbourhood (0.6702), then average (0.6518), well-to-do (0.5737) then rural (0.5269) neighbourhoods, i.e. the reported accident mean for the poor neighbourhood increased from 0.5892 to 0.6702. For the non U.K. children the highest reported accident mean was for the average neighbourhood (0.5056), then rural (0.4878), well-to-do (0.4818) and poor (0.3754) neighbourhoods. This result may also be confounded with other aspects of misreporting between specific groups of parents.

It seems clear that accident underreporting exists in the non U.K. ethnic groups, so for homogeneity it would be sensible to discard the relatively small ethnic sub-populations and use only those children of U.K. ethnic origin. Hence, our analysis will be based on those children of U.K. ethnic origin only. In using only these the male population was reduced from 6,808 to 6,310, a loss of just over 7% of the total male population and the female population was reduced from 6,327 to 5,830, a loss of just under 8% of the total female population.

We have now identified a number of suspected deficiencies to be aware of when analysing the reported accident data. These are the systematic misreporting of accidents by particular groups of parents, the accident recall problem over the five year period, the misreported accident times which produce a 'heaping' affect, the different results obtained from using 'all accidents' and 'hospitalised accidents' as outcome variables, the suspected bias of the mother's assessment of her child's personality and the possible bias of the Health Visitor's assessment of the child's home. We now discuss the simple reported accident trends found in the data.

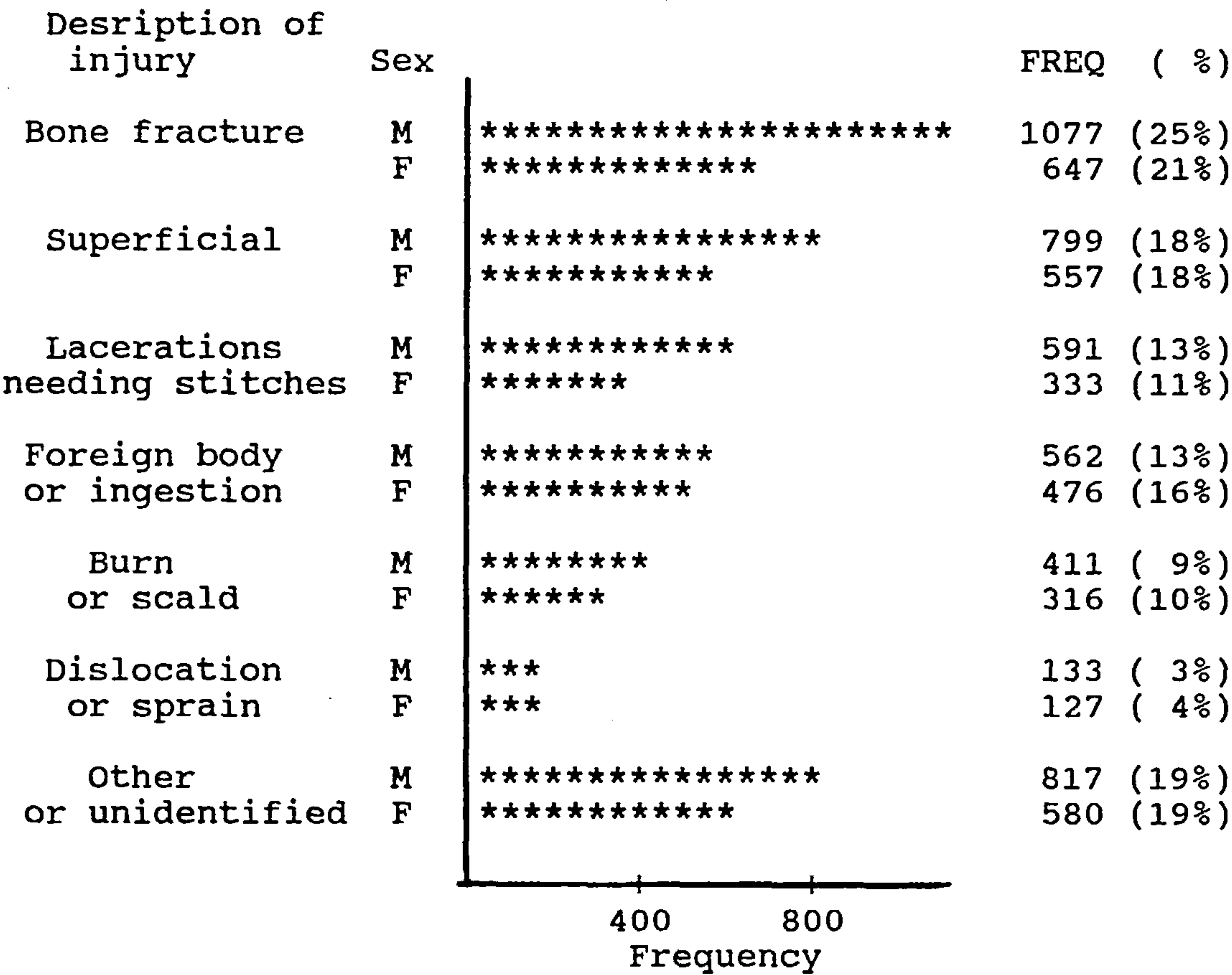
1.3.4 SIMPLE REPORTED ACCIDENT TRENDS IN THE DATA

A better understanding of the underlying accident distribution may be obtained if we first examined the available accident data in more detail. In particular we would like to identify variables which appear in reported accident frequency trends or associations. Reported accident frequency graphs were therefore constructed with respect to accident details, for example, type of accident, place of accident, body area injured. The following five figures give breakdowns of the accident details by gender.

Figure 1 shows a small difference between gender with respect to the types of injuries incurred from the

reported accidents. Proportionally, from the 4,390 reported male injuries and 3,036 reported female injuries shown in Figure 1, we can see that males had slightly more bone fractures and slightly more lacerations requiring stitches than females. It was also seen that proportionally, females had slightly more reported foreign body/ingestions, dislocations/sprains and burns/scalds than males. For both males and females bone fractures were the most frequent type of reported injury followed next by superficial injuries.

FIGURE 1. Types of injury by gender with percentages in brackets.



One would suspect that injuries such as bone

fractures, dislocations and lacerations requiring stitches would be reported more accurately than accidents resulting in minor injuries, such as superficial injuries, or sprains. The reason for this might be because of the seriousness of these injuries and the necessity for specialised medical attention. We should also recall that accidents which resulted in both a bone fracture and a laceration requiring stitches, were coded as bone fractures. This would mean that the proportion of reported accidents which resulted in lacerations requiring stitches was actually higher than seen in Figure 1. For example, a child may break his wrist and severely lacerate his arm so that stitches are required, hence the injury would be coded as a bone fracture.

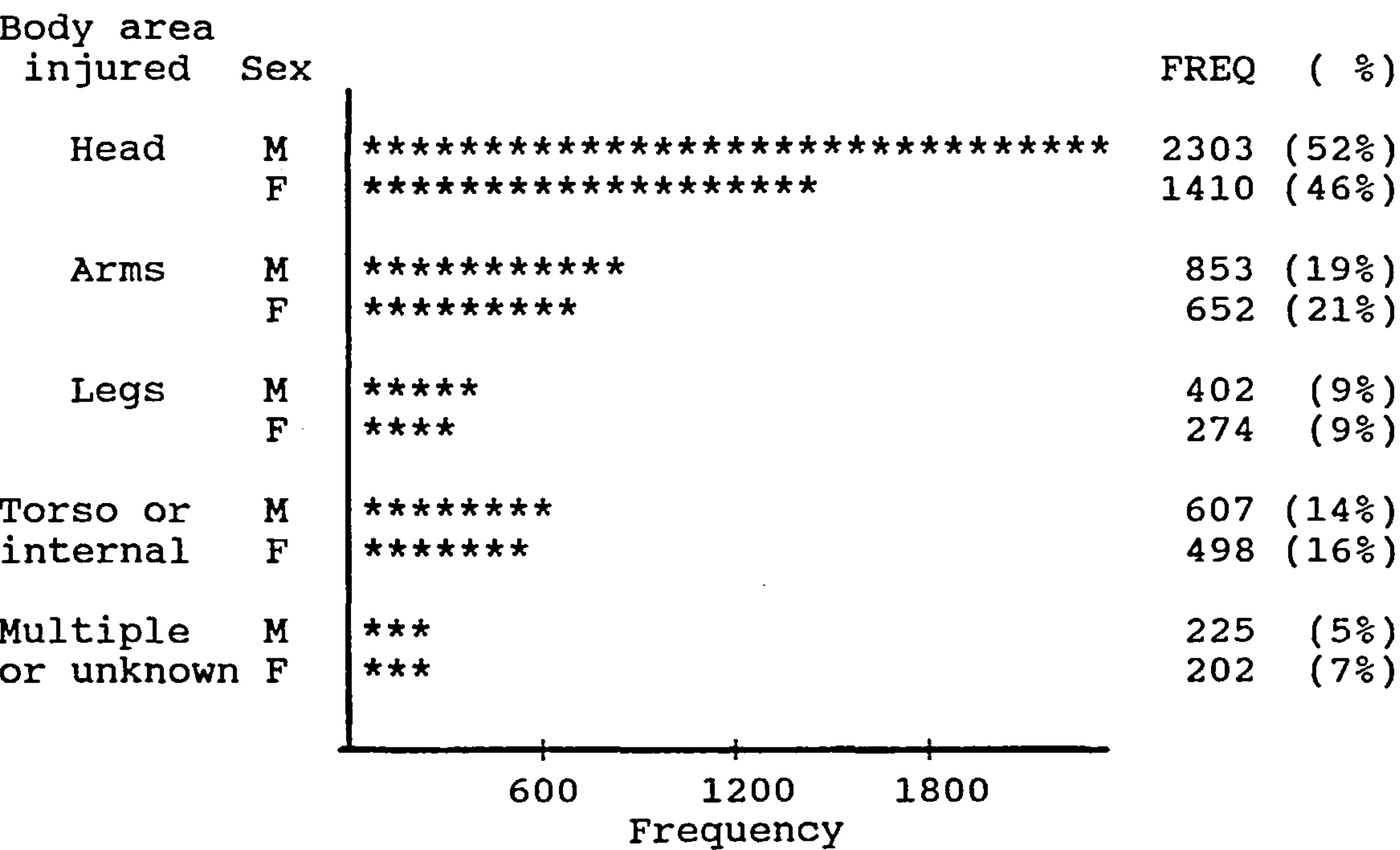
We also see that 18% of male and 18% of female reported accidents were classified as superficial accidents. It might be that the parents' threshold of deciding whether an injury requires medical advice/treatment or not varies according to the parents and may also be dependent on the child's previous accident history. For example, if the child's first accident is a serious accident resulting in a prolonged hospital stay then this accident may be used as a yardstick for all subsequent accidents. Hence, the parents will not seek medical advice until the accident is quite serious. Alternatively, some parents may have low thresholds until their child actually has a serious accident at which point they may then redefine their

threshold level to take into account the serious accident.

When looking at the body area injured one would expect head injuries to feature strongly because children under five years old are still learning to use and develop their physique. In early childhood the head is oversized in comparison to the rest of the body and is particularly vulnerable. Head injuries are also potentially more serious than other injuries since internal head injuries can often go unnoticed without any outward sign of injury.

From Figure 2 we see the breakdown of body area injured by gender and see that head injuries do feature strongly in both the male and female populations.

FIGURE 2. Body area injured by gender with percentages in brackets.

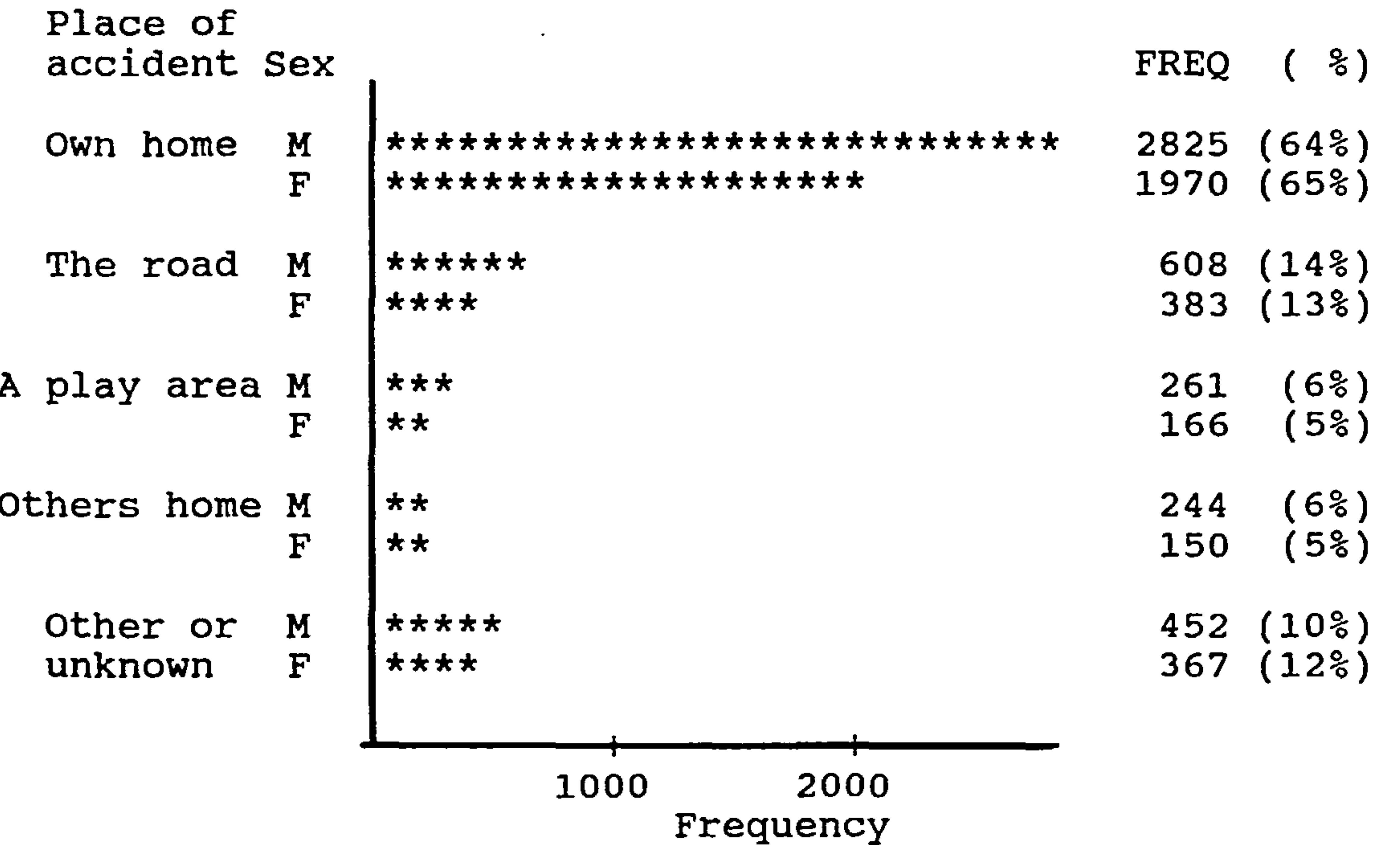


By comparing the results between gender there is

evidence to suggest that males had proportionally more reported head injuries than females, whilst females had proportionally more reported arm injuries and more torso injuries than males. For both males and females the most frequently reported body area injured was the head followed by the arms, the torso and then the legs.

As noted earlier the cohort consists of very young children, therefore, regardless of gender, we would expect that the most common place where accidents occurred would be the child's home. This is confirmed from Figure 3 which gives a breakdown of the places where the reported accidents occurred.

FIGURE 3. The place where the accident occurred by gender with percentages in brackets.

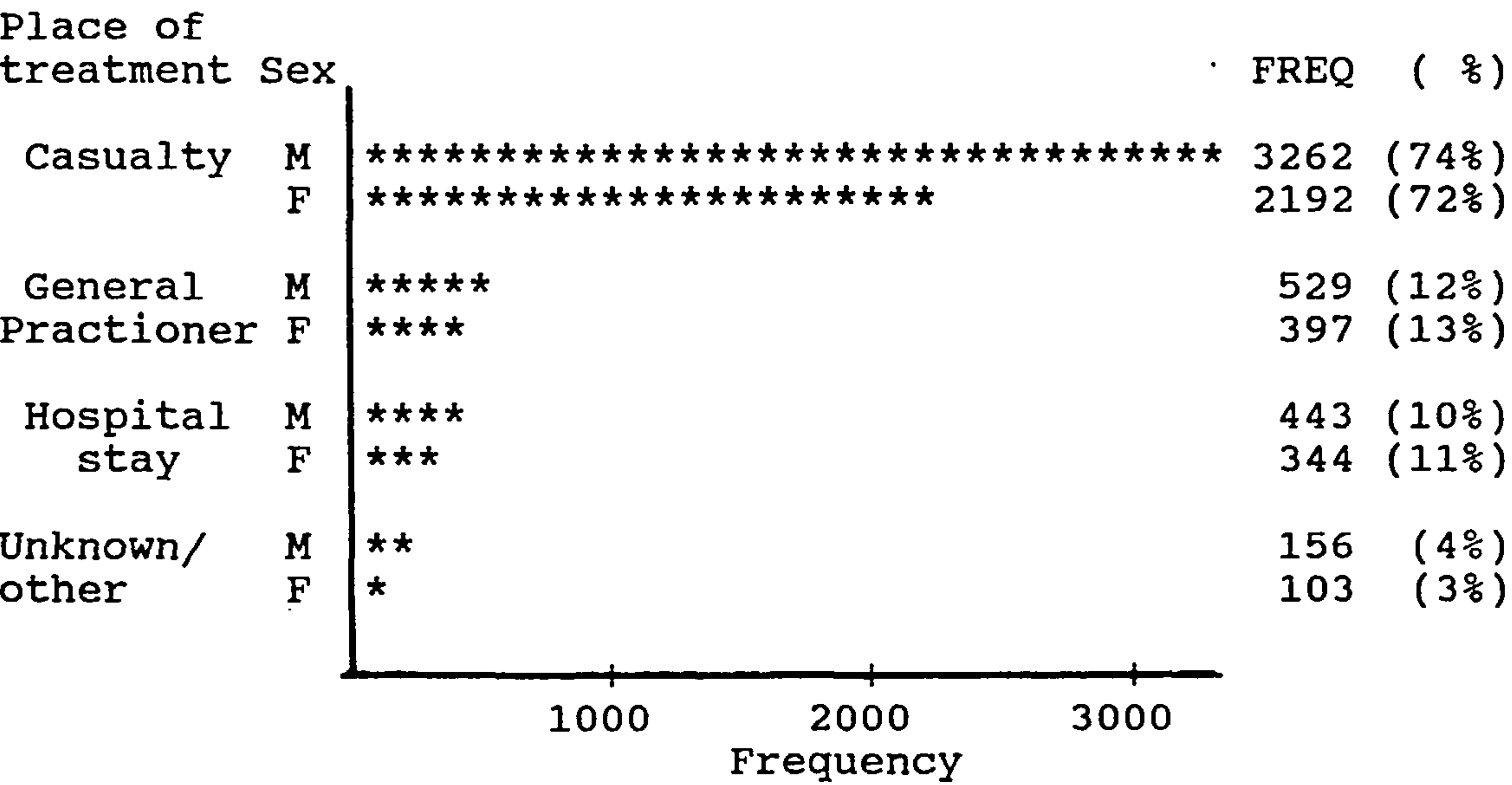


We emphasise that the children were under five years old and would have been under very close supervision

in their main environment, their own home. A child under five years old does not have the same freedom of movement say as a child of eight years old. We would therefore expect accidents in the home to feature strongly. For both males and females the majority of reported accident occurred in the child's own home, followed next by accidents on the road.

Since each reported accident resulted in an injury that required medical advice and often, in addition, specialised medical treatment, for example, stitches or the setting of a bone fracture, one would expect that most of the injuries would be treated in a hospital casualty department. This can be verified by examining Figure 4.

FIGURE 4. The place where treatment was given, by gender with percentages in brackets.



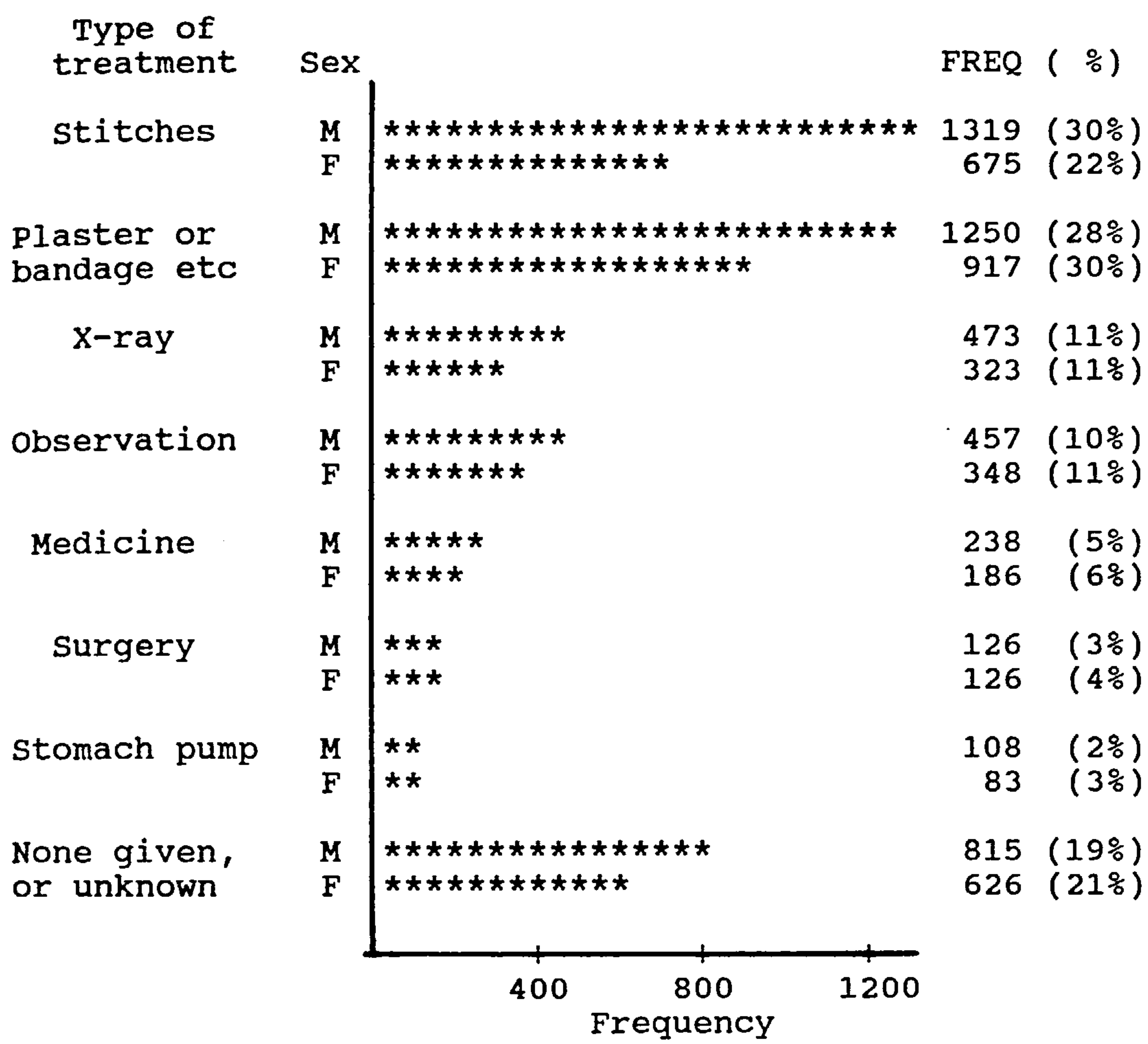
From Figure 4 we see that the majority of reported injuries for both males and females were treated in a casualty department. The next most frequented place for

treatment was the General Practitioner's followed next by the hospital during a stay of one night or more. There was no striking difference between the places where the reported injury was treated for males and females.

When looking at the type of treatment given for each injury (Figure 5), we must recall that an accident could have resulted in more than one injury. For example, a male child may trip over whilst running in the street and severely lacerate his knee and also break his ankle. Hence, as a result of one accident he would have stitches in his knee and his broken ankle set in plaster, i.e. two different types of treatment for one accident.

From examining Figure 5 we note some differences between gender for the types of treatment given for each injury. For males the most frequently reported treatment was stitches, then the application of a bandage/plaster or splint, followed next by an x-ray and then observation (30%, 28%, 11%, and 10% of all reported male injuries respectively). For females the most frequently reported treatment was the application of a bandage/plaster or splint, followed next by stitches, observation, and then x-ray (30%, 22%, 11% and 11% of all female injuries respectively). Males had proportionally more reported injuries requiring stitches than females (30% and 22% respectively) which suggests a difference in the male and female reported accident distributions.

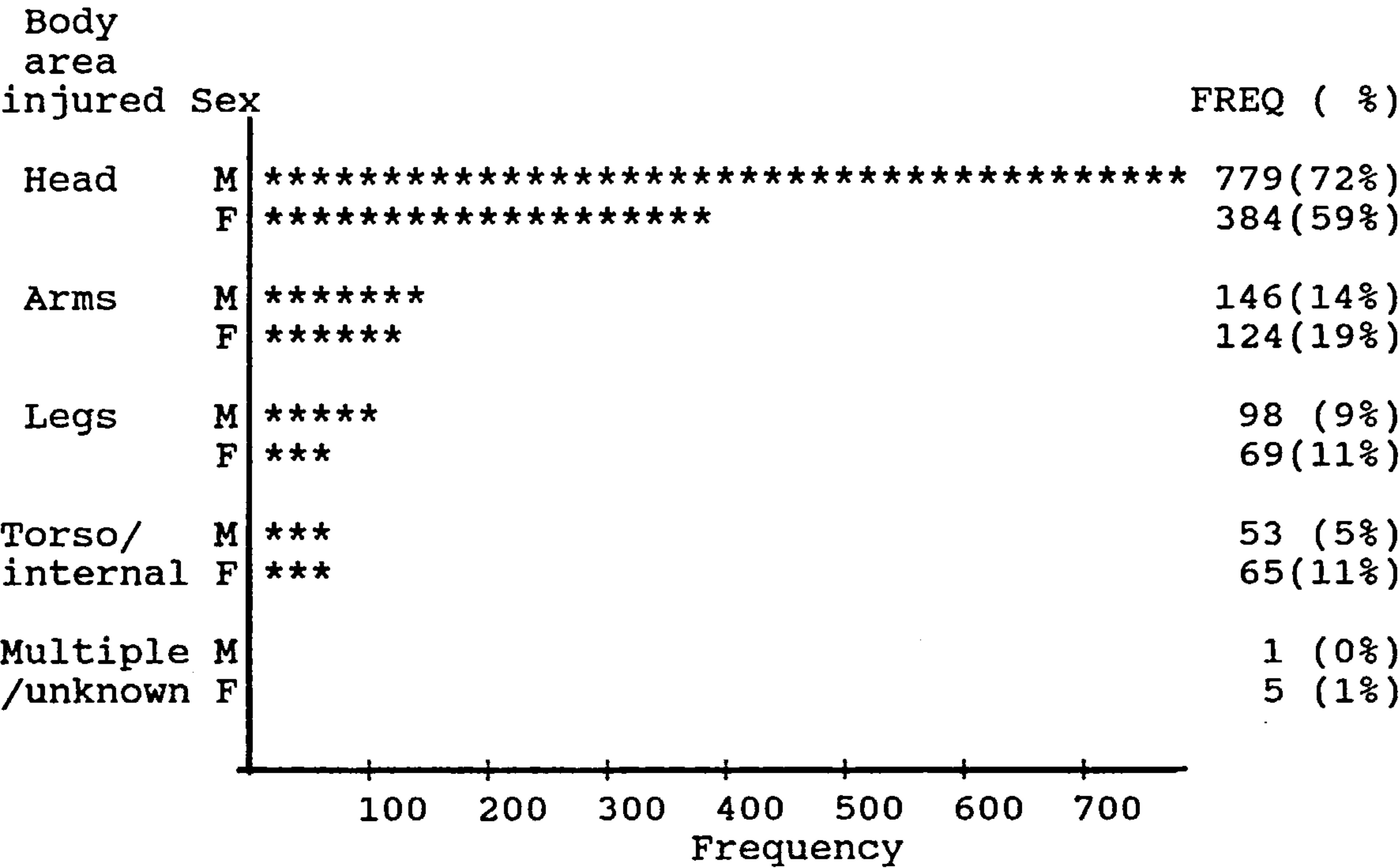
FIGURE 5. The types of treatment given by gender with percentages in brackets.



The results from Figure 5 may also seem to contradict the results shown in Figure 1 until we recall that an accident which resulted in both a bone fracture and a laceration requiring stitches was coded as a bone fracture. In addition, 817 (19%) of the 4,390 reported reported male injuries and 580 (19%) of the 3,036 reported female injuries were either not treated at all or the parents had forgotton the type of treatment that had been given. This is an example of the recall problems associated with retrospective data collection.

The next six figures are breakdowns of the types of injury with respect to the body area injured. Figure 6 gives a breakdown of the 1,077 reported male bone fractures and the 647 reported female bone fractures with respect to the body area injured and for both males and females it was observed that head fractures were reported the most frequently. Proportionally, 72% of all reported male fractures were head fractures in comparison to 59% of all reported female fractures. A possible explanation for this difference could be that males played rougher or more physically demanding games than females. As the head is very vulnerable during early childhood we would therefore expect to see proportionally more reported male head fractures.

FIGURE 6. A breakdown of bone fractures and area of the body fractured by gender, with percentages in brackets.



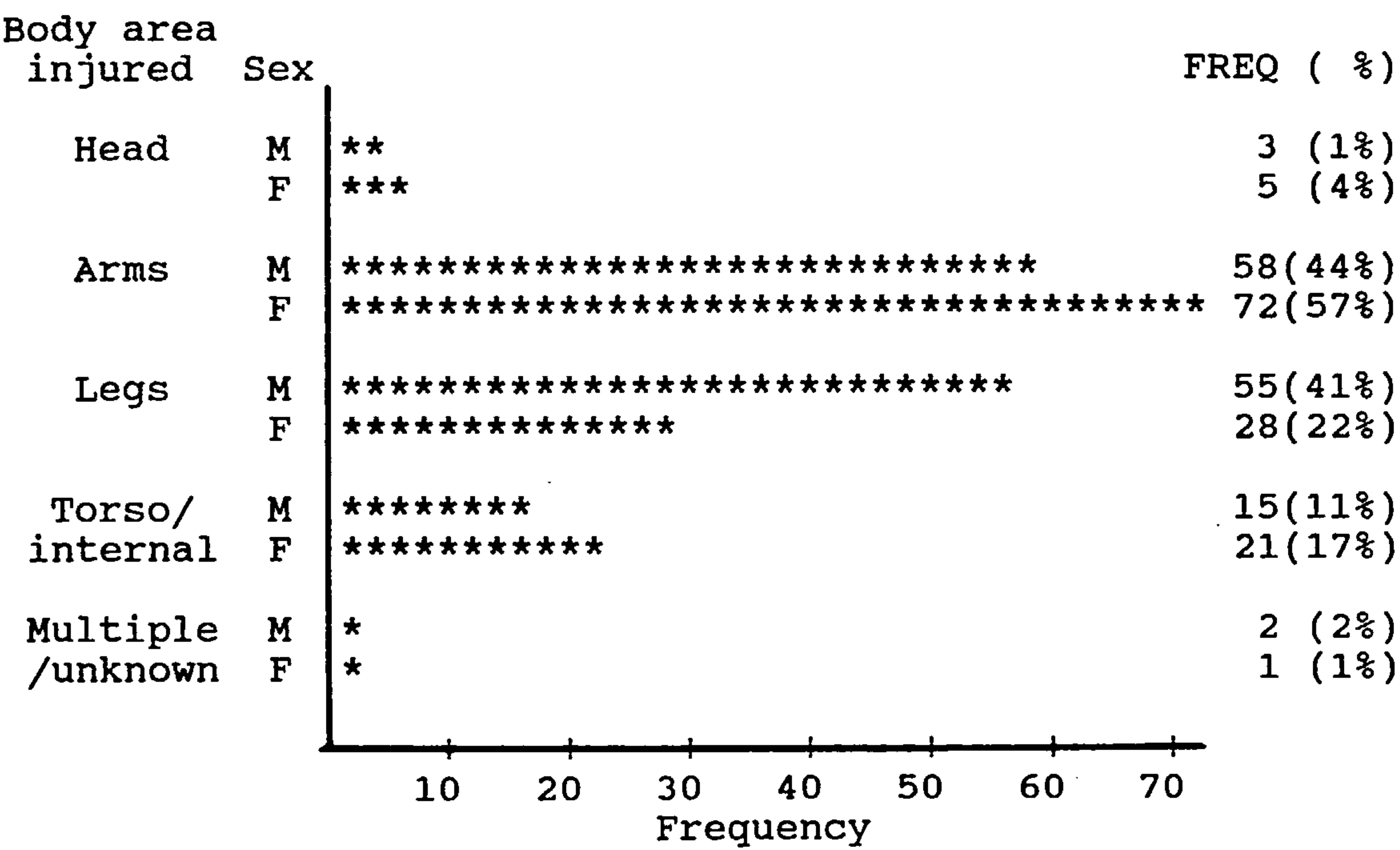
It was also found that females had proportionally more reported arm fractures than males (19% and 14% respectively), and also had proportionally more reported torso fractures than males (11% and 5% respectively). This result is not suprising if we accept that females played games less physically demanding on the upper body than males. We could then assume that young females did not develop their upper physique as quickly nor as strongly as males and would therefore be injured far more easily in less demanding situations.

The breakdown of 133 reported male dislocations/sprains and 127 reported female dislocations/sprains with respect to the body area dislocated/sprained is given in Figure 7 and we see that females had proportionally more reported arm dislocations/sprains than did males (57% and 44% respectively), and proportionally more reported torso dislocations/sprains than males (17% and 11% respectively). Males had more reported leg dislocations/sprains than females (41% and 22% respectively).

A possible explanation for this particular result is that the activities and games that males played strengthened the upper body and torso, yet were not sufficient to strengthen the legs at the same rate. It is also possible that males overestimated the strength of their legs and that females overestimated the strength of their arms. This would mean that reaction times might be

slower for different parts of the child's body with respect to gender, hence the above results. Males might also be more cavalier in their attitude towards danger and therefore get into dangerous accident situations which resulted in more serious injury. Hence, we again have to consider the different play activities of children.

FIGURE 7. A breakdown of dislocations/sprains and area of the body injured by gender, with percentages in brackets.

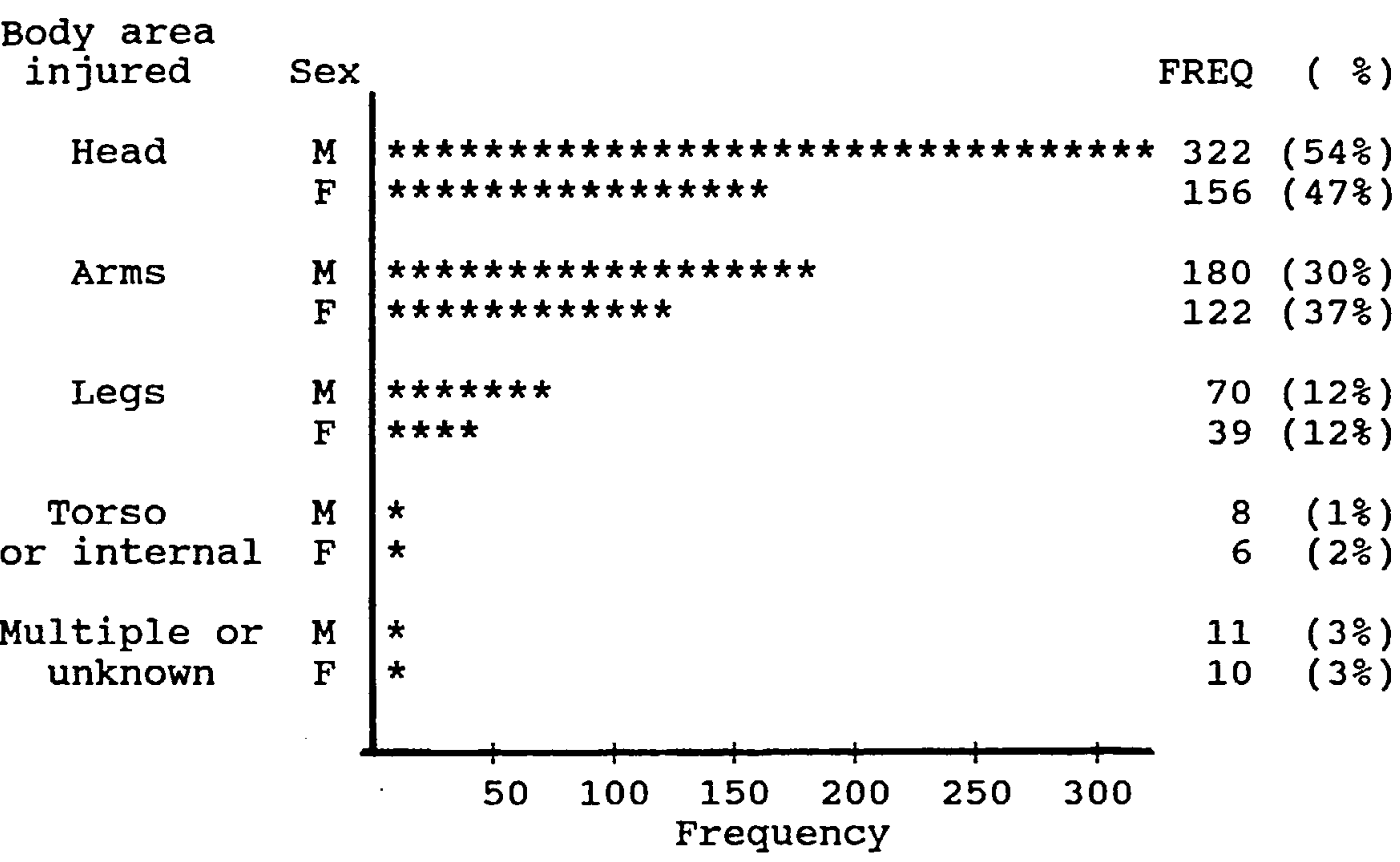


Differences between gender were also found when looking at the 591 reported male lacerations requiring stiches and the 333 reported female lacerations requiring stitches shown in Figure 8.

Males had a greater proportion of reported head lacerations than females (54% and 47% respectively) and females had a greater proportion of reported arm lacerations than males (37% and 30% respectively). This

result supports the earlier hypothesis that males and females are involved in different play activities, leading to different types of injury.

FIGURE 8. A breakdown of lacerations requiring stitches and area of the body injured by gender, with percentages in brackets.

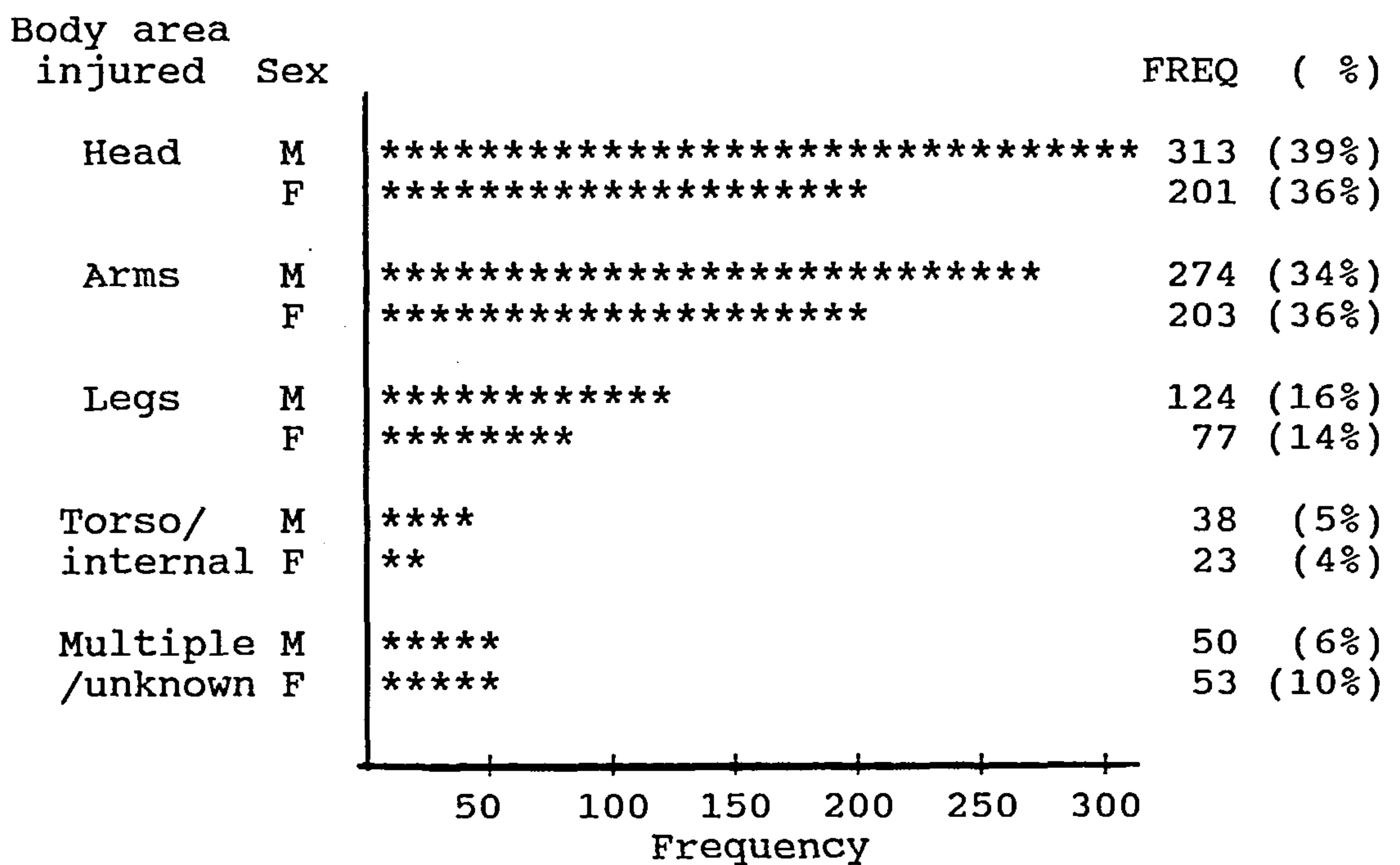


When looking at the breakdown of reported superficial injuries (799 in the male population and 557 in the female population) with respect to the body area injured there did not appear to be any great difference between the sexes. This is shown in Figure 9.

Proportionally, males had slightly more reported superficial head injuries than did females (39% and 36% respectively) and for both males and females, reported superficial head injuries were the most frequent superficial injuries followed next by reported superficial arm injuries and then reported superficial leg injuries.

It seems likely that superficial injuries would be underreported by parents because these injuries would simply not be remembered in comparison to the more serious injuries. In addition, it is thought that the parent's definition of a superficial injury could range from being very trivial to bordering on serious so there could be a great deal of accident misreporting.

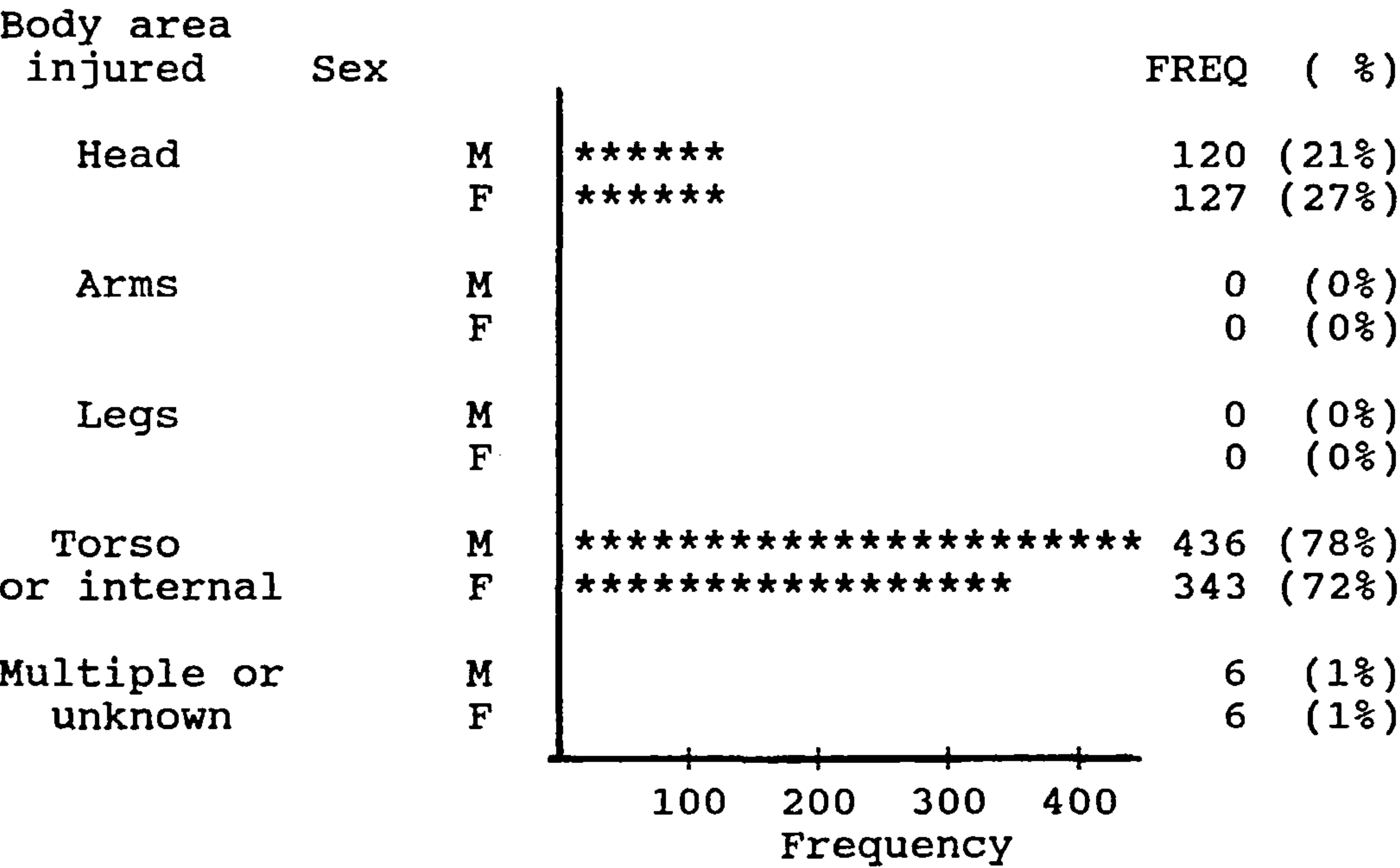
FIGURE 9. A breakdown of superficial injuries and area of the body injured by gender, with percentages in brackets.



As would be expected from the breakdown of reported foreign body and ingestion type injuries shown in Figure 10 (562 in the male population and 476 in the female population), the torso and the head were the two most common sites of injury. This is because ingestions relate to the digestive system in the torso, and most of the foreign body type injuries involved insertion of an

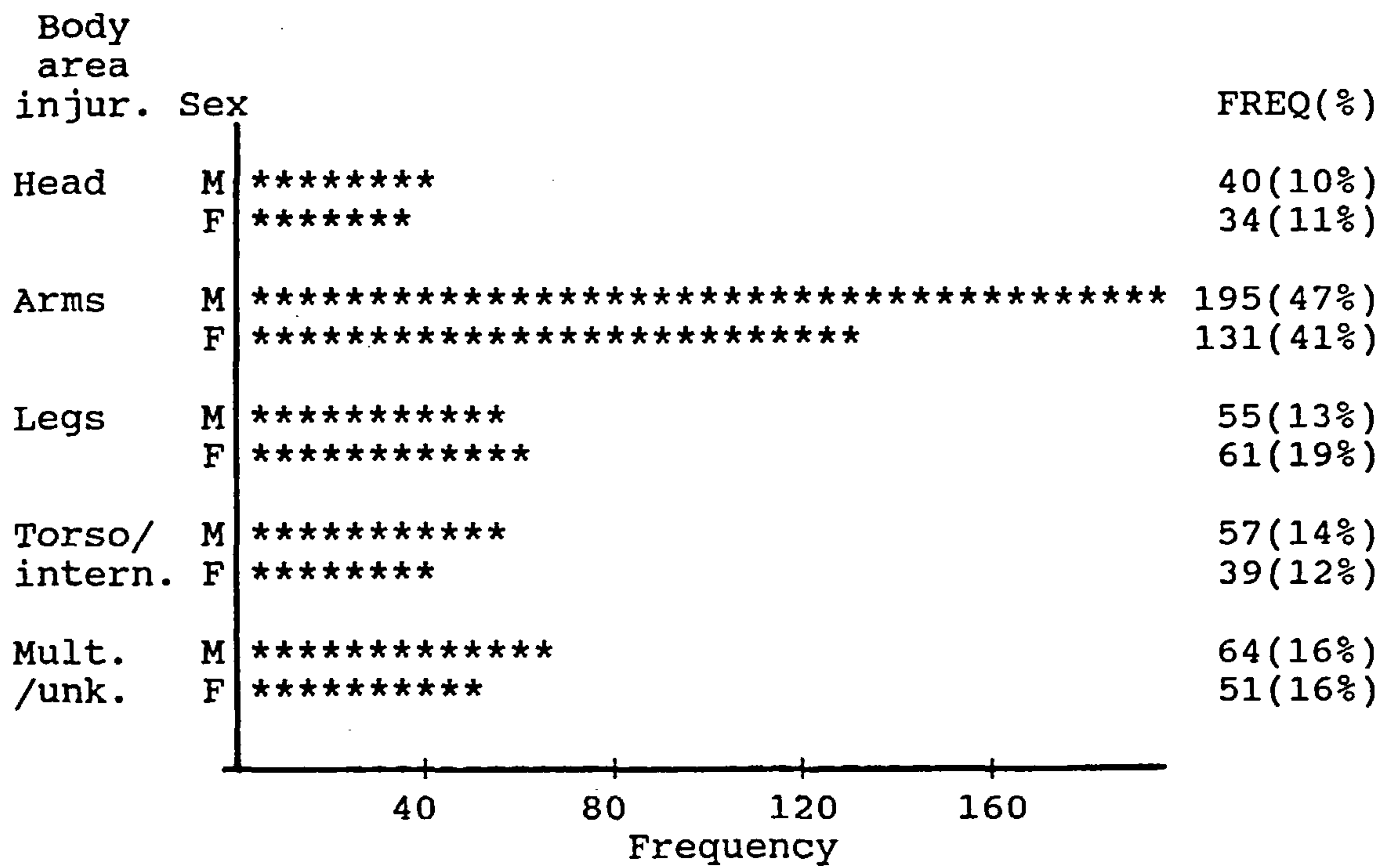
object(s) into the ears, eyes, nose or mouth. For both males and females, reported foreign body/ingestions involving the torso were the most frequent area of the body involved followed next by reported foreign body/ingestions involving the head. Females had proportionally more of this type of reported injury involving the head than males (27% and 21% respectively) but males had proportionally more involving the torso than females (78% and 72% respectively).

FIGURE 10. A breakdown of foreign body and ingestion type injuries and area of the body injured by gender, with percentages in brackets.



From Figure 11 we see the breakdown of the 411 reported male burns/scalds and 316 reported female burns/scalds with respect to body area injured and again we note some differences between gender.

FIGURE 11. A breakdown of burns/scalds and area of the body injured by gender, with percentages in brackets.



Proportionally, males had more reported arm burns/scalds than females and females had proportionally more reported leg burns/scalds than males. For males the most common area of the body reported to be burned/scalded was the arms, followed by the torso and then the legs (47%, 14%, and 13% of all reported male burns/scalds respectively). For females, the most common area of the body to be burned/scalded was the arms, then the legs followed by the torso (41%, 19% and 12% of all reported female burns/scalds repsectively).

These burns/scalds results could be due to the type of clothing that children wore in the 1970's since traditionally girls wore dresses or skirts which left their legs exposed. Hence, it would be logical to assume

that the severity of a burn or a scald would be reduced if the skin were protected in some way, i.e. by trousers. It is also possible that in comparison to females, males had stronger arms/hands and could open hot water taps, fire protection guards, etc., and generally get into more burn/scald accident situations which result in injury to their upper torso.

As a summary of these reported accident frequency breakdowns, we note that most of the reported accident frequency trends identified in the exploratory analysis were not surprising. In particular, the relatively large proportion of reported home accidents and the large proportion of reported head injuries were expected. In terms of types of reported injuries such as bone fractures, dislocations/sprains, lacerations requiring stitches and superficial injuries it appears that proportionally, females tended to injure their arms more than males but in terms of burns/scalds males tended to injure their arms more than females. Males also tended to dislocate/sprain their legs more than females but females tended to burn/scald their legs more than males. It was also observed that males had proportionally more head injuries than females. Clearly there is strong evidence suggesting that the underlying features of the male reported accident distribution are different from the underlying features of the female reported accident distribution.

Before attempting to unravel some of the differences

we should first recall that the results refer to data collected retrospectively in 1975 for the previous five years. It would therefore be safe to assume that males of that era were brought up to participate in more physically demanding play activities and a more rough and tumble life-style than females.

During those years males were encouraged to play adventurous and boisterous games whilst females were encouraged in more placid activities. Hence, we would certainly expect some differences in the features of the reported accident distributions between the sexes. This would account for the reported head fractures and reported lacerations requiring stitches which were proportionally more common in males than females. It could also account for the way in which males had proportionally more reported leg dislocations/sprains than females and yet had proportionally fewer reported arm dislocations/sprains, i.e. males perhaps participated in games and activities that were more physically demanding and which strengthened the upper body whilst females participated in activities that did not develop the upper physique as much.

When looking at the trend for males to have proportionally more arm burns/scalds and proportionally fewer leg burns/scalds than females we must recall that in the 70's females upto the age of five years old were most likely to be dressed in a skirt or a dress. Hence, females' legs were less protected than males' legs and we would therefore expect females to have proportionally more

leg burns than males, a phenomenon which was observed. We might also expect that over all types of accidents females would have proportionally more burns/scalds than males, however this phenomenon was not observed to any great extent (see Figure 1.).

A further breakdown of the accident information was obtained by splitting the data set with selected variables. We now examine those results.

1.3.5 SIMPLE REPORTED ACCIDENT ASSOCIATIONS IN THE MALE POPULATION

The variables used in the search for accident associations were selected on the basis of earlier research, curiosity and discussion with medics. The 269 variables available for the analysis were not all examined and only 27 variables were selected for this preliminary, exploratory analysis. The variables used in this preliminary analysis are as follows.

- 1) The child's mother's age in April 1970, (less than 20 years old, 20-30, over 30 years old).
- 2) The child's father's age in April 1970, (less than 20 years old, 20-30, over 30 years old).
- 3) The social class of the child's father based on his occupation, (manual, non-manual).
- 4) The standard of the household equipment in the child's home, (adequate, well equipped, low/very low standard).
- 5) The level of tidiness of the child's home, (average, very/over tidy, untidy/chaotic).
- 6) The type of neighbourhood of the child's home, (average, poor, well-to-do or rural).
- 7) The number of household moves during 1970-1975, (0, 1, 2, 3, more than 3).

- 8) Whether the child's mother smoked, (smoker or non-smoker).
- 9) Whether the child's father smoked, (smoker or non-smoker).
- 10) The type of accommodation in which the child lived in, (owned outright, being bought, a council property, other).
- 11) The number of younger siblings, (0, 1, more than 1).
- 12) The number of older siblings, (0, 1, 2, more than 2).
- 13) The person-per-room ratio, (0.5 or less, 0.501-0.75, 0.751-1.0, 1.001-1.50, 1.501-2.0 or more than 2.0).
- 14) The child's social index, (most advantaged, advantaged, average, disadvantaged, most disadvantaged).

The following variables refer to characteristics of the child's personality and are assessed by the child's mother. There are three responses as follows; not applicable, applied somewhat and certainly applied.

- 15) Whether the child was restless.
- 16) Whether the child squirmed or fidgeted.
- 17) Whether the child destroyed belongings.
- 18) Whether the child fought with other children.
- 19) Whether the child was not much liked.
- 20) Whether the child was irritable.
- 21) Whether the child was miserable.
- 22) Whether the child took things.
- 23) Whether the child bit his/her nails.
- 24) Whether the child was disobedient.
- 25) Whether the child could not settle.
- 26) Whether the child told lies.
- 27) Whether the child bullied others.

The male and female populations were split with respect to one of the above variables at a time into non-overlapping sub-populations and the reported accidents in each sub-population were then broken down with respect to either the type of accident or the body area injured.

It should be recalled that some of the variables such as the tidiness of the child's home or the standard of the

household equipment, were assessments made by the Health Visitor to the child's home and may be biased. Similarly, characteristics of the child's personality were assessed by the child's mother and may also be biased.

Some data splits resulted in small sub-populations of less than 100 which have been excluded from this part of the analysis but from the remaining sub-populations we were able to identify several reported accident trends/associations. In the male population for example, when examining the 591 reported lacerations which required stitches we observed that proportionally, this serious reported injury occurred more often in males from a poor urban neighbourhood than males from any other neighbourhood. The proportions of reported lacerations with respect to neighbourhood were as follows; poor(18%), average(14%), well-to-do(13%) and rural(11%). It is thought that males from poor neighbourhoods would have fewer facilities in the house, i.e. no play room, lower quality toys, harder floor coverings, etc., which would place them in a higher accident risk group.

This association can be contrasted to the association found when looking at the 799 reported male superficial injuries broken down with respect to neighbourhood. It was found that proportionally, reported superficial injuries were less frequent in males from a poor neighbourhood than from other neighbourhoods. The proportions of reported superficial injuries with respect to neighbourhood were as follows; average(19%), well-to-

do(19%), rural(17%) and poor(13%).

A possible explanation for these two peculiar results could be that when a male from a poor neighbourhood had an injury it was usually not superficial but a more serious injury. Alternatively, parents in poor neighbourhoods may not have reported superficial injuries as often as parents from other areas, the reason being that superficial injuries were more frequent and not regarded as accidents but just part of growing up in a poor neighbourhood. For example, consider a household in a poor neighbourhood with lino floor coverings throughout the home. If a male hurts himself every time he falls over onto the hard floor covering then his parents are less likely to seek medical advice/treatment or report such an injury until the injury is severe, i.e. a laceration requiring stitches.

As discussed earlier we suspect that the parents' threshold of deciding whether an injury requires medical advice/treatment or not could also be dependent on the child's previous accident history as well as environmental variables. Indeed, the peculiar results highlighted above support this and provide strong evidence of differential reporting across the population. This phenomena of differential accident reporting has implications for any kind of accident analysis and is investigated further in Chapters 4 and 5.

The same type of reporting phenomenon was also seen in the breakdown of social index. Looking first at the breakdown with respect to the 591 reported lacerations

requiring stitches we found that proportionally this type of injury was reported more often for males from the most disadvantaged social index group (group 5) than for males from other social index groups. The proportions of reported lacerations from each social index group were as follows; group five(21%), group four(13%), group two(13%), group one(13%) and group three(12%), where group five is the most disadvantaged group and group one the most advantaged group.

Now examining the 799 reported superficial injuries broken down with respect to social index we see that proportionally, superficial injuries were reported least frequently for males from the most disadvantaged index group than for males of the other social index groups. The proportions of reported superficial injuries for each social index group were as follows; group three(20%), group one(20%), group four(17%), group two(17%) and group five(14%). It is quite likely that social index is closely associated with the child's neighbourhood, hence the similar results.

Still looking at the 799 reported superficial injuries we also observed that proportionally, superficial injuries were reported less frequently in males who had more than two older siblings than for males with two or less older siblings. The proportions of reported superficial injuries with respect to the number of older siblings were as follows; two(21%), none(19%), one(18%) and more than two(14%).

A possible explanation for this could be that older siblings looked after the younger male and reduced his risk of a superficial injury because of the extra attention and supervision. Alternatively it is possible that superficial injuries did occur more often but were not reported by the parents because less attention was given to superficial injuries in larger families. For example, we would assume that overall there would be more accidents in a family with four children in it than in a single child family, and it seems likely that the less serious accidents will be forgotten or discounted in the larger family. It is also likely that this variable is confounded with other variables such as the number of younger siblings, the sex of the siblings, the age of the parents, etc..

We now discuss the reported accident trends with respect to the body area injured. Examining the 2,303 reported head injuries with respect to the social index of the males we observed that males in the most advantaged social index group had proportionally more reported head injuries than males from the other social index groups. The proportions of reported head injuries with respect to social index group were as follows; group one(56%), group two(53%), group three(53%), group four(51%) and group five(48%), where group five is the most disadvantaged group and group one the most advantaged group.

This is a suprising result since one would naturally assume that children from poor urban neighbourhoods would

have fewer advantages than children from other areas and hence would be in a higher accident risk group. One would have expected children from the most-advantaged social index group to have better housing, a play-room, better toys, better floor coverings, better furnishings and better facilities, etc., than children from other social index groups.

It is possible that children from the most advantaged social index group actually do have proportionally more head injuries than children from other social index groups, however, this seems unlikely. In fact, the most likely explanation for this peculiar result is differential accident reporting.

The possibility which must not be overlooked is that males have proportionally the same number of head injuries across social index group but that parents from the most advantaged social index group recognise that child head injuries are potentially more serious than injuries to other parts of the child's body. These parents therefore report head injuries more often than parents from other social index groups.

Another alternative is that head injuries actually are proportionally more frequent in the most disadvantaged group, but because they are more frequent, they are not readily recalled and the seriousness of the injury is not recognised. This same type of misreporting was discussed for superficial accidents and the child's neighbourhood, where it was found that there was a lower proportion of

superficial injuries in males from the poorest neighbourhood than in males from other neighbourhoods.

Out of the three possibilities presented here we would favour either the second or third or a combination of them as the correct interpretation of this peculiar result. Again we have strong evidence that differential accident reporting is present.

Looking at particular characteristics of the child's personality it was observed that males who certainly fought other children had proportionally fewer of the reported 2,303 head injuries than males who didn't fight or who occasionally fought. The proportions of reported head injuries with respect to whether the male fought were as follows; did not fight(54%), sometimes fought(52%) and certainly fought(46%). This is an unexpected result since we would automatically expect males who certainly fought other children to place themselves in a higher accident risk group and have more reported head injuries than other males. A male who fought others would have more opportunity to injure himself simply because he places himself more often in situations where accidents are likely to occur.

One explanation for this result could be that males who fought others became more physically adept and developed a stronger physique than other children, hence they were injured less. Alternatively, it is possible that parents of males in this category actually underreported head injuries because they were common

occurrences. The most logical explanation would seem to be a combination of these two.

Continuing to look at characteristics of the child's personality we found another unexpected result when examining the breakdown of the 2,303 reported head injuries with respect to whether the male settled. Males who had difficulty settling had proportionally fewer reported head injuries than males who had no difficulty settling. The proportions of reported head injuries with respect to the male settling or not were as follows; did not apply(53%), applied somewhat(53%) and certainly applied(46%). This is a puzzling result since one would have expected that males who had trouble settling would be more active and be at greater risk of having an accident. Hence we should once again be aware of the strong possibility of accident misreporting.

We would expect the number of household moves to have some type of effect on the reported accident frequencies since the home is a young child's main environment. It is suspected that any household move would disorient a child. When looking at the number of household moves with respect to the area of the body injured, it was observed that males who had never moved home had proportionally fewer of the 607 reported torso injuries than males who had moved. It seems likely that the age of a child would be confounded with this variable since a child would become more aware of its surroundings as it got older and became more mobile. The proportions of reported torso injuries

with respect to the number of moves were as follows; two moves(20%), one move(15%), more than three moves(15%), three moves(14%) and no moves(11%).

We now discuss the reported accident associations seen in the female population and then compare the results across gender.

1.3.6 SIMPLE REPORTED ACCIDENT ASSOCIATIONS IN THE FEMALE POPULATION

In this section we examine and discuss the simple reported accident trends/associations identified in the female population. The reported accident frequency trends/associations identified in the male population were not consistently reproduced in the female population. Several variables which appeared in the male reported accident associations were also present in this analysis of the female population but were associated with different accident trends.

There is only one association seen in the male reported accident associations which is almost identically repeated in the female population. From the breakdown of 1,410 reported female head injuries and social index group it was found that reported head injuries were proportionally more frequent in females from the most advantaged (group one) category than the other categories. The proportions of reported head injuries with respect to

social index group were as follows; group one(51%), group two(48%), group three(47%), group five(45%) and group four(41%), where group one is the most advantaged group and group five the most disadvantaged group. The implications of this association for the issue of misreporting were discussed in the previous section.

A variable which appeared in the reported male accident associations and also appears in this section but with respect to a different association is the child's neighbourhood. From a breakdown of the child's neighbourhood with respect to the 316 reported female burn/scald injuries it was found that reported burns/scalds were proportionally less frequent in females from well-to-do neighbourhoods than females from other neighbourhoods. The proportions for reported burns/scalds with respect to the child's neighbourhood were as follows; rural(14%), poor(13%), average(10%) and well-to-do(7%). A simple explanation for this association could be that open fires were more common as a form of heating in rural, poor and average neighbourhoods. Houses in well-to-do neighbourhoods were perhaps equipped with other, better methods of central heating. If this were the case then children in well-to-do neighbourhoods were not exposed to the same risks to burns/scalds as children from other areas. This result may also be linked to the result seen in Figure 11 where it was observed that females had proportionally more leg burns/scalds than males.

The number of household moves also appears as an

influencing variable and after examining the breakdown of the 557 reported female superficial injuries with respect to the number of household moves it was found that proportionately, reported superficial injuries were more frequent for females with more than three household moves than for other females. This is not unexpected since the affects of a household move were discussed earlier and would have a disorienting affect on the child. The proportions of reported superficial injuries with respect to household moves were as follows; more than three moves(25%), three moves(22%), no moves(18%), two moves(18%) and one move(17%).

When looking at the 476 reported female foreign-body/ingestion injuries, we observed that these injuries were proportionally more frequent in females whose mother's were very young (less than 20) than for females born of older mothers. The results indicated that as mother's age increased so the proportion of reported foreign-body/ingestion type injuries decreased. The proportions of reported foreign body/ingestion injuries with respect to mother's age were as follows; less than 20 years old(21%), 20-30 years old(16%) and over 30 years old(11%).

A similar association was seen when looking at reported female foreign-body/ingestion type injuries with respect to the father's age. Reported foreign-body/ingestion type injuries were proportionally more frequent in females whose father's were very young (less

than 20) than for females with older fathers. The proportions of reported foreign body/ingestion injuries with respect to father's age were as follows; less than 20 years old(24%), 20-30 years old(16%) and greater than 30 years old(15%).

It is thought that younger parents in particular are less aware and less responsive to potential accident situations than older parents so it is not suprising that parents' age appears in the analysis. It is thought that mother's age and father's age are closely associated and that the effects might be confounded with the effects of other variables such as the size of the family, the neighbourhood or the standard of household equipment, parent's income level.

The same age-related phenomenon was also seen when looking at mother's age with respect to the 498 reported female torso injuries. The results indicated that females whose mothers were young (less than 20) had proportionally more reported torso injuries than females with older mothers. As mother's age increased the proportion of torso injuries decreased. The proportions of reported torso injuries with respect to mother's age were as follows; less than 20 years old(23%), 20-30 years old(16%) and greater than 30 years old(13%). This result is in line with the previous result with respect to foreign body/ingestion type injuries. For our purposes ingestions are considered to be injuries of the digestive system in the torso, hence the association between these

reported injuries and the torso.

We now examine the 274 reported female leg injuries in a breakdown with respect to the level of tidiness in the child's home. We observed that females from a household with an untidy/chaotic level of tidiness had proportionally more reported leg injuries than females from other homes. The proportions of reported leg injuries with respect to the tidiness of the household were as follows; untidy/chaotic (16%), over/very tidy(10%) and average(8%). This is quite a logical result since children in an untidy household have more objects to bump into and trip over than children in tidier households. It is suprising that this result was not present when examing the same breakdown in the male population. It was further observed that females from poor neighbourhoods had proportionally more reported leg injuries than females from other neighbourhoods. The proportions with respect to neighbourhood were as follows; poor(14%), average(10%), rural(8%) and well-to-do(7%).

Accident associations with respect to neighbourhood were mentioned earlier and the child's neighbourhood is thought to be closely associated with other variables. For example, a household in a poor neighbourhood is more likely to have poor quality floor coverings than a household in a better neighbourhood. Hence, a child falling in a household in a poor neighbourhood is more likely to injure itself than a child falling on a carpeted floor in a house in a better neighbourhood. This result

might also be linked to the previous result since a child in an untidy household may frequently trip over and fall onto the floor. These accidents would be less likely to result in injury in households with better and softer floor coverings. In addition, it was further observed that reported arm injuries were proportionally less frequent in females from poor neighbourhoods than for females from other neighbourhoods. The proportions of reported arm injuries for females from each neighbourhood were as follows; average(23%), rural(22%), well-to-do(19%) and poor(14%). This is an unusual result unless we assume that females from poor neighbourhoods had stronger arms, or stronger arm reflexes which helped them in accident situations. We could alternatively assume that female arm injuries were so common and frequent in females from the poor neighbourhoods that they were not recalled accurately, hence the possibility of accident misreporting.

We now compare the results across gender.

1.3.7 A SIMPLE COMPARISON OF THE REPORTED ACCIDENT ASSOCIATIONS BETWEEN GENDER

Having discussed the results summarised in the previous two sections we now look at the similarities and differences between gender when taking into account accident trends and associations in specific variables.

When comparing reported accident frequency trends and associations across gender we see that males and females have one particular result in common, i.e. that head injuries are proportionally more common in children in the most-advantaged social index group than in the other social index groups, particularly the lower social index groups.

It is thought that this peculiar result was mainly due to accident misreporting, i.e. that parents of children in the lower social index groups did not report child head injuries as accurately as parents in the most advantaged social index group. It seems extremely unlikely that children in the most advantaged social index group do actually have proportionally more head injuries than other children and it is possible that parents of children from the lower social index groups do not recognise the potential dangers of a child head injury. This peculiar result provides strong evidence that differential reporting exists across the population. This phenomena has implications for any kind of accident analysis and is investigated further in Sections 3.5, 5.2, 6.2 and 6.3.

Table 6 and Table 7 give the breakdown of the reported head accidents with respect to the type of accident that was described for the male and female populations, respectively.

Concentrating on the reported male head accidents only, shown in Table 6, we observe that males from the highest social index group (one), had proportionally more

reported superficial head accidents than males from any other social index grouping. From this we could assume that parents from social index one were perhaps more sensitive about their child's 'head accidents' than parents from the other social index groups. We also note that males from the lowest index grouping, group five, had proportionally more reported head accidents requiring stitches than males from the other social index groupings.

TABLE 6. A breakdown of all male reported head accidents with respect to social index. Social index category one is the most advantaged index group, two is the advantaged group, three is average, four is disadvantaged and five is the most disadvantaged index group. The percentages in brackets are in terms of the total number of reported head accidents that occurred in the male population.

	SOCIAL INDEX OF THE MALE				
	ONE	TWO	THREE	FOUR	FIVE
Fracture	74(34%)	218(34%)	297(34%)	112(33%)	78(33%)
Stitches	33(15%)	94(15%)	108(12%)	43(13%)	44(19%)
Disl/spr.	1 (0%)	1 (0%)	1 (0%)	0 (0%)	0 (0%)
Burn/scald	3 (1%)	9 (1%)	14 (2%)	8 (2%)	6 (3%)
Ingest/for.	13 (6%)	28 (4%)	52 (6%)	14 (4%)	11 (5%)
Superficial	37(17%)	80(13%)	125(14%)	44(13%)	27(11%)
Other	57(26%)	202(32%)	281(32%)	116(34%)	70(30%)
Total head accidents	218	632	878	337	236
Total accidents	387	1198	1648	667	488
Pop. size	621	1747	2331	945	656

We do not observe the same trends in Table 7 as seen in the male population. We see that females from the highest social index group (one), had proportionally fewer

reported head accidents requiring stitches than females from any other social index grouping. We also see that females from group one also had proportionally more reported ingestion/foreign body type accidents than females from the other social index groupings. Proportionally, females from group two had more head fractures than females from the other groups, and females from group four had more reported superficial head accidents than females from the other groups.

TABLE 7. A breakdown of all female reported head accidents with respect to social index. Social index category one is the most advantaged index group, two is the advantaged group, three is average, four is disadvantaged and five is the most disadvantaged index group. The percentages in brackets are in terms of the total number of reported head accidents that occurred in the female population.

	SOCIAL INDEX OF THE FEMALE				
	ONE	TWO	THREE	FOUR	FIVE
Fracture	38(27%)	119(31%)	133(25%)	57(29%)	37(25%)
Stitches	12 (8%)	48(13%)	60(11%)	19(10%)	17(12%)
Disl/spr.	1 (1%)	0 (0%)	2 (0%)	2 (1%)	0 (0%)
Burn/scald	4 (3%)	5 (1%)	13 (2%)	7 (4%)	5 (3%)
Ingest/for.	18(13%)	33 (9%)	60(11%)	7 (4%)	9 (6%)
Superficial	19(13%)	54(14%)	78(14%)	31(16%)	19(13%)
Other	51(36%)	123(32%)	193(36%)	76(38%)	60(41%)
Total head accidents	143	382	539	199	147
Total accidents	281	797	1146	484	325
Pop. size	558	1627	2146	868	598

From this we conclude that the characteristics of the male and female reported accident distributions are

certainly not the same and that misreporting trends seen in the male population may not be duplicated in the female population and vice versa.

It was found that some variables appeared in associations with both sexes but with different accident trends. For example, for both males and females the type of neighbourhood and the number of household moves appeared to be influential variables. In males, the neighbourhood was associated with a trend in reported lacerations which required stitches, and in females, neighbourhood appeared in a trend with respect to reported burns/scalds. In males, the number of household moves appeared in an association with reported accident resulting in injuries of the torso and for females, the number of household moves also appeared but in an association with reported superficial injuries.

In addition, variables appeared with respect to associations with only one sex. For example, associations were identified in the male population with respect to the number of older siblings and reported accidents which resulted in superficial injuries, also to the males personality and head injuries, i.e. whether the male fought others and whether the male was solitary. For females there were associations which involved the level of tidiness of the home and reported accidents which resulted in leg injuries. Also, parents' age in April 1970 were both in associations with reported foreign body/ingestion type injuries.

In summary, some reported accident trends were identified for which the most likely explanation is differential accident reporting. For example, reported accidents which resulted in superficial injuries and the child's neighbourhood. The results presented here in the preliminary, exploratory analysis also suggest that the male reported accident distribution is very different to the female reported accident distribution.

This misreporting phenomena has implications for the all the following analysis. We suspect that fitting the accident probability distributions directly to the reported accident distribution will produce poor fitting models unless the probability distributions are modified to allow for accident misreporting. This is investigated in Chapters 3 and 4.

It is believed that accidents which result in serious injury are recalled and reported more accurately than accidents which resulted in superficial injury. Hence, in Chapters 3 and 4 we also examined the cumulative affects of adding in successfully less serious accidents to the dependent variable. Since the implication for the regression analysis in Chapter 5 is that differential misreporting could produce erroneous regression models when modelling 'all reported accidents', this was also investigated in the regression modelling in Chapter 5.

During the first stage of the analysis we have attempted to identify as many deficiencies and inadequacies in the data set as possible. If we are aware

from the outset of what 'surprises' to expect then there would be less difficulty in interpreting the results.

It is generally suspected that in any accident or sickness study, the results are based on remembered or reported events rather than on all events that occurred (Newbold, 1927, and Stewart-Brown, Golding, Bijur and Butler, 1983) so underreporting of accidents is to be expected. From the preliminary analysis it has been discovered that specific types of accidents have been reported differently by different groups of parents. Several reported accident trends and associations were identified and high-risk sub-populations were highlighted. The accident information contained in the 1975 British Birth Cohort was collected retrospectively for the previous five years and in these circumstances it would be fair to assume that there is an underreporting of accidents that occurred during the five year period.

The next stage of the methodology is to model the reported accident distribution with the traditional accident modelling probability distributions, the Poisson and Negative Binomial probability distributions.

We now define the aims of the thesis.

1.4 THE AIMS OF THIS THESIS

In this Chapter we have focussed on previous accident analyses to highlight the problems already uncovered and

to examine the implications for this analysis. Not all areas of our analysis were fruitful. For example, there are several types of recall problems and it was found that any analysis based on the reported recalled times of the accidents was unreliable. Accident misreporting was also identified but was overcome by introducing 'misreporting' parameters into the probability distribution modelling.

The nature of the data allows us to examine the occurrence of different types of reported accidents in more detail. For example, we have information on the seriousness of each accident and can examine reporting thresholds and differential reporting across the population. It was found that under certain conditions, modelling probability distributions fitted the data well and not in other circumstances. For example, the Negative Binomial model fits the reported accident distribution of the number of male bone fractures very well, yet when reported accidents such as lacerations requiring stitches are added into the reported distribution, the model fails.

In addition we have information on the 'personality' of each child as assessed by its mother, and are therefore able to examine the proneness issue in more depth. For example, children with particular personality characteristics (such as destroying belongings, say) have higher reported accident means than children without those characteristics.

The objectives of this project are therefore to fit statistical models to the British Birth Cohort data of

1975 in order to identify and quantify the effect of variables influencing accident rates in children up to the age of five years with a view to prevention. Several methods of analysis are investigated and compared. The objectives can be summarised as follows.

- 1) Fitting probability distributions directly to the reported accident distribution and investigating the extension of these models to include accident misreporting.
- 2) To examine the major problem of accident misreporting and to identify the most reliable subset of the reported accident data.
- 3) Fitting regression models to investigate differential accident reporting using different reported accident outcome variables with a view to pinpointing methodology.
- 4) Comparing and contrasting the results from the different model fitting techniques.
- 5) To identify sub-populations of high-risk.
- 6) To explore the types and severity of injuries and whether high-risk sub-populations are more at risk to particular types and severity of injuries than other groups.

CHAPTER 2

PROBABILITY MODELS FOR THE REPORTED ACCIDENT DISTRIBUTION

In this Chapter we introduce the methodology and the probability distributions used to model the reported accident distribution in the later Chapters. The analysis was completed in several stages and we will briefly summarise each stage here before going into more detail in the rest of the Chapter.

2.1 INTRODUCTION

During the first stage of the analysis we discussed reported accident associations or trends and uncovered deficiencies in the data set which could affect the later stages of the analysis. In particular we discovered strong evidence of accident misreporting in Section 1.3.4.

The next stage of the analysis consisted of splitting the male and female populations into sub-populations with respect to a single variable at a time, then modelling the reported accident distribution from each sub-population using the traditional accident modelling probability distributions, the Poisson and Negative Binomial distributions. The gender split was important and necessary because the evidence from the preliminary

analysis suggested major differences between the male and female reported accident distributions (see Sections 1.3.3 to 1.3.7). In addition, by splitting the male and female populations into smaller sub-populations with respect to a single variable we hoped to be able to isolate the sub-populations in which misreporting was most evident (see Section 3.2).

The simple, one parameter Poisson probability distribution is traditionally the first model to be fitted in any accident study but it is sometimes observed that the sample variance is larger than the sample mean, that is, 'over-dispersion' is present. The next logical step in the modelling is therefore to use a model which allows for over-dispersion, such as the Negative Binomial distribution. This is a natural modelling progression which has often been used in the past (Arbous and Kerrich, 1951, and Adelstein, 1952). Hence the next stage of the accident analysis was the fitting of the two parameter Negative Binomial distribution to the full reported accident distribution.

Since the preliminary analysis highlighted several cases of accident misreporting it seemed unlikely that either of the traditional modelling distributions would provide a good fit to the full reported accident distribution. This view was supported by the results from the traditional modelling described in Section 3.1.

When modelling the reported accident distributions we compared the reported accident counts with the fitted accident counts to identify where the models failed. When

modelling the full reported accident distribution we found that the poor-fitting Negative Binomial models consistently overestimated the reported two accident count and underestimated the reported single accident count (see Section 3.1). This is evidence of misreporting when weighed against the successful modelling of accident data by the Negative Binomial in many previous studies and the theoretical explanation of the appropriateness of the model (see Section 1.2).

One way of catering for this specific type of accident misreporting is to generalise the probability distribution models by introducing a 'misreporting' parameter. Generalisations of the traditional probability distributions to include misreporting parameters are developed in Section 2.3. A similar approach is used to develop an alternative probability distribution in Section 2.4. The results from fitting these models are described in Section 3.8. Therefore, in this stage of the analysis, generalised traditional accident probability distributions were fitted to the full reported accident distribution to take into account accident misreporting between specific accident counts.

When the traditional accident probability distributions fail to provide an adequate model for the reported accident distribution there are several other probability distributions which can be used as alternatives. Four alternative models are fitted to the full reported accident distribution of each sub-population and these models are discussed in Section 2.4. The

'alternative' accident models produced similar goodness-of-fit results to the Negative Binomial models and hence only the best fitting alternative distribution was generalised to take into account specific accident misreporting. This generalised distribution is derived in Section 2.4. The results are given in Chapter 4, Section 4.2.

In addition to the four alternative accident modelling distributions the reported accident distribution could be truncated at zero and the truncated traditional accident probability distributions can be fitted to this reported accident distribution (see Section 2.5). The results of modelling the truncated reported accident distributions are given in Chapter 4.2 and justify the use of the generalised traditional distributions.

The hypothesis was put forward that there was a smaller chance of misreporting serious accidents because parents were more likely to remember the occurrence of such events. To investigate if such differential reporting had occurred the full reported accident distribution was broken down into smaller reported accident distributions according to the degree of 'seriousness' of each reported accident. The 'cumulative' reported accident distributions were then defined by adding in successively less serious reported accidents into the reported accident outcome variable. These cumulative distributions were then modelled with the traditional accident modelling distributions in Section 5.2. A good fit of a traditional accident model to a subset of data corresponding to

accidents of a particular degree of seriousness would suggest that misreporting is minimal, that is that unreliable data has been excluded. Section 5.5 addresses the issues of how much data could then be retained for analysis and the validity of conclusions which could be drawn.

We also searched for evidence of proneness by calculating the correlation coefficients between the numbers of accidents occurring in consecutive time periods. Three different sets of time periods were used and the results in Section 4.3 strongly suggest that accident times were misreported.

The final stage of the analysis was the regression modelling. The different dependent outcome variables and the two different types of regression modelling used are discussed later in this Chapter (see Section 2.6). The first approach involved the use of the GLIM statistical package and the results are given in Section 7.2. The second type of regression modelling involved the use of SAS to fit logistic regression models and the results are given in Section 7.3. The cumulative reported accident distributions were also modelled and it was found that different outcome variables and different regression techniques produced different regression models.

We now discuss the methodology of the analysis in more detail.

2.2 FITTING THE TRADITIONAL ACCIDENT DISTRIBUTIONS

- AN INVESTIGATIVE TOOL

The Poisson process is the simplest form of a continuous time stochastic process for describing the occurrence of single events over time. The underlying assumptions are that the accident rate L is constant over time which ensures that the process is time homogeneous. In addition, the numbers of accidents which occur in non-overlapping time intervals are independent, that is the accidents occur at random. If we also specify that $Y(t)$ is the number of events in a time interval of length t , then $Y(t)$ has a Poisson distribution with mean $L \times t$.

It is unrealistic to expect the Poisson distribution to be an appropriate model for the underlying reported accident distribution over the entire population. From the preliminary analysis we know that the assumption of a constant accident rate L for all children is unreasonable. For example, the parameter L obtained from the Poisson model fitted to the male population was 0.6957, and for the females was 0.5208 (see Table 5, Section 1.3.3).

A natural progression from this single parameter model is to assume that different sub-populations are described by distinct and separate, independent Poisson models. This would mean that the overall distribution would be a mixture of several Poisson distributions, each with its own parameter L_1 , L_2 , etc.. For instance, the males could have a different accident rate from the females, or

children whose fathers are manual workers could have a different accident rate from children whose fathers are non-manual workers.

Suppose there are several sub-populations such that the i th sub-population has a Poisson distribution for the number of reported accidents in time t with parameter λ_i , and suppose the parameters λ_i are random variables with a Gamma distribution. If the reported accidents are independent from each other then the probability distribution for the number of reported accidents $Y(t)$, in time t is the Negative Binomial distribution (Greenwood and Yule, 1920), originally called Greenwood and Yule's model of unequal liabilities. It has been shown by Feller (1943) that the Negative Binomial distribution can also be obtained by assuming that accidents or events are not independent of each other, i.e. by assuming that a learning process exists which depends on the number of previous accidents.

An alternative generalisation of the Poisson distribution is to use a Gamma distribution for the parameters λ_i , giving a Neyman Type A distribution for $Y(t)$. The Neyman Type A distribution has been used successfully in the past as an alternative to the Negative Binomial distribution and has often been used to describe models of heterogeneity (Johnson and Kotz, 1969). The method we used to obtain the parameter estimates of the Negative Binomial distribution can be found in Johnson and Kotz (1969, p. 131) and is described as follows;

Consider the Negative Binomial distribution in the

following form;

$$P(k) = \frac{(N+k-1)! \times (P)^k \times (1+P)^{-(N+k)}}{(N-1)! \times k!}, \quad (2.1)$$

$$k = 0, 1, 2, \dots,$$

where, N is the parameter to be estimated and $P=M/N$.

The simplest way of obtaining the parameters N and P (where $P=M/N$) is the method of moments, that is equating the sample mean and variance to the corresponding population values. Thus if $y_1, y_2, \dots, y_n$ are the n observed and independent values then we calculate the solutions of N and P from the equations,

$$N \times P = \bar{y}, \quad N \times P \times (1+P) = s^2,$$

where $\bar{y}$ is the mean number of events and s^2 the sample variance.

$$\text{Thus, } P = s^2 / \bar{y}^2 - 1, \quad \text{and, } N = \bar{y}^2 / (s^2 - \bar{y}^2).$$

Note that P is negative if $s^2 < \bar{y}^2$, hence if this occurs the Negative Binomial distribution is inappropriate.

In place of $N \times P \times (1+P) = s^2$, an equation obtained by equating the observed and expected zero frequencies among the y 's may be used. This equation is,

$$f_0 = (1+P)^{-N},$$

where f_0 is the proportion of zero values in the sample.

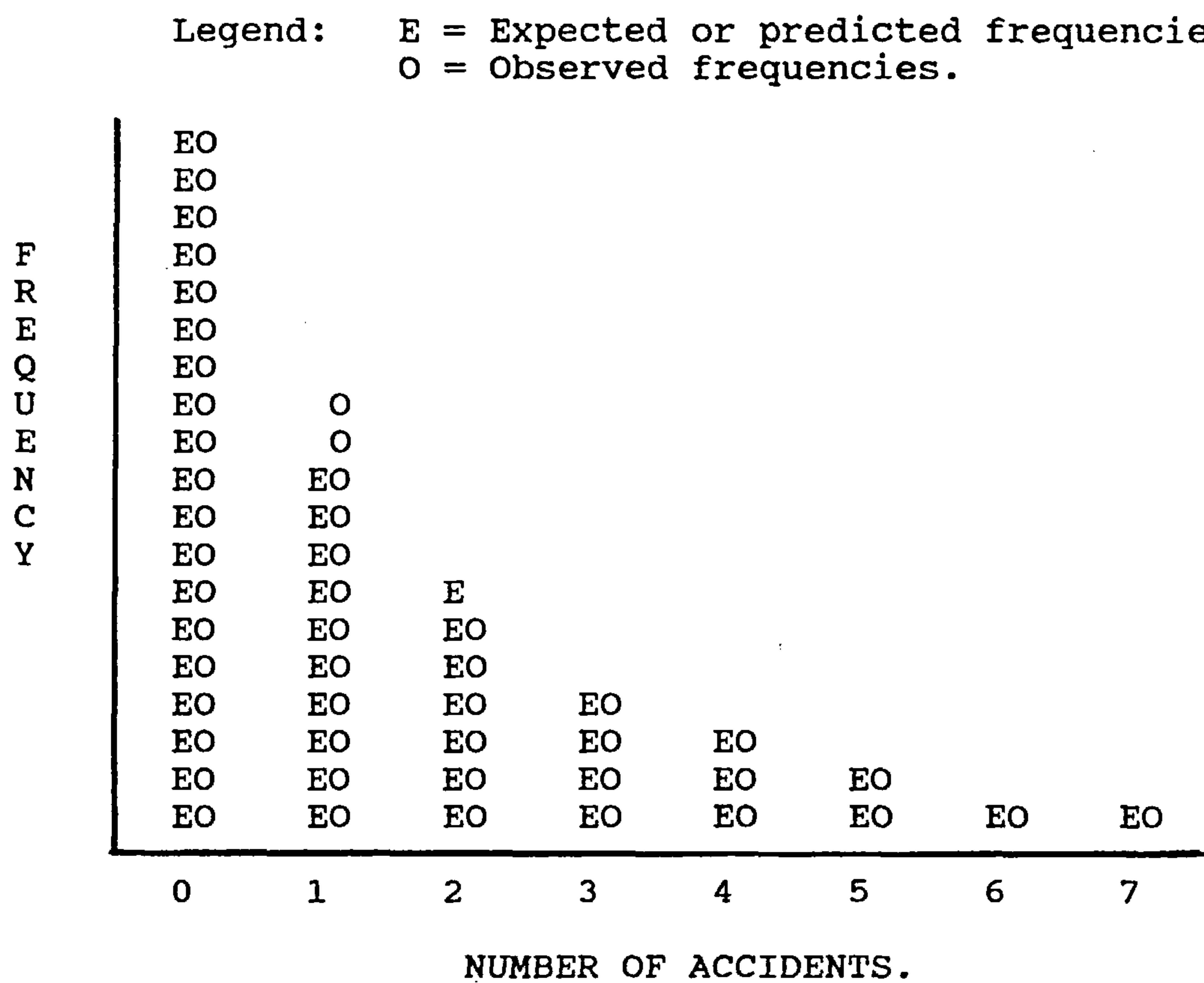
Combining this equation with the equation $N \times P = \bar{y}$, gives,

$$\left(\frac{P}{\log(1+P)} \right) = \left(\frac{\bar{y}}{-\log(f_0)} \right).$$

This equation has to be solved iteratively for P , and provided that $\bar{y} > -\log(f_0)$, there is always a unique solution for P .

Accident misreporting had already been identified (see Section 1.3.3) so it was anticipated that neither the Poisson nor the Negative Binomial distribution would adequately describe the reported accident distribution. This was confirmed by fitting the traditional models described in Section 3.1. However, fitting these models was useful for detecting further evidence of misreporting in two ways. Firstly, in looking for trends in misreporting we examine Figure 12 which shows a typical comparison of the reported accident counts and the predicted accident counts when a Negative Binomial model is fitted to a sub-population. In this example the sub-population is the male population. The model is a poor fit and the figure suggests a trend in misreporting from the two accident count to the one accident count. This indicates that two accidents were sometimes reported as one accident. No similar trend was seen when the predicted counts from the poor fitting Poisson models were similarly examined.

FIGURE 12. An illustration of what is generally seen when fitting the Negative Binomial distribution to a sub-population.



A second role for fitting the traditional models is in investigating if accidents which result in serious injury are recalled and reported more accurately than accidents which result in superficial injury. We therefore examined the cumulative effects of adding in successively less serious accidents to the dependent variable.

As the misreporting of specific types of accidents had been identified in the preliminary analysis, we examined the 'serious' reported accident distribution in some detail. The males and females were split into sub-populations as before and the traditional modelling was repeated.

The results from fitting the Poisson and Negative Binomial distributions to the serious reported accident distribution provided evidence of unusual reporting phenomena. For example, in the poor fitting Negative Binomial models the model usually overestimated the two accident count and underestimated the tail of the distribution. This suggests accident misreporting from the lower serious accident counts to higher serious accident counts. Since we assume that misreporting can occur only in the opposite direction, i.e. from higher reported accident counts to lower reported accident counts, the serious reported accident distribution as we have defined must be revised.

The reported accident distribution was therefore redefined to include all accidents except those resulting in either superficial or unidentified injuries. This distribution was called the 'non-superficial' accident distribution and lay between the full and the serious reported accident distributions. The full reported accident distribution therefore comprises of non-superficial reported accidents and superficial/unidentified reported accidents.

The Poisson and Negative Binomial distributions were fitted to both the non-superficial and superficial/unidentified reported accident distributions of the male and female sub-populations separately and the Negative Binomial model produced excellent Chi-squared goodness-of-fit results in each case. For example, for the non-superficial reported accident distribution the

Negative Binomial distribution provided an adequate model at the one percent significance level for 70 out of 73 male sub-populations and 58 out of 62 female sub-populations.

2.3 EXTENSIONS OF PROBABILITY MODELS TO INCLUDE MISREPORTING PARAMETERS

Another method of taking into account accident misreporting is to generalise a probability distribution in order to take into account the misreporting of particular accident counts. For example, in a series of accidents one or more accidents may be forgotten or not considered serious enough in relation to the other accidents that occurred, i.e. an accident was not regarded or remembered as an accident. Hence, one accident might be reported as no accident, two accidents reported as one accident, three accidents reported as two accidents, and so on. A second example is when parents are unable to recall any accidents at all, i.e. there is a heaping of children with zero accidents. Hence, we would then get misreporting from the one accident count to zero accidents and from two accidents to zero accidents, from three accidents to zero accidents, and so on. We therefore investigate generalisations of the traditional probability distributions.

From the earlier examination of the poor fitting Negative Binomial models in Figure 18 in the previous

section we speculated that the poor fit was due a trend in misreporting from the two accident count to the one accident count. No similar trend was seen when the predicted counts from the poor fitting Poisson models were similarly examined. Hence, in the third stage of the analysis the Poisson and Negative Binomial distributions were generalised to take into account accident misreporting between specific accident counts. The rest of this section will describe each of the generalised models used in Section 3.2. For each model discussed we use the maximum likelihood method for estimating the parameters and solve the respective equations using iterative procedures. The programs to do this are written in FORTRAN.

Three Generalised Poisson models were developed sequentially to investigate this approach. Firstly we investigated accident misreporting from one to zero accidents, then in a separate model we investigated accident misreporting from two to one accidents. In the third model we investigated accident misreporting from two to one accidents together with misreporting from one to zero accidents, i.e. we estimated two accident misreporting parameters. The third model was further extended by using one parameter to estimate all the accident misreporting and hence two more Generalised Poisson models were developed. These two Generalised Poisson models only contained two parameters which makes them easier to interpret and computationally easier to estimate.



When defining the generalised distributions we assumed that misreporting occurred in only one direction, i.e. two accidents could be misreported as one accident and one accident could be misreported as no accident. To do otherwise would mean that an imaginary accident was reported when in fact no accident occurred.

2.3.1 GENERALISED POISSON DISTRIBUTIONS

In this section we formally develop the Generalised Poisson distributions which were used in the distribution modelling of the accident data.

The first Generalised Poisson distribution that we develop allows for misreporting from the reported one accident count to the reported zero accident count. Rao and Rubin (1964) have investigated this particular Generalised Poisson distribution when they examined errors in recording a variable which in reality did have a Poisson distribution. In the case where there is misreporting from the one accident count to the zero accident count the distribution is defined as follows;

$$P(0) = (1 + BxL) \times \exp(-L),$$

$$P(1) = L \times (1 - B) \times \exp(-L),$$

$$P(j) = L^j \times \exp(-L)/j!, \quad \text{for } j > 1,$$

where L is the usual Poisson parameter and B is the parameter which describes the accident misreporting from

one accident to no accidents.

This distribution is called Model I and In terms of our reported accident data it describes a situation where there is an overreporting of zero accidents that should have been reported as single accidents, i.e. a single accident is overlooked or misreported as no accident when in reality an accident did occur. We shall also assume the rest of the reported accident distribution was reported correctly.

Cohen (1960) obtained the following maximum likelihood estimators of L and B respectively.

$$L = 0.5 \times [\bar{y}-1 + f_0 + \{(\bar{y} - 1 + f_0)^2 + 4 \times (\bar{y} - f_1)\}^{0.5}],$$

$$B = \left(\frac{(f_0 - f_1/L)}{(f_0 + f_1)} \right),$$

where $\bar{y}$ is the sample mean, f_0 and f_1 the proportions of the reported counts equal to zero and one respectively, and B is the parameter of misreporting from one accident to no accident.

To expand upon this concept of misreporting we again generalise the simple Poisson distribution but this time allow for the mis-reporting of two accidents as one accident. We assume that the rest of the reported accident distribution was reported correctly. This distribution was also investigated by Rao and Rubin (1964) and is defined as follows;

$$P(0) = \exp(-L),$$

$$P(1) = L \times (1 + \theta x L/2) \times \exp(-L),$$

$$P(2) = (L \times L/2) \times (1 - \theta) \times \exp(-L),$$

$$P(j) = L^j \times \exp(-L)/j!, \quad \text{for } j > 2,$$

where L is the usual Poisson parameter and θ the parameter of misreporting from two accidents to one accident.

This is called Model II and the maximum likelihood estimators of L and θ are now

$$L = 0.5 \times [(\bar{y} - 2 + f_1) + \{ (\bar{y} - 2 + f_1)^2 + 8 \times (\bar{y} - f_2) \}^{0.5}],$$

$$\theta = \left(\frac{(f_1 - 2 \times f_2 / L)}{(f_1 + f_2)} \right),$$

where $\bar{y}$ is the sample mean, f_1 and f_2 the proportions of the reported counts equal to one and two respectively.

We have made a further generalisation of the Poisson distribution by extending the concept of misreporting to include the misreporting of one accident as no accident, together with the misreporting of two accidents as one accident. We again assume that the rest of the reported accident distribution was reported correctly. This is Model III and is defined as follows.

$$P(0) = (1 + L \times B) \times \exp(-L),$$

$$P(1) = L \times (1 - B + \theta \times L/2) \times \exp(-L),$$

$$P(2) = (L^2 / 2) \times (1 - \theta) \times \exp(-L),$$

$$P(j) = L^j \times \exp(-L)/j!, \quad (j > 2),$$

where L is the usual Poisson parameter,
 θ is the parameter of misreporting from two to one accident and
 B is the parameter of misreporting from one to zero accidents.

In terms of our accident data this distribution would describe a situation where there is an overreporting of single accidents some of which should have been reported as two accidents, and an overreporting of zero accidents some of which should have been reported as one accident. It does not allow for the misreporting of two accidents as zero accidents and we also assume the rest of the reported accident distribution was reported accurately. The log likelihood function for this distribution is as follows.

$$n_0 \times \log(1 + BxL) + n_1 \times \log(1 - B + \theta xL/2) + n_2 \times \log(1 - \theta) \\ - TxL + K \times \log(L) + \log\left(\prod_k 1/k!\right),$$

where K = the sum of the T observations,
 T = the total number of observations,
 n_0 = the number reported to have zero accident,
 n_1 = the number reported to have one accident,
 n_2 = the number reported to have two accidents,
and L , B and θ are the parameters to be estimated.

To obtain maximum likelihood estimators for L , B and θ we partially differentiate the log-likelihood function with respect to the parameters and solve for the following three simultaneous equations.

$$\text{i) } \left(\frac{n_0 x L}{(1 + BxL)} \right) - \left(\frac{n_1}{(1 - B + \theta xL/2)} \right) = 0,$$

$$\text{ii) } \left(\frac{n_0 x B}{(1 + BxL)} \right) + \left(\frac{0.5 x n_1 x \theta}{(1 - B + 0.5 x \theta x L)} \right) - T + K/L = 0,$$

$$\text{iii)} \left(\frac{0.5 \times n_1 \times L}{(1 - B + 0.5 \times \theta \times L)} \right) - \left(\frac{n_2}{(1 - \theta)} \right) = 0,$$

where T = the total number of observations,
 K = the sum of the T observations,
 n_0 = the number reported to have zero accidents,
 n_1 = the number reported to have one accident,
 n_2 = the number reported to have two accidents, (2.2)
and L , B and θ are the parameters to be estimated.

The next logical step in the model development is to define a model which uses a single parameter to estimate the misreporting from one to no accident and also the misreporting from two to one accident. This has the advantage of using fewer degrees of freedom in the Chi-squared goodness-of-fit test and may perhaps improve the goodness-of-fit of the model.

After examining the estimates for B and θ obtained from Model III (the Generalised Poisson distribution with misreporting from one to zero accident together with misreporting from two to one accident) it was decided to define B in two different ways. The first was to define B as follows;

$$B = a \times \theta,$$

where B is the parameter of misreporting from one to zero accident, θ is the parameter of misreporting from two to one accident, and a is a constant.

In terms of our accident data this distribution describes a situation where there is an overreporting of both single accidents which should have been reported as two accidents, and of zero accidents which should have been reported as one accident. It does not allow for the

misreporting of two accidents as zero accidents. We also assume accurate reporting of the rest of the reported accident distribution. This model will be called Model IV and is defined as follows.

$$P(0) = (1 + ax\theta xL) \times \exp(-L),$$

$$P(1) = L \times [1 - \theta x(a - L/2)] \times \exp(-L),$$

$$P(2) = (L^2/2) \times (1 - \theta) \times \exp(-L),$$

$$P(j) = L^j \times \exp(-L)/j!, \quad \text{for } j > 2,$$

where L is the usual Poisson parameter,
 θ is the parameter of misreporting from two to one accident, and
 a is a known constant.

The log-likelihood of this distribution is:

$$\begin{aligned} & n_0 \times \log(1 + ax\theta xL) + n_1 \times \log(1 + \theta x(L/2 - a)) \\ & + n_2 \times \log(1 - \theta) - TxL + K \times \log(L) + \log \prod (1/k!) \end{aligned}$$

where

K = the sum of the T observations,
 T = the total number of observations,
 n_0 = the number reported to have zero accident,
 n_1 = the number reported to have one accident,
 n_2 = the number reported to have two accidents, (2.3)
 a is a known constant and L and θ are the parameters to be estimated.

After differentiating the log-likelihood with respect to the parameters we obtain maximum likelihood estimators for L and θ from solving the following simultaneous equations;

$$i) \quad \left(\frac{\theta x a x n_0}{1 + ax\theta xL} \right) + \left(\frac{n_1 x \theta x L/2}{1 + \theta x(L/2 - a)} \right) - T + K/L = 0,$$

$$ii) \left(\frac{n_0 x a x L}{1 + a x \theta x L} \right) + \left(\frac{n_1 x (L/2 - a)}{1 + \theta x (L/2 - a)} \right) - \left(\frac{n_2}{1 - \theta} \right) = 0.$$

Where L and θ are the parameters to be estimated.

We now consider an alternative relationship between B and θ which may be applicable when both types of misreporting occur, namely:

$$B = \theta^a,$$

This distribution is called Model V and has the same assumptions as Model IV. The model is defined as follows:

$$P(0) = (1 + \theta^a x L) x \exp(-L),$$

$$P(1) = L x [1 - \theta x (\theta^{a-1} - L/2)] x \exp(-L),$$

$$P(2) = (L^2 / 2) x (1 - \theta) x \exp(-L),$$

$$P(j) = L^j x \exp(-L) / j!, \quad \text{for } j > 2,$$

where L is the usual Poisson parameter, θ is the parameter of misreporting from two to one accident, and a is a constant.

The log-likelihood of this distribution is as follows.

$$\begin{aligned} & n_0 x \log(1 + \theta^a x L) + n_1 x \log(1 + \theta x (L/2 - \theta^{a-1})) \\ & + n_2 x \log(1 - \theta) - T x L + K x \log(L) + \log \prod (1/k!) \end{aligned}$$

where L and θ are the parameters to be estimated.

After differentiating the log-likelihood with respect to the parameters we obtain maximum likelihood estimators

for L and θ from solving the following simultaneous equations.

$$\begin{aligned}
 \text{i)} \quad & \left(\frac{\theta^a \times n_0}{1 + \theta^a \times L} \right) + \left(\frac{n_1 \times (\theta/2)}{1 + \theta \times (L/2 - \theta^{a-1})} \right) - T + K/L = 0, \\
 \text{ii)} \quad & \left(\frac{n_0 \times L \times \theta^{a-1}}{1 + \theta^a \times L} \right) + \left(\frac{n_1 \times (L/2 + \theta^{a-1})}{1 + \theta \times (L/2 - \theta^{a-1})} \right) - \left(\frac{n_2}{1 - \theta} \right) = 0.
 \end{aligned}$$

2.3.2 GENERALISED NEGATIVE BINOMIAL DISTRIBUTIONS

Moving on from the Poisson distribution we now generalise the Negative Binomial distribution. If we use the Negative Binomial distribution in the form already expressed in Section 2.1, where N and P are the parameters to be estimated, we again introduce the parameter θ , the parameter of accident misreporting from two to one accident. Assuming that misreporting occurs from two accidents to one accident only and the rest of the reported accident distribution was reported accurately we obtain the following distribution.

$$P(0) = (1 + P)^{-N}$$

$$P(1) = N \times P \times (1 + P)^{-(N+1)} \times \left(1 + \frac{P \times \theta \times (N + 1)}{2 \times (1 + P)} \right),$$

$$P(2) = \frac{N \times (N + 1) \times P^2 \times (1 - \theta) \times (1 + P)^{-(N+2)}}{2!},$$

$$P(j) = \frac{(N + j - 1)! \times P^j \times (1 + P)^{-(N+j)}}{(N - 1)! \times j!}, \quad \text{for } j > 2,$$

This is called Model VI and the log-likelihood of this distribution is as follows.

$$n1 \times \log[2x(1 + P) + \theta x P x (N + 1)] - n1 \times \log[2x(1 + P)]$$

$$+ n2 \times \log(1 - \theta) + K \times \log(P) - (TxN + K) \times \log(1 + P)$$

$$+ \sum_{j=1} \log \frac{(N+j-1)!}{(N-1)! \times j!},$$

where K, T, n1, n2, are defined in (2.2) and N, P and θ are the parameters to be estimated.

After the partial differentiation of this function with respect to the three parameters we obtain the maximum by solving for the following three simultaneous equations.

$$i) \left(\frac{n1 \times P \times (N + 1)}{2x(1 + p) + \theta x P x (N + 1)} \right) - \left(\frac{n2}{(1-\theta)} \right) = 0,$$

$$ii) \left(\frac{n1 \times (2 + \theta \times (n + 1))}{2x(1 + P) - \theta x P x (N + 1)} \right) + \left(\frac{K}{P} \right) - \left(\frac{(TxN + K + n1)}{(1 + P)} \right) = 0,$$

$$iii) \left(\frac{n1 \times \theta \times P}{2x(1 + P) + \theta x P x (N + 1)} \right) - T \times \log(1 + P)$$

$$+ \sum_{z=1} \sum_{j=1} (N + j - 1)^{-1} = 0.$$

No further generalisation of the Negative Binomial distribution was attempted since it was found that Model VI gave favourable results and any further generalising of the Negative Binomial distribution would not be necessary.

If the traditional accident probability distributions fail to provide an adequate model for the reported accident distribution then there are several other probability distributions which can be used as alternatives. We therefore discuss the next stage in the methodology which was the fitting of alternative probability distributions to the full reported accident distribution.

2.4 ALTERNATIVE PROBABILITY DISTRIBUTIONS AND EXTENSIONS

There are several alternative probability distributions available which have often been used in the past to model accident data when the Poisson or Negative Binomial distributions have failed to provide an adequate model. Stage four of our analysis was therefore the fitting of four alternative accident probability distributions to the full reported accident distribution. These were Neyman's Type A distribution (1939), the Polya-Eggenberger distribution (1956), the Generalised Poisson distribution of Consul and Jain (1973), and the compound Poisson Binomial distribution (1957).

In terms of goodness-of-fit these four alternative distributions produced models which performed better than

the traditonal Poisson model but no better than the Negative Bionomial model (see Section 4.2). Hence, only the best fitting of the alternative models was generalised to take into account specific accident misreporting. The rest of this section introduces each of the alternative models in turn.

2.4.1 THE NEYMAN TYPE A DISTRIBUTION

The Neyman Type A distribution was one of the earliest models of 'contagion'. It was first described by Neyman (1939) where it was used to define a model of the distribution of larvae in a randomly chosen area of a field. He assumed that the number of groups of eggs per area could be represented by a Poisson distribution, while the number of larvae developing from a group could be represented by independent variables, each having a common Poisson distribution. This is generally now considered as a model representing heterogeneity rather than a true 'contagion' model. A true contagion model is one where events being observed depend on the previous pattern of occurrences, e.g. the Polya-Eggenberger (1923) model. The Neyman Type A distribution can be obtained from compounding a Poisson distribution with a Poisson distribution and is defined as follows.

$$P(k) = \sum_{j=1}^{\infty} \exp(-L_1) \times (L_1^j / j!) \times \exp(-j \times L_2) \times (j \times L_2)^k / k!,$$

$k > 0,$

and,

$$P(0) = \exp(-L1) \times \sum_{j=1}^{\infty} \exp(-L1) \times (L1^j / j!) \times \exp(-j \times L2),$$

$$= \exp\{-L1 \times [1 - \exp(-L2)]\}.$$

The maximum likelihood estimation of the two parameters $L1$ and $L2$ can be quite complicated in form and estimators based on equating the first and second sample and population moments are frequently used. These estimators are as follows.

$$L2 = (s^2 - \bar{y}) / \bar{y},$$

$$L1 = \bar{y} / L2.$$

2.4.2 THE POLYA-EGGENBERGER DISTRIBUTION

We now discuss the Polya-Eggenberger distribution (1923) which is considered a model of true contagion with the underlying assumption that an event is not independent of previous events. The distribution was originally described in terms of random drawings of coloured balls from an urn. A ball is drawn at random and then returned to the urn with more balls of the same colour. Hence events are not independent and a contagion is established. In terms of accidents this can be applied so that an individual who has had no accidents is less likely to have an accident than an individual who has had more

and,

$$P(0) = \exp(-L1) \times \sum_{j=1}^{\infty} \exp(-L1) \times (L1^j / j!) \times \exp(-j \times L2),$$

$$= \exp\{-L1 \times [1 - \exp(-L2)]\}.$$

The maximum likelihood estimation of the two parameters $L1$ and $L2$ can be quite complicated in form and estimators based on equating the first and second sample and population moments are frequently used. These estimators are as follows.

$$L2 = (s^2 - \bar{y}) / \bar{y},$$

$$L1 = \bar{y} / L2.$$

2.4.2 THE POLYA-EGGENBERGER DISTRIBUTION

We now discuss the Polya-Eggenberger distribution (1923) which is considered a model of true contagion with the underlying assumption that an event is not independent of previous events. The distribution was originally described in terms of random drawings of coloured balls from an urn. A ball is drawn at random and then returned to the urn with more balls of the same colour. Hence events are not independent and a contagion is established. In terms of accidents this can be applied so that an individual who has had no accidents is less likely to have an accident than an individual who has had more

accidents since the more accidents that you have, the more you are likely to have.

In generalising the Hypergeometric distribution, Kemp and Kemp (1956) obtained the Polya-Eggenberger distribution and defined it as follows.

$$P(y=k) = \frac{\left(\frac{(-1)!}{(-1-k)! \times k!} \right) \times \left(\frac{(-m)!}{(n-k)! \times (-m-n+k)!} \right)}{\left(\frac{(-1-m)!}{(-1-m-k)! \times n!} \right)},$$

for $k = 0, 1, 2, 3, \dots, n,$

where l and m are the parameters to be estimated, and using extended factorial notation we define for real $a < 0$, and real $b < 0$ with b an integer.

$$\left(\frac{a!}{(a+b)!} \right) = (-1)^b \times \left(\frac{(-a-b-1)!}{(-a-1)!} \right) = (-1)^b \times \left(\frac{\Gamma(-a-b)}{\Gamma(-a)} \right)$$

The Negative Binomial distribution is a limiting case of the Polya-Eggenberger distribution.

Kemp and Kemp (1956) calculated the maximum likelihood estimators for l and m through iteration, using their moment estimators as the initial estimates. The moment estimators of l and m are as follows;

$$l = \left(\frac{(n - M - V/M)}{(V/M + M/n - 1)} \right) \times (M/n),$$

$$m = \left(\frac{(n - M - V/M)}{(V/M + M/n - 1)} \right) \times (1 - M/n),$$

where M is the observed mean and V is the observed variance.

The maximum likelihood equations are derived in the

usual way and become;

$$EQ1 = \sum_{k=0}^{n-1} \left(\frac{A(k)}{1+k} \right) - A(-1) \sum_{k=0}^{n-1} \left(\frac{1}{1+m+k} \right) = 0,$$

$$EQ2 = \sum_{k=0}^{n-1} \left(\frac{B(k)}{m+k} \right) - A(-1) \sum_{k=0}^{n-1} \left(\frac{1}{1+m+k} \right) = 0.$$

where $A(k)$ = number of observations $\geq k + 1$,
and $B(k)$ = number of observations $\leq n - k - 1$.

Hence, EQ1 and EQ2 have no explicit solution for l and m and iteration is necessary. Given the initial moment estimates l_1 and l_2 we obtain the values for EQ1 and EQ2 and obtain better estimates for l and m by solving for the following equations;

$$EQ1 = (l_2 - l_1) \sum_{k=0}^{n-1} \frac{A(k)}{(l_1 + k)^2} - (l_2 - l_1 + m_2 - m_1) \times A(-1) \sum_{k=0}^{n-1} \frac{1}{(l_1 + m_1 + k)^2},$$

$$EQ2 = (m_2 - m_1) \sum_{k=0}^{n-1} \frac{B(k)}{(m_1 + k)^2} - (l_2 - l_1 + m_2 - m_1) \times A(-1) \sum_{k=0}^{n-1} \frac{1}{(l_1 + m_1 + k)^2},$$

where l_2 and m_2 are the second estimates used in the iterations.

2.4.3 THE CONSUL/JAIN DISTRIBUTION

We now examine the distribution introduced by Consul and Jain (1973) which is a generalisation of the Poisson distribution with two parameters L_1 , and L_2 . It can be

obtained as a limiting form of the Generalised Negative Binomial distribution. The variance of the distribution can be less than, equal to or greater than the mean according to whether L_2 is negative, zero or positive. The distribution gives a very close fit to both Poisson data and Negative Binomial data. To derive the distribution Consul and Jain began with a Generalised Negative Binomial distribution of the following form:

$$P(y) = \left(\frac{n \times \text{Gamma}(n + Bxy)}{y! \times \text{Gamma}(n + Bxy - y + 1)} \right) \times a^y \times (1 - a)^{(n+Bxy-y)},$$

where, $1 > a > 0$, $B \geq 0$, $n > 0$, $|aB| < 1$, and B need not be an integer.

The following conditions were imposed: n and B are large, a is very small (such that $nxa = L_1$, $axB = L_2$, where L_1 is finite and positive and $|L_2| < 1$). This Generalised Negative Binomial distribution can be approximated using James Stirling's formula on the two Gamma functions and can be simplified to the following form;

$$L_1 \times (L_1 + yxL_2)^{y-1} \times \exp[-1 \times (L_1 + y \times L_2)] / y!,$$

which can be considered a Poisson type distribution. Consul and Jain therefore defined the distribution as:

$$P(y) = L_1 \times (L_1 + yxL_2)^{y-1} \times \exp[-1 \times (L_1 + y \times L_2)] / y!,$$

for $y = 0, 1, 2, 3, \dots$,

such that $P(y) = 0$ for $y \geq m$ if $L_1 + mxL_2 \leq 0$, (where $m = \text{mean}$).

Using the moment method of parameter estimation the parameter L2 can easily be estimated using the sample mean and variance. Hence,

$$L2 = 1 - (\text{mean}/\text{variance})^{0.5},$$

and,

$$L1 = \text{mean} \times (1 - L2).$$

2.4.4 COMPOUNDING THE POISSON DISTRIBUTION WITH THE BINOMIAL DISTRIBUTION

A further distribution which has been used successfully in the past as an alternative to the Negative Binomial distribution was first used by McGuire, Brindley and Bancroft (1957) to represent variation in numbers of corn-borer larvae in a randomly chosen area. This distribution has two parameters and is obtained by compounding the Poisson distribution with the Binomial distribution.

The compounded Poisson-Binomial distribution is defined as follows.

$$P(k) = \exp(-l) \times \sum_{j=0}^k \frac{(nxj)!}{(nxj-k)! \times j!} \times p^k \times q^{(nxj-k)} \times (1/j!),$$

where $(p+q)=1$, n is the largest observed count, and l and p are the parameters to be estimated.

We have used Katti and Gurland's (1962) method to estimate the parameters l and p for ease of calculation.

They use the first two moments with the observed first frequency. This approach gives

$$n \times l \times p = \bar{y},$$

$$\exp[-l \times (1 - q^n)] = f_0.$$

where, $(p+q)=1$, n is the largest observed count, f_0 is the proportion of zeros in the population and l , p are the parameters to be estimated.

From these two equations we get the following equation which is solved by iteration,

$$\bar{y} / [-\log(f_0)] = n \times p / (1 - q^n)$$

2.4.5 GENERALISATION OF THE CONSUL/JAIN DISTRIBUTION

Modelling the reported accident distributions obtained from splitting the data with respect to a single factor at a time, the best fitting of these four alternative models was found to be the Consul/Jain distribution (see Section 4.2). The poor fitting Consul/Jain models failed in the two and one accident counts so the distribution was generalised to take into account the misreporting of two accidents as one accident. The generalised distribution is defined as follows.

$$P(0) = \exp(-L_1),$$

$$P(1) = L_1 \times \exp[-(L_1+L_2)] \times [1+0.5 \times \theta \times (L_1+2 \times L_2) \times \exp(-L_2)],$$

$$P(2) = L_1 \times (L_1 + 2 \times L_2) \times \exp[-(L_1 + 2 \times L_2)] \times (1 - \theta)/2,$$

$$P(y) = L1 \times (L2 + y \times L2)^{y-1} \times \exp[-1 \times (L1 + y \times L2)] / y!,$$

for $y = 3, \dots,$

where $L1$, $L2$ and θ are the parameters to be estimated and θ is the parameter of misreporting between two and one accident.

The log-likelihood equation is as follows;

$$\begin{aligned} & -L1 \times T + (T - n0) \times \log(L1) - K \times L2 \\ & + n1 \times \log[1 + 0.5 \times \theta \times (L1 + 2 \times L2) \times \exp(-L2)] \\ & + n2 \times \log(1 - \theta) + \sum_{y=2}^{\infty} (y - 1) \times \log(L1 + y \times L2) \\ & - \log\left(\prod_{y=2}^{\infty} y!\right), \end{aligned}$$

where K , T , $n0$, $n1$, $n2$ are defined in (2.3) and L , B and θ are the parameters to be estimated.

To obtain maximum likelihood estimators for $L1$, $L2$ and θ we partially differentiate the log-likelihood function with respect to the parameters and solve the following three simultaneous equations.

$$\begin{aligned} \text{i)} \quad & \theta \times (n1 + n2) \times ZED - n1 \times ZED + n2 = 0, \\ \text{ii)} \quad & -T + (T - n0)/L1 + 0.5 \times n1 \times \theta \times \left(\frac{\exp(-L2)}{(1 + \theta \times ZED)} \right) \\ & + \sum_{y=2}^{\infty} \left(\frac{(y - 1)}{(L1 + y \times L2)} \right) = 0, \\ \text{iii)} \quad & -K + n1 \times \theta \times (1 - L2 - L1/2) \times \left(\frac{\exp(-L2)}{(1 + \theta \times ZED)} \right) \end{aligned}$$

$$\sum_{y=1}^{\infty} \left(\frac{(y-1) \times y}{(L1 + y \times L2)} \right) = 0,$$

where $ZED = 0.5 \times (L1 + 2 \times L2) \times \exp(-L2)$.

In addition to fitting the alternative probability distributions, the reported accident distribution was truncated at zero and the truncated Poisson and truncated Negative Binomial distributions were fitted to the shortened accident distribution. This is analagous to analysing accident data collected from a hospital casualty department where the zero accident count is not collected. The models considered are described in the following section.

2.5 TRUNCATING THE TRADITIONAL DISTRIBUTIONS

A particular example of accident misreporting is when parents report that no accidents occurred when in reality one or more accidents did occur. This leads to a 'heaping' of frequencies at the zero reported accident count. A simple way to incorporate this phenomenon into a model is to modify a distribution say the Poisson distribution, by adding an arbitrary proportion of zeros to the zero count and decreasing the remaining counts in an appropriate constant ratio. This is called a 'Poisson with zeros' distribution (Johnson and Kotz, 1969), and is analagous to predicting the zero reported accident count after fitting the truncated Poisson distribution to the

truncated reported accident distribution. This section describes the methods used in the truncation and estimation of the traditional accident distributions.

Truncating the Poisson distribution at zero, David and Johnson (1952) obtained the maximum likelihood estimate of the Poisson parameter by equating the mean value of y to its expectation, and so L is the solution of:

$$\bar{y} = L / [1 - \exp(-L)].$$

To find the parameters of the truncated Negative Binomial distribution we used the Fortran computer program written by Wyshak (1974). The program uses maximum likelihood estimators obtained as follows.

If we consider the Negative Binomial distribution in the form previously referred to in equation (2.1), we have;

$$P(k) = \frac{(N+k-1)! \times (P)^k \times (1+P)^{-(N+k)}}{(N-1)! \times k!}, \quad k = 0, 1, 2, 3, \dots,$$

where, N and P are the parameters to be estimated, and,
 $P = (M/N)$,

where M is the mean of the distribution. Then,

$$P(0) = (1 + M/N)^{-N}.$$

To obtain the corresponding probabilities for the truncated distribution, $P(k)$ must be divided by

$$[1 - P(0)].$$

By letting $w = 1/(1 + P)$ we get;

$$P(k) = \frac{w^N \times (1 - w)^{-k} \times (N + k - 1)!}{(N - 1)! \times k!}, \quad k = 0, 1, 2, 3, \dots$$

The log likelihood is then differentiated with respect to the parameters to obtain the parameter estimates for the truncated distribution.

This concluded the direct modelling of the reported accident distribution using accident probability distributions and we now discuss the regression analysis.

2.6 REGRESSION ANALYSIS

In the final stage of the analysis two different types of regression modelling were carried out, namely GLIM regression and logistic regression modelling. This section describes the two types of regression models which were fitted and defines the different dependent reported accident outcome variables used. The GLIM regression analysis was the preliminary regression analysis and thirty variables were included in a stepwise regression procedure. Consequently, many children were removed from this part of the analysis because of missing data.

Different reported accident outcome variables were used in both types of regression modelling. These were crudely defined as there was no medical judgement of seriousness available. The hierarchy of reported accident

seriousness is defined as follows.

- (i) Head fractures together with lacerations of the head requiring stitches.
- (ii) (i) + other bone fractures/dislocations.
- (iii) (ii) + other lacerations requiring stitches.
- (iv) (iii) + ingestions, burns, scalds and foreign body type of accidents.
- (v) (iv) + other lacerations requiring stitches.
- (vi) (v) + sprains, i.e. non-superficial accidents.
- (vii) Superficial and unidentified injuries.
- (viii) (vi) + (vii), i.e. all reported accidents.

The male and female populations were modelled separately because GLIM places restrictions on the size of the data that can be modelled. One of the facilities of GLIM is that the user is able to specify an error structure and so for ease of calculation it was assumed that the error structure was Poisson with a log link. Once the final models for 'all accidents' had been selected the error structure was redefined to be a Negative Binomial error and the parameter estimates were recalculated. There was very little difference in the model parameter estimates obtained using the different error structures as Maycock and Hall (1984) have already concluded in the analysis of their data.

When using the Negative Binomial error structure we assume that the population comprises of smaller sub-populations each of which has a separate Poisson accident rate λ_i . These parameters λ_i are from a Gamma distribution. Hence, we have a Negative Binomial

distribution to describe the accident distribution for the whole population. This is essentially the same technique that was used by Maycock and Hall and we make use of their GLIM macros to fit the Negative Binomial error. The distribution can be formally derived as follows.

If $P(Z)$ is the probability density function of Gamma form for the distribution of mean accident frequencies between sub-populations (Z being a continuous positive random variable), then,

$$P(Z) = \frac{\mu^S \times Z^{S-1} \times \exp[(-S/\mu) \times Z]}{\Gamma(S)},$$

where μ is the mean of the distribution and S the second parameter.

The variance of this distribution can be shown to be μ^2 / S .

The sampling distribution (i.e. the probability of y accidents in a given period) in any sub-population with a mean accident frequency z is Poisson since accidents are assumed to be independent and occur at random throughout the population. Hence,

$$P(y|z) = \frac{z^y \times e^{-z}}{y!}, \quad y = 0, 1, 2, 3, \dots$$

Hence, the sampling distribution over all sub-populations will be;

$$P(y) = \int_0^{\infty} P(z) \times P(y|z) dz,$$

(regarding y as a constant whilst integrating over z)
which gives;

$$P(y) = \frac{\frac{S^y}{\mu^y \times \Gamma(S) \times y!}}{\int_0^{\infty} z^{S+y-1} \times \exp\{-(1+S/\mu) \times z\} dz} dz.$$

Substituting $w = z \times (1 + S/\mu)$ and integrating with respect to w gives;

$$P(y) = \frac{\Gamma(S+y)}{\Gamma(S) \times y!} \times \left(\frac{S}{\mu + S} \right)^S \times \left(\frac{\mu}{\mu + S} \right)^y,$$

the Negative Binomial distribution with mean value μ , and variance $\mu \times (\mu + S)/S$.

The scaled deviance used in GLIM is a measure of goodness-of-fit of a particular model and is defined as a likelihood ratio statistic equal to $-2 \times \log(L)$, where L is equal to:

$$\frac{(\text{maximized log-likelihood for the current model})}{(\text{maximized log-likelihood for the full model})}$$

In terms of the Negative Binomial error the model deviance becomes,

$$2 \sum \left(y \times \log \left(\frac{y}{\mu} \right) - (y + S) \times \log \frac{y + S}{\mu + S} \right)$$

As S approaches infinity the distribution becomes the Poisson distribution and the between sub-population variation disappears. In addition, the Poisson deviance

will equal the Negative Binomial deviance.

The value of S can be estimated as that value for which the sample deviance equals its expected value. In order to do this calculation it is necessary to estimate the sample expectation value of deviance as a function of the mean value and the parameter S . This was done iteratively within GLIM using macros written by Maycock and Hall (1984).

The Negative Binomial error structure was fitted using the 'own' fit facility in GLIM and resulted in parameter estimates which were very similar to those obtained using a Poisson error but with larger standardised errors. As one would expect, the use of the Negative Binomial error in the model also resulted in a reduced model deviance.

The GLIM regression models fitted are of the following form.

$$M = a + X_i \times B_i,$$

where M is the linear predictor and a , B_i ($i=0, 1, \dots$), are parameters to be estimated.

The log link was used and we also used the GLIM offset facility to offset the constant five. The number of accidents are assumed to be Poisson distributed, but by offsetting the constant five we are able to model the mean number of accidents without affecting the error distribution.

The modelling procedure used in the model selection process was the forward stepwise method:

- STEP 1. The general mean or intercept was first fitted to the male or female population.
- STEP 2. A single variable was temporarily added to the model and the improvement in fit of the model was recorded.
- STEP 3. Step 2 was repeated for each variable and the variable which gave the most significant improvement in fit was identified.
- STEP 4. The variable which gave the most significant improvement was permanently added to the model and STEP 2 is repeated with the new model.

These steps were repeated with the addition of further variables into the model until no further significant improvement in fit at the 0.1% significance level was possible.

The test of significance used in assessing the effect of the addition of a variable into the model was determined by calculating the difference in the model deviance caused by the introduction of the new variable. Theoretically this difference follows a Chi-squared distribution with the degrees of freedom being equal to the difference in the degrees of freedom from the two model fits.

Once the final models had been selected the mean deviance reductions for each variable were calculated as follows.

$$\text{M.D.R.} = \frac{(\text{deviance difference}) / (\text{df1} - \text{df2})}{(\text{final model scaled deviance}) / (\text{df1})},$$

where the deviance difference is the difference between the scaled deviance of the final model (degrees of freedom = df1) and the scaled deviance of the model with a particular variable omitted (degrees of freedom = df2). This ratio was then compared to the values from the F-tables.

Only fifteen variables were included in the stepwise logistic regression modelling and these were selected from the results of the GLIM regression modelling analysis and the earlier analyses.

We obtain the standard logistic model by assuming that there are just two accident response outcomes, i.e. no accidents or some accident(s). Hence,

$$\text{Prob}(Y=0) = 1-p = \text{probability of no-accident.}$$

$$\text{Prob}(Y=1) = p = \text{probability of accident(s).}$$

The logistic regression models fitted are of the following form.

$$M = a + X_i \times B_i,$$

where M is the linear predictor and a, B_i ($i=0, 1, \dots$), are parameters to be estimated.

Choosing the logit response transformation we obtain

$$\text{logit}(p) = \log[p/(1-p)].$$

Hence, $p/(1-p) = \exp(a + X_i \times B_i)$, and therefore,

$$\text{Prob}(Y=0) = \frac{1}{(1 + \exp(-a - (X_i \times B_i)))}.$$

The best fitting logistic model was selected using the same stepwise selection process described and used in the stepwise GLIM regression analysis. Variables were added into the logistic regression model based on the most significant improvement in fit at the 0.1% significance level. Once the final models were obtained, the mean deviance reduction was calculated for each variable in the

model.

In our analysis the logistic regression modelling requires a binary response variable. When modelling the full reported accident distribution four different binary outcome response groupings were used and these are defined as follows.

- 1) [Zero reported accidents] or [one or more reported accidents].
- 2) [Zero or one reported accident] or [two or more reported accidents].
- 3) [Zero, one or two reported accidents] or [three or more reported accidents].
- 4) [Zero reported accidents] or [three or more reported accidents].

When modelling the other outcome variables such as reported bone fractures, the binary response variable was defined as 1) above. The final logistic models were compared across gender in Chapter 5, Section 2.

Only main affects were examined in each type of regression modelling and the final models from using the two types of regression modelling are compared in Chapter 5, Section 3.

2.7 DISCUSSION

In this Chapter we have laid out the analysis in progressive stages. The first step was the simple exploratory analysis which resulted in strong evidence to support our hypothesis of accident misreporting and also

indicated the presence of accident proneness in children who displayed particular personality characteristics.

This simple analysis was followed by the traditional probability distribution modelling of the full reported accident distribution of sub-populations of the data. Again we found evidence of accident misreporting and saw that the Negative Binomial model consistently overestimated the reported two accident count and underestimated the one reported accident count. By introducing a parameter of misreporting into the Negative Binomial distribution model we discovered a good model for the data.

The use of alternative modelling distributions was also investigated. The generalised models which allowed for accident misreporting provided a good fit to the data. Truncating the data at zero was also investigated which is analogous to the situation where the zero accident count cannot be recorded, for example, data collected at hospital admissions.

A problem arises when fitting a large number of probability distributions to data since a good fit may be obtained from more than one distribution. In addition, the assumptions and conclusions inherent in fitting any particular distribution may not be correct. For example, suppose a good fit is obtained using a traditional Negative Binomial model. This distribution can be derived by modelling either proneness or contagion in the population. Hence we can form no conclusion on whether or not proneness is present. Great care must therefore be

taken when selecting the final model and thought must be given to a sensible interpretation.

In an attempt to find evidence of proneness we calculated the coefficients of correlation between the number of reported accidents that occurred in consecutive time periods. Three different sets of time periods were used and in addition to using the full reported accident distribution, the data was truncated at the zero accident count and the correlation analysis was repeated using those children with one or more reported accidents. The problems of accident recall were very obvious in this part of the analysis since the reported accident times were heaped in six monthly intervals.

By breaking down the reported accidents into smaller sub-groups according to the seriousness of each accident and repeating the earlier distribution modelling of sub-populations we produced further evidence of accident misreporting. From this we must conclude that particular types of accidents are reported more accurately than other types of accidents.

The purpose of the regression analysis was to support the earlier stages of the analysis. The regression results confirm the earlier findings and give support to the hypothesis that differential accident reporting is present.

CHAPTER 3

ANALYSIS OF THE BRITISH BIRTH DATA SET, PART I; VARIANTS ON THE POISSON AND NEGATIVE BINOMIAL PROBABILITY DISTRIBUTIONS

3.1 INTRODUCTION

In this Chapter we present the results of fitting the models which were developed in Chapter 2, to sub-populations of the full reported accident data set from the British Birth Study. By detailed examination of the data it is shown that fitting the Poisson and Negative Binomial distributions is unsatisfactory. However, the new Model VI, developed by including a misreporting parameter that takes into account misreporting two accidents as one accident, is shown to provide a much better fit to the data than the other generalisations of the traditional accident models. Further, Model VI leads to results with sensible interpretation.

We begin by identifying and discussing key variables which had a large effect on the mean reported accident rate and through which accident proneness could be defined. The variables were used to define 88 male sub-populations and 84 female sub-populations, described in Section 3.2. The distribution of the mean accident rates across the sub-populations was examined in order to

identify trends and anomalies in the data.

We start the modelling of the full reported accident distribution using the simple Poisson distribution and the Negative Binomial distribution. Both these distributions have traditionally been used in the past to model accident data.

The goodness-of-fit of each distribution to each sub-population was measured using the Chi-squared goodness-of-fit test. A problem concerning large sub-population sizes was encountered, namely that large sample sizes usually produced poor goodness-of-fit results, yet smaller sample sizes generally produced satisfactory results. This is discussed in Section 3.4..

It is suspected that different types of accidents are recalled with different degrees of accuracy. Hence, in this Chapter the reported accident outcome variable is redefined to include only the most serious of the reported accident injuries and the earlier traditional accident modelling is repeated using this outcome variable. The results from this modelling indicate the need for further investigation and this is pursued in Chapter 5 where we repeat the traditional distribution modelling using several different accident outcome variables. In addition, in Chapter 6 we also model different reported accident outcome variables using different regression modelling techniques.

From our analysis we have evidence of accident misreporting in the full reported accident distribution. Hence, the traditional modelling distributions were

generalised to take into account accident misreporting between particular accident counts. For this part of the analysis we returned to using the full reported accident distribution.

To investigate possible types of accident misreporting we also systematically developed three separate Generalised Poisson distributions (see Section 2.3.1) and two Generalised Negative Binomial models (see Section 2.3.2).

We now discuss the results of the reported accident distribution modelling.

3.2 FITTING THE TRADITIONAL DISTRIBUTIONS TO THE FULL REPORTED ACCIDENT DISTRIBUTION

In this section we describe the variables which were used to split the data into mutually exclusive sub-populations. The male and female populations were split into non-overlapping sub-populations with respect to a single variable at a time, such as neighbourhood say, and a separate Poisson distribution and a separate Negative Binomial distribution were fitted to the reported accident distribution of each sub-population in an attempt to identify the underlying accident distribution. The variables were selected on the basis of previous studies and sub-population sizes.

The variables selected to split the data into smaller non-overlapping sub-populations, together with their

values, were as follows.

- 1) The child's mother's age in April 1970, (less than 20 years old, 20-30, over 30 years old).
- 2) The child's father's age in April 1970, (less than 20 years old, 20-30, 30-40, 40-65 years old).
- 3) The child's mother's highest educational qualification, (none, typing, 'O' levels, 'A' levels).
- 4) The child's father's highest educational qualification, (no father, none, trade, 'O' levels, 'A' levels, degree, other).
- 5) The family's social class based on the occupation of the child's father, (manual, non-manual).
- 6) Whether the child's mother smoked, (non-smoker, smoker).
- 7) Whether the child's father smoked, (non-smoker, smoker).
- 8) The region of the U.K. in which the child lived, (North, York/Humberside, North West, East Midlands, West Midlands, East Anglia, South East, South West, Wales, Scotland).
- 9) The number of subsequent births in the family (0, 1 or 2).
- 10) The number of household moves during 1970-1975 (0, 1, 2, 3, more than 3).
- 11) The number of older siblings (0, 1, 2, 3, more than 3).
- 12) The number of younger siblings (0, 1, more than 1).

In addition, we included three variables which are assessments made by the Health Visitor of the household, namely:

- 13) The standard of the household equipment in the child's home, (luxury/high standard, adequate, low/very low standard).
- 14) The level of tidiness of the child's home, (very/over tidy, average, untidy/chaotic).
- 15) The neighbourhood of the child's home, (poor urban area, average, well-to-do, rural).

The following four variables are assessments made by the child's mother of characteristics of her child's nature and personality. The responses are: not applicable, applied somewhat, certainly applied.

- 16) Whether the child squirmed and fidgetted.
- 17) Whether the child destroyed belongings.
- 18) Whether the child was solitary.
- 19) Whether the child bullied other children.

The child's mother's has been indicated as an important factor in their child's accident rate in the past, (Langley and Silva, 1982). Therefore seven variables which are assessments made by the child's mother of her own health were included. The responses are: applied, did not apply.

- 20) Whether she often felt miserable or depressed.
- 21) Whether she often got into a violent rage.
- 22) Whether she was easily upset or irritated.
- 23) Whether she was often keyed-up and jittery.
- 24) Whether every little thing got on her nerves and wore her out.
- 25) Whether she had ever suffered a nervous breakdown.
- 26) Whether she had other health problems not specifically mentioned in the questionnaire.

The preliminary exploration into the data set revealed differences in the nature of the male and female reported accident distributions. These findings were confirmed in Section 1.3.7 by making a comparison of the reported accident means between gender for sub-population categories within each variable where it was seen that in every category the male reported accident mean was greater than the female reported accident mean.

For both males and females the earlier preliminary analysis also revealed a reporting trend with respect to social index grouping and head injuries. The association we observed was that for both males and females, head injuries were proportionally more frequent in the most advantaged social index grouping than in the other

groupings. For this reason we will look at the results from the social index split first and then examine the results for the sub-populations obtained from splitting by the other variables.

After splitting by social index the lowest reported male accident mean was seen in the most advantaged social index sub-population and the highest reported male accident mean was seen in the most disadvantaged social index sub-population. For females we found that the lowest reported female accident mean was seen in the most advantaged social index sub-population and the highest reported female accident mean was seen in the disadvantaged social index sub-population. This result was to be expected since children from the less advantaged households would certainly have less facilities and more opportunities to have accidents themselves than children from other social index groups.

A summary of the reported accident means for the sub-population splits across gender is given in Table 8.

From Table 8 we observe that the majority of the reported accident means for males were between 0.65 and 0.75. The lowest reported accident mean of 0.5625 was for males whose father's were between 40 and 65 years old in April 1970, and the highest of 0.8880 was for males from a household with a low/very low standard of household equipment.

Fifty of the male sub-populations had reported accident means that were above 0.6957 (the reported accident mean for the full male population), and 15 of the

groupings. For this reason we will look at the results from the social index split first and then examine the results for the sub-populations obtained from splitting by the other variables.

After splitting by social index the lowest reported male accident mean was seen in the most advantaged social index sub-population and the highest reported male accident mean was seen in the most disadvantaged social index sub-population. For females we found that the lowest reported female accident mean was seen in the most advantaged social index sub-population and the highest reported female accident mean was seen in the disadvantaged social index sub-population. This result was to be expected since children from the less advantaged households would certainly have less facilities and more opportunities to have accidents themselves than children from other social index groups.

A summary of the reported accident means for the sub-population splits across gender is given in Table 8.

From Table 8 we observe that the majority of the reported accident means for males were between 0.65 and 0.75. The lowest reported accident mean of 0.5625 was for males whose father's were between 40 and 65 years old in April 1970, and the highest of 0.8880 was for males from a household with a low/very low standard of household equipment.

Fifty of the male sub-populations had reported accident means that were above 0.6957 (the reported accident mean for the full male population), and 15 of the

male sub-populations had reported accident means above 0.8.

TABLE 8. Reported accident means for the sub-populations across gender.

SAMPLE MEAN RANGE	Frequency	
	MALE	FEMALE
0.35 to < 0.40	0	2
0.40 to < 0.45	0	1
0.45 to < 0.50	0	18
0.50 to < 0.55	0	39
0.55 to < 0.60	4	11
0.60 to < 0.65	8	6
0.65 to < 0.70	28	3
0.70 to < 0.75	25	3
0.75 to < 0.80	8	0
0.80 to < 0.85	9	1
0.85 to < 0.90	6	0
TOTAL	88	84

The males with the five highest reported accident means are as follows with the sub-population reported accident mean given in brackets.

- i) Males from a household with a low/very low standard of household equipment (0.8880).
- ii) Males who certainly destroyed belongings (0.8779).
- iii) Males whose mothers often had violent rages(0.8675).
- iv) Males whose fathers were under 20 years old in April 1970 (0.8663).
- v) Males who certainly squirmed and fidgeted (0.8648).

It is not suprising that these five male sub-populations include males with undesirable personality characteristics since these males would naturally get into more situations where accidents could occur. It is also not suprising that these five sub-populations also include males whose mothers often had violent rages since these

males would live in constant fear of their mothers erupting into violence. Males with young fathers also have a high reported accident mean and this is a reasonable finding since father's age would be associated with mother's age, neighbourhood, standard of household equipment, etc., all of which seem to have an effect on the nature of the reported accident distribution.

For females we observe from Table 8 that the majority of the reported accident means were between 0.45 and 0.60. The lowest reported accident mean of 0.3506 was for females whose home was in East Anglia, and the highest of 0.8095 was for females who bullied others. Forty three of the female sub-populations had reported accident means that were above 0.5208 (the reported accident mean for the full female population), and four of the female sub-populations had reported accident means above 0.7. The females with the five highest reported accident means are as follows with the sub-population reported accident mean in brackets.

- i) Females who certainly bullied others (0.8095).
- ii) Females whose fathers were under 20 years old in April 1970 (0.7320).
- iii) Females whose mothers often had violent rages (0.7200).
- iv) Females who certainly destroyed belongings (0.7154).
- v) Females with more than three household moves (0.6680).

As in the male population it is not suprising that these five sub-populations also include females with personality problems or females whose mothers often had violent rages. We also see that females with more than

three household moves are included and it is thought that frequent changes of home have a disorienting effect on the child, although the effects of this variable would also be confounded with other variables such as the child's age.

Overall it was also observed that only four of the female sub-populations had reported accident means that were higher than the reported accident mean for the full male population, 0.6957. In addition, the reported accident mean for each male sub-population was always larger than the reported accident mean of the full female population, 0.5208.

There are some similarities across gender when comparing the sub-populations with the five highest reported accident means and for both males and females these include children whose fathers were under 20 years old in April 1970, children whose mothers often had violent rages and children who destroyed belongings.

When splitting the separate male and female populations into nonoverlapping sub-populations we identified several reported accident frequency trends. It was observed for both males and females that as grouped mother's age increased so the reported accident mean decreased. The same trend was also observed when looking at grouped father's age for each sex.

This parental age-related result is not unexpected since we would expect that in comparison with older parents, younger parents would be less aware and less responsive to potential accident situations. Older parents would be more mature and responsible in their

outlook than the younger parents. In addition mother's age and father's age are closely associated and the overall affects could be confounded with other variables such as parents' income, neighbourhood, the standard of household equipment, etc..

Another reported accident frequency trend which was observed for both parents and for both sexes was with respect to parent's smoking habits. It was observed that those children whose mother's smoked had a higher reported accident mean than those children whose mothers did not smoke. This same trend was also found when looking at children whose fathers smoked/did not smoke.

One would have thought that children whose parents smoked would be at higher risk to burns in particular, however in a further breakdown of type of injury with these 'smoking' variables no such trend was identified. It is also possible that burns from cigarettes/cigars, etc., might be considered trivial by parents who are smokers yet serious by parents who do not smoke, hence possible accident misreporting.

When examining specific characteristics of the child's personality as assessed by the child's mother we found that higher reported accident means were present in children who displayed certain undesirable personality characteristics. In particular, looking at children who bullied others we observed that for both males and females, those who certainly bullied others had a higher reported accident mean than those who sometimes bullied others, who in turn had a higher reported accident mean

than those who never bullied others. This pattern was also seen for each sex when looking at children who certainly squirmed, sometimes squirmed and never squirmed, also for children who certainly destroyed belongings, sometimes destroyed belongings, and never destroyed belongings.

These results are not unexpected since children who bullied others, squirmed, or destroyed belongings placed themselves at higher risk simply by exposing themselves more frequently to the types of situations where accidents are likely to occur.

If we recall that Haddon, Suchman and Klein (1964) attributed proneness to 'an enduring or stable personality characteristic that predisposes an individual towards having accidents', then using this definition we see possible evidence of accident proneness. For example, by this definition we see that children who display certain personality characteristics have a higher reported accident mean than children without those characteristics. Hence, they could be considered as an accident prone group.

One result seen only in the male population was that males who were certainly solitary had a higher reported accident mean than males who were sometimes solitary, who in turn had a higher reported accident mean than males who were never solitary. We have assumed that a child who is not solitary is in the company of adults or other children or a mixture of both, the questionnaire is not specific on

this. Hence, the child may be supervised or influenced by the other children, or the adults, or both. We hypothesise that a child who prefers his own company would not have the same supervision or attention that might otherwise have been available and this would put the child in a higher risk group.

Mother's health was also found to be an important variable when examining reported accident means for both males and females. Malaise conditions of the child's mother which when present were associated with a high reported accident mean for both males and females were as follows; miserable or depressed moods, often had violent rages, generally being easily upset, being keyed-up and jittery, nervous, having suffered at one time from a nervous breakdown, and other problems not mentioned specifically in the Questionnaire.

These results are not suprising since one would expect a child to be greatly influenced by the health of its parents and in particular the health of the mother. For example, if the mother is feeling unwell or suffering say from a migraine then the child would be given less attention and less supervision than if the mother was fit and healthy. In addition, if the mother was unable to give her full attention or supervision to her child then the child's activities would have to be restricted.

It is also possible that the presence of particular mother's malaise conditions may directly affect the child's personality. For example, a child whose mother often suffers from migraines may develop a solitary

personality, or alternatively the child may develop a bullying nature as a way of demanding more attention.

Taking into account the size of the child's family we continue to find reported accident frequency trends. When looking at the number of younger siblings in the female's family we found that the reported accident mean increased as the number of younger siblings increased. For males we found that those who were born last in the family had a lower reported accident mean than those with subsequent births in the family. The highest reported accident mean was for males who had just one younger sibling.

Looking at the number of older siblings in the female's family we see that those females with two older siblings had the lowest reported accident mean and those with more than three older siblings had the highest reported accident mean.

It is suspected that the number of older and younger siblings does have an effect on the reported accident means for both males and females but it is unclear as to the exact affects since a child can have both older and younger siblings. One would have thought that the larger a family the less attention or supervision a particular child would get from the parents. However, as a type of compensation a child with older siblings may get attention or supervision from the older children in the family. In terms of accidents one would have thought that the presence of older siblings would decrease the reported accident mean because of the attention or supervision that the older siblings would give to the younger child.

In contrast, the presence of younger siblings might have the reverse effect since less attention or supervision might be given to the older child. It is also possible that children with siblings get less individual attention or supervision than single children so accident misreporting might occur. Larger families might also have more accidents occurring in the household which could result in poor accident recall with respect to individual children. In addition, it is thought that the number of siblings could be closely associated with parents age because of the time involved in building up a larger family.

The effect of a combination of older and younger siblings is therefore unclear from these results and is also likely to be confounded with other variables such as parents' age and parents' income level and the gender of the siblings.

We have already seen in the data that most childhood reported accidents occurred in the child's own home so it is no surprise that we also observed reported accident trends with respect to variables associated with the child's home. For both males and females we saw that the reported accident mean increased as the number of household moves increased. This is understandable since we would expect a child to become increasingly disoriented in new surroundings as the number of household moves increased, a result which was observed. This effect may also be confounded with the child's age since a child becomes more mobile and more aware of its surroundings as

it gets older. The older a child gets the wider the range of activities that he gets involved in.

It was also found that the reported accident mean for both males and females decreased as the standard of household equipment in the household increased. This is not suprising since we would expect that children from households with a high standard of household equipment, including such things as good quality floor coverings, good quality furniture, central heating, better quality toys, etc., would be less likely to injure themselves say from falling onto thick carpet than a child from a household with a poor standard of household equipment falling onto a harder floor covering. Clearly this variable is closely associated with other variables such as parents' income, parents' age and neighbourhood.

For both sexes we also observed that the reported accident mean increased as the level of tidiness of the child's home decreased. This result is not suprising since a child in an untidy/chaotic home would have more things to fall over, or to bump into, than a child whose home is very tidy and where everything is organised and in its own specific place, i.e. there are fewer supprises for the child. This result may also be linked with the previous result since a child falling over in an untidy/chaotic household may injure himself more often if falling onto a hard flooring than a child falling over onto a heavy carpet floor covering.

On a wider dimension we considered the child's neighbourhood and found that males from a poor urban area

had the highest reported accident mean, followed by males from an average neighbourhood, then well-to-do and rural neighbourhoods. In the female population we observed a slightly different trend in that females from a poor urban area had the highest reported accident mean, followed by females from an average neighbourhood, then rural and well-to-do neighbourhoods.

These results are quite plausible since the child's neighbourhood would be closely associated with variables such as social class, the standard of household equipment, the age of the parents, the mother's malaise condition, etc., and from the results already described, children from poor urban areas would most certainly be in a higher accident risk exposure group than children from other types of neighbourhoods. It was also found that for both males and females, those children whose father's were in the manual social class had a higher reported accident mean than those children whose father's were in the non-manual class. This variable would also be closely associated with neighbourhood, parent's age and other factors.

On a regional basis we observed that for both males and females the region with the lowest reported accident mean was the East Anglian region. The region with the highest reported accident mean for both sexes was the Northern England region.

In summary we note that some of the results from this section are not suprising, for example, children from low social index groups have a higher reported accident mean

than children from higher reported social index groups. We have also uncovered evidence of reported accident proneness in that children with undesirable personality characteristics have higher reported accident rates than those without those characteristics. This evidence of reported accident proneness supports the use of the Negative Binomial modelling in Section 3.4 since the Negative Binomial model can be interpreted as a model of accident proneness where the population comprises of sub-populations with different accident rates (see Section 2.2).

3.3 FITTING THE POISSON ACCIDENT DISTRIBUTION

In this section and the following section we present the results obtained from fitting the Poisson distribution and Negative Binomial distribution to sub-populations of the data set. The first step in this part of the analysis was to split the male and female populations into non-overlapping sub-populations with respect to one variable at a time. A separate Poisson distribution and a separate Negative Binomial distribution were fitted to the reported accident distribution from each sub-population.

Several statistics were calculated for each reported accident distribution. For instance, the reported accident mean and variance, the sample moments, the moments obtained from using the Poisson parameter and the Negative Binomial parameters and the Chi-squared goodness-

of-fit test statistics obtained from fitting the Poisson and Negative Binomial distributions.

We now examine the results of modelling the reported accident distribution directly using the simple one parameter Poisson probability distributions. The goodness-of-fit results from fitting the Poisson and Negative Binomial distributions to the male and female sub-populations are given in Table 9. Fitting a separate Poisson model to each of the male sub-populations resulted in an acceptable fit at the one percent significance level for 36 (42%) of 88 male sub-populations. For the female sub-populations we obtained acceptable fits at the one percent level for 35 (42%) of the 84 female models. By examining the observed and predicted counts from the poor fitting Poisson models we see where the model performed badly, however no pattern was identified when examining the poor fitting Poisson models.

TABLE 9. Goodness-of-fit results from fitting the Poisson and Negative Binomial distributions to the full reported accident distributions of the male and female sub-populations.

		Level of significance for the Chi-squared values obtained from the fitting the Poisson and Negative Binomial distributions				
Sex	Distrib.	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
MALE	POISSON	52	10	7	19	88
	N.B.	14	26	11	34	85
FEMALE	POISSON	48	12	2	22	84
	N.B.	22	18	5	34	79

When answering the question 'How well does the Poisson model fit the sub-populations?' we used the Chi-squared goodness-of-fit test statistic calculated from the predicted and reported accidents and compared this with the standard Chi-squared tables with the appropriate degrees of freedom. The Poisson distribution was acceptable at the ten percent significance level for some of the male sub-populations. For example;

- i) The 172 males who had no father (p-value of 0.383).
- ii) The 384 males whose fathers were over 40 years old in April 1970 (p-value of 0.409).
- iii) The 874 males whose mother's highest educational qualification was a typing qualification (p-value of 0.341).
- iv) The 259 males whose home had a low or a very low standard of household equipment (p-value of 0.445).
- v) The 1,452 males whose home was in a well-to-do neighbourhood (p-value of 0.158).

In the females the Poisson model was also acceptable at the ten percent level for a number of sub-populations. For example;

- i) The 1,365 females whose mothers were over 30 years old in April 1970 (p-value of 0.242).
- ii) The 153 females whose fathers less than 20 years old in April 1970 (p-value of 0.917).
- iii) The 1,008 females whose mother's highest educational qualification was 'O' level standard (p-value of 0.107).

We would expect the sample (mean/variance) ratio for the male and female sub-populations mentioned above to be very close to one since in any good-fitting Poisson model the mean equals the variance. As expected, the sample (mean/variance) ratios for these sub-populations were high ranging from 0.7947 to 0.9182. However, it should also be

noted that there were sub-populations with higher sample (mean/variance) ratios but for which the p-value from the Poisson model indicated a poor fit. From the male population for example;

- i) The 4,230 males whose mothers were 20-30 years old in April 1970. We observed a sample (mean/variance) ratio of 0.8784 and a p-value of 0.000.
- ii) The 1,952 males whose fathers were 30-40 years old in April 1970. We observed a sample (mean/variance) ratio of 0.9010 and a p-value of 0.000.

From the female sub-population;

- i) The 900 females whose mother's highest educational standard was 'O' levels. This group had a sample (mean/variance) ratio of 0.8636 and a p-value of 0.002.
- ii) The 2,447 females whose mother's smoked had a sample (mean/variance) ratio of 0.9057 and a p-value of 0.001.

As we would expect the sample (mean/variance) ratio to be very close to one it is suprising to find sub-populations with approximately the same (mean/variance) ratio and yet very different p-values. This could be a reflection on the size of the sub-populations since it was generally seen that smaller sub-populations produced better Poisson fits according to the Chi-squared goodness-of-fit test than larger sub-populations.

It should also be noted that we are 'forcing' the Poisson parameter on the assumption that the sub-populations do actually follow separate Poisson distributions. An alternative method of obtaining a Poisson fit would be to use a range of values for the Poisson parameter and fit a series of models to each sub-

population using those parameters. The final selection of the parameter would then be dependent on the best Chi-squared goodness-of-fit test.

From the results of the Poisson model fitting completed in this section it seems unlikely that the Poisson accident model does accurately describe the data. This is not suprising since we have already uncovered evidence of accident misreporting and accident proneness in the earlier analysis (see Section 1.3.4 and Section 3.2.)

3.4 FITTING THE NEGATIVE BINOMIAL ACCIDENT DISTRIBUTION

We now examine the results of fitting the Negative Binomial distribution to the male and female sub-populations (described in Section 3.2).

When fitting a separate Negative Binomial model to each of the male sub-populations we obtained a satisfactory fit at the one percent level for 71 (84%) of the 85 male sub-populations and 57 (72%) of the 79 female sub-populations. These results can be seen in Table 9. At the five percent level we accept 45 (54%) of the 85 male sub-populations and 39 (49%) of the 79 female sub-populations.

The poor fitting Negative Binomial models were examined to see where the model failed and in almost every case the poor fit was attributable to the one accident and two accident counts. It was observed that the Negative

Binomial model consistently underestimated the one accident count and overestimated the two accident count.

In comparison to the simple, single parameter Poisson distribution we estimate two parameters for the Negative Binomial distribution and by estimating the extra parameter we reduce the degrees of freedom available for the Chi-squared goodness-of-fit test. For example, if we consider the male sub-population from the York/Humberside region we find that the Poisson model produced a Chi-squared value of 4.7767 (2 d.f.) and a p-value of 0.092, whereas the Negative Binomial model produced a Chi-square value of 4.1946 (1 d.f.) and a p-value of 0.041. The Chi-squared test statistic from each fit are very similar but the degrees of freedom are fewer in the Negative Binomial model which resulted in a more significant p-value.

As a further example, consider the sub-population of females whose fathers were less than 20 years old in April 1970. The Poisson model produced a Chi-squared value of 0.1737 (2 d.f.) with a p-value of 0.917 whereas the Negative Binomial model for the same sub-population produced a Chi-squared value of 0.5251 (1 d.f.) with a p-value of 0.469.

We also observed that the Negative Binomial model generally produced an acceptable fit to sample sizes of under 1,000 and a poor fit to larger sample sizes. It was thought that sample size had a great influence on the Chi-squared test statistic, hence, Table 10 was constructed to investigate the possible effects.

In Table 10 the male and female populations were

successively reduced by multiplying the reported frequencies for each count by a constant factor. Hence, we can see what happens to the significance of the Chi-squared statistic when separate Negative Binomial distributions are fitted to scaled down versions of the reported data.

TABLE 10. A listing of Chi-squared values for the Negative Binomial models for each sex showing what happens to the Chi-squared test statistic when the male and female populations are successively reduced by a constant factor.

SEX	Sample size	N.B. parameters		Chi-squared value(d.f)	p-value
		P	N		
M A L E	6310	0.1101	6.3214	15.718(3)	0.0013
	5509	0.1106	6.2896	13.584(3)	0.0035
	4812	0.1130	6.1677	12.532(3)	0.0058
	4201	0.1127	6.1835	10.851(3)	0.0126
	3671	0.1165	5.9972	9.802(3)	0.0203
	3205	0.1186	5.8941	8.951(3)	0.0230
	2786	0.1004	6.8860	6.190(3)	0.1027
	2443	0.1004	5.2080	5.841(2)	0.0539
	2132	0.1036	6.6917	5.126(2)	0.0771
	1862	0.1072	6.4777	4.607(2)	0.0999
	1626	0.1059	6.5545	3.787(2)	0.1506
	1419	0.1001	6.9054	3.430(2)	0.1800
F E M A L E	1238	0.0931	7.3906	2.806(2)	0.2459
	1081	0.0958	7.1947	3.397(2)	0.1829
	5830	0.0936	5.5655	15.520(2)	0.0004
	5091	0.0942	5.5282	13.600(2)	0.0011
	4445	0.0950	5.4864	12.032(2)	0.0024
	3883	0.0954	5.4637	10.406(2)	0.0055
	3390	0.0988	5.2858	9.668(2)	0.0080
	2961	0.1003	5.2149	8.639(2)	0.0133
	2584	0.0875	5.9296	6.690(2)	0.0353
	2257	0.0864	6.0021	5.709(2)	0.0576
	1970	0.0879	5.9011	5.111(2)	0.0777
	1720	0.0881	5.8932	4.827(2)	0.0895
	1501	0.0834	6.2013	3.589(2)	0.1662
	1312	0.0894	5.8029	3.026(1)	0.0820
	1146	0.0885	5.8673	2.802(1)	0.0942
	1000	0.0876	5.9137	2.232(1)	0.1352

From Table 10 we see that as the population size decreases by a constant factor the Chi-Squared test statistic decreases and the degrees of freedom decrease. We also note that the parameter estimates for P and N do not follow any particular trend and actually fluctuate in magnitude as the population decreases. It seems therefore, that when using the Chi-squared test as a goodness-of-fit test for large sample sizes we see a p-value that is larger than we would get if we took a smaller sample.

To further illustrate this, random samples were taken from the male and the female populations and a separate Negative Binomial distribution was fitted to each random sample. The Chi-squared test statistic was then calculated to measure the goodness-of-fit of the Negative Binomial model for each random sample. The size of the samples ranged from approximately 500 to 4,000 children.

The results confirm that generally, a good fit occurred in samples of 1,000 or less, and that for larger samples the fit was usually poor. The poor fits that occurred in samples of around 500 are usually due to the loss of a degree of freedom when fitting the extra parameter for the Negative Binomial distribution.

The phenomenon that we are observing is a typical consequence of using large sample sizes. It is suspected that if we were to reduce the sample sizes by splitting the data with respect to more than one variable at a time, then the Negative Binomial distribution would provide a good fit to the smaller sub-populations.

This was investigated and we found that as the sample sizes were reduced, so the fit of the Negative Binomial model improved. For example, if we split the male population with respect to two variables such as neighbourhood and grouped mother's age we obtain 12 male sub-populations ranging in size from 49 to 2,020 males. The Poisson distribution is fitted to the reported accident distribution of each of these 12 male sub-populations and gives an adequate fit at the one percent level to 11 of them. The Negative Binomial distribution is fitted to the reported accident distribution of eight of the sub-populations and gives an adequate fit at the one percent level to all of them.

From comparing the sample kurtosis and skewness values calculated from the sub-populations and the kurtosis and skewness values calculated using the parameters obtained from the Poisson and Negative Binomial fitted distributions, we made the following observation.

$$A < B < C$$

where A = the kurtosis calculated using the Poisson parameter,
 B = the kurtosis calculated using the Negative Binomial parameters,
 C = the sample kurtosis.

The same relationship was seen for skewness.

We now examine the parameter estimates for P and N obtained from fitting the Negative Binomial model. Figures 12 and 13 show the relationship between the parameters P and N obtained from fitting the Negative Binomial distribution to the male and female sub-

populations respectively.

FIGURE 12. The parameters P and N obtained from fitting the Negative Binomial distribution to the male sub-populations.

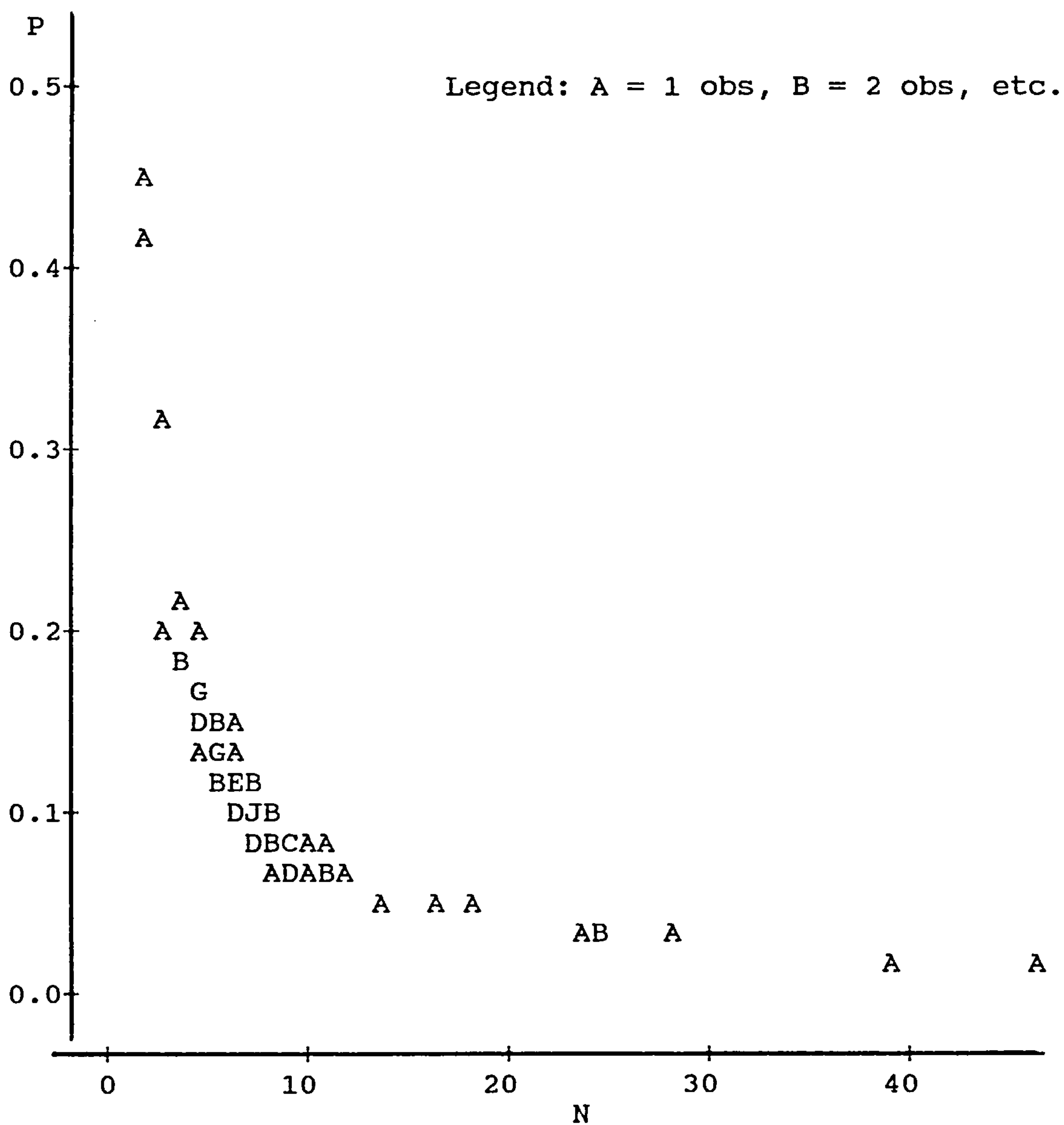
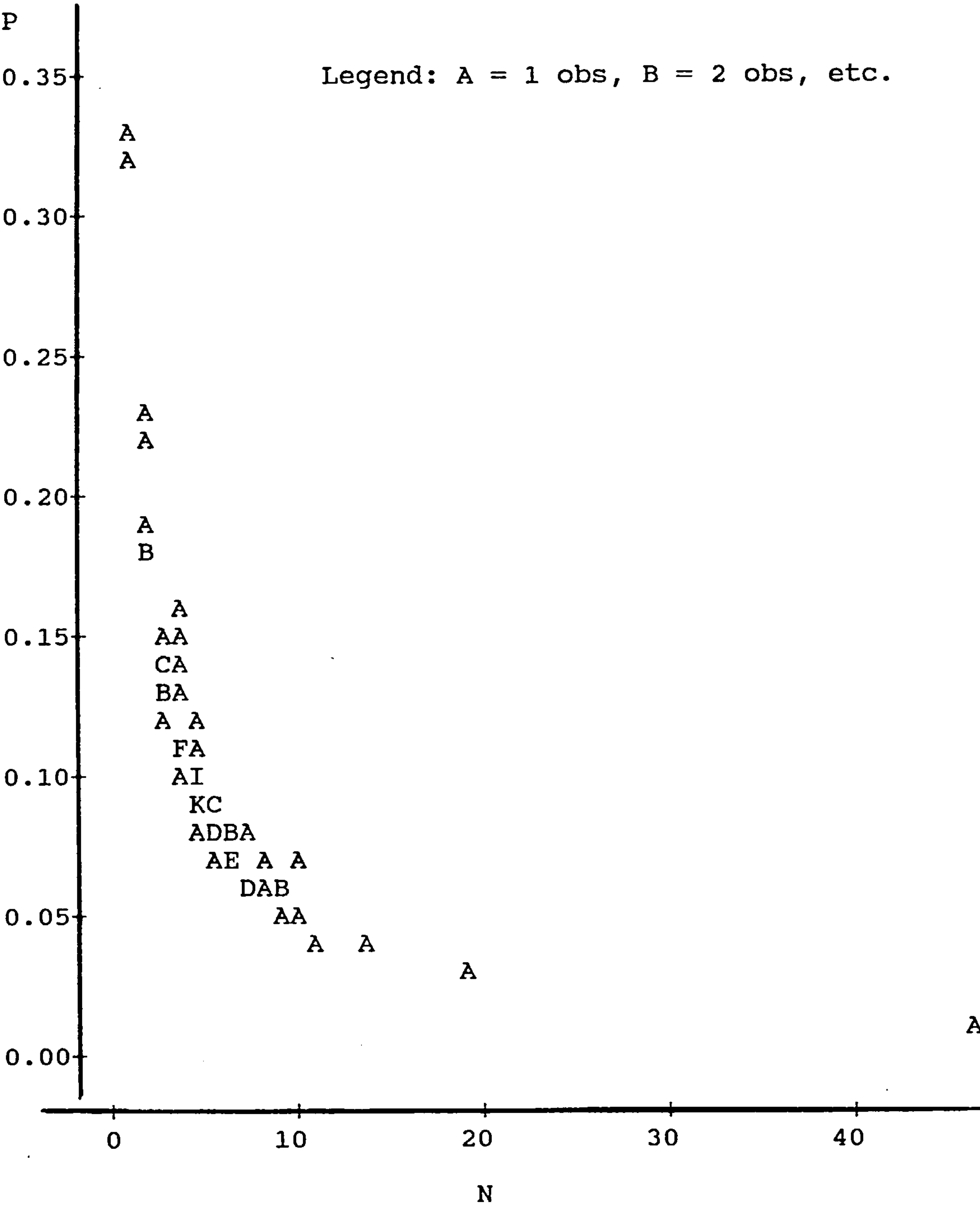


FIGURE 13. The parameters P and N obtained from the Negative Binomial distributions fitted to the female sub-populations.



From Figure 12 and Figure 13 we observed that the parameter estimates for P were generally larger in the

extreme values for males than for females. For both males and females there is strong evidence of a relationship between the parameter estimates for P and N and hence the reported accident mean since the mean equals $N \times P$. Looking at the results across gender it was observed that the large estimates for P correspond to the small parameters for N , and that the small estimates for P correspond to the large estimates for N .

In summary, in this and the previous section we have identified several reported accident trends and shown that the full reported accident distribution cannot be modelled adequately using the traditional accident modelling distributions, the Poisson and Negative Binomial distributions. When examining the poor fitting Negative Binomial models across gender we found strong evidence of accident misreporting between particular accident counts. The same pattern of misreporting could not be seen when examining the poor fitting Poisson models.

Higher reported accident means were seen in males and females who displayed particular undesirable personality characteristics and this can be considered as evidence of accident proneness. This type of phenomenon would favour the fitting of the Negative Binomial model (see Section 2.2). However, small sample sizes also mean that the tail of the accident distribution is too small to be modelled using the two parameter Negative Binomial accident model.

We now examine the results obtained when modelling a subset of the full reported accident distribution, i.e. the serious reported accident distribution.

3.5 FITTING THE TRADITIONAL MODELLING DISTRIBUTIONS TO THE SERIOUS REPORTED ACCIDENT DISTRIBUTION

We suspect that different types of accidents are recalled with different degrees of accuracy so in this section we define a different reported accident outcome variable based on the level of seriousness of the reported accident. This reported accident distribution is then modelled with the traditional accident probability distributions.

We define the serious reported accident distribution to include only fractures, dislocations and lacerations requiring stitches, i.e. accidents resulting in serious injury. We omit sprains, burns, scalds, ingestions, foreign body type injuries, superficial injuries and injuries not identified. This definition of 'serious' reported accidents was arbitrary and provided us with a convenient starting point.

The reasoning behind this line of investigation is that we have assumed that the reporting of serious accidents is more accurate than the reporting of all accidents that occurred. On this assumption we therefore expect the reported serious accident distribution to be very like the 'true' serious accident distribution.

A separate Poisson and a separate Negative Binomial distribution were fitted to the serious reported accident distribution for each male and female sub-population as defined earlier and the results from this distribution

modelling are compared with the earlier modelling results from Chapter 3.1.

For each sub-population we would expect the serious reported accident mean to be lower than the full reported accident mean, and for the tail of the serious reported accident distribution to be shorter than the tail of the full reported accident distribution. Without doubt this will effect the distribution modelling since there will be some sub-populations with a serious reported accident distribution too short to be modelled by the Negative Binomial model.

We will again look at the results from the social index split first and then move on to examine the results for all the splits.

For males we found that after splitting by social index the lowest reported serious accident mean was seen in the most advantaged social index sub-population and the highest reported serious accident mean was seen in the most disadvantaged social index sub-population, results which are similar to modelling the full reported accident distribution. A Negative Binomial model was not fitted to the disadvantaged sub-population because the tail of the reported serious accident distribution was too short to be modelled.

For females we found that the lowest reported serious accident mean was seen in the most advantaged social index sub-population and the highest reported serious accident mean was seen in the disadvantaged social index sub-population. Again, results which are similar to modelling

the full reported accident distribution.

The Negative Binomial model was not fitted to any of the serious reported accident distributions from the female sub-populations obtained from the social index split because the tail of each reported serious accident distribution was too short to be modelled.

Now examining the overall results of the serious reported accident distribution modelling we observe that the reported accident mean for males was higher than for females. Table 11 gives a summary of the reported serious accident means by gender.

TABLE 11. Reported serious accident means of the male and female sub-populations.

SAMPLE MEAN RANGE	Frequency	
	MALE	FEMALE
0.12 to < 0.14	0	1
0.14 to < 0.16	0	2
0.16 to < 0.18	0	35
0.18 to < 0.20	0	22
0.20 to < 0.22	3	8
0.22 to < 0.24	2	2
0.24 to < 0.26	14	0
0.26 to < 0.28	36	0
0.28 to < 0.30	15	0
0.30 to < 0.32	6	0
>= 0.32	8	0
TOTAL	84	70

From Table 11 we observe that the majority of the male serious reported accident means were between 0.24 and 0.30. The range of male serious reported accident means was from 0.2059 to 0.3605 which is smaller than the range for the male full reported accident means which was 0.5625

to 0.8880. The males with the five highest serious reported accident means are as follows with the sub-population serious reported accident mean in brackets.

- i) Males who certainly destroyed belongings (0.3605).
- ii) Males from a household with a low/very low standard of household equipment (0.3464).
- iii) Males from a home which is untidy/chaotic (0.3373).
- iv) Males who certainly squirmed and fidgeted (0.3366).
- v) Males from the East Midlands (0.3333).

Comparing the five highest male reported accident mean sub-populations for the serious and the full reported accident distributions we see that they include males who came from an untidy/chaotic household, males who certainly squirmed and fidgeted and males who certainly destroyed belongings.

For the females we observe from Table 11 that the majority of the serious reported accident means were between 0.16 and 0.20 with a range from 0.1386 to 0.2306 which is smaller than the range for the female full reported accident means which was from 0.3506 to 0.8095.

The females with the five highest serious reported accident means were from the following sub-populations with the sub-population accident mean in brackets.

- i) Females who certainly squirmed and fidgeted (0.2306).
- ii) Females whose mothers often had violent rages (0.2225).
- iii) Females from an untidy/chaotic household (0.2158).
- iv) Females who certainly destroyed belongings (0.2157).
- v) Females whose mothers were nervous (0.2044).

Comparing the five highest female reported accident mean sub-populations for the serious and the full reported accident distributions we see that they include females

whose mothers often had violent rages and females who certainly destroyed belongings.

For both males and females the five highest serious accident mean sub-populations include children who came from an untidy/chaotic household, children who certainly squirmed and fidgeted and children who destroyed belongings.

Several serious reported accident frequency trends were identified which were similar to the earlier trends identified in the full reported accident distribution. For example, for both males and females an increase in the age of a parent was associated with a decrease in the serious reported accident mean. Similarly, the parents smoking habits were associated with the same accident frequency trends seen earlier for each sex.

When considering the child's neighbourhood we found that for males the earlier result was repeated but that for females the result was slightly different. It was seen that females from an average neighbourhood had the highest reported serious accident mean, then well-to-do and rural neighbourhoods (the serious reported accident distribution for females from a poor neighbourhood was too short to be modelled).

When examining specific characteristics of the child's personality as assessed by the child's mother we found that higher reported accident means were associated with children who displayed particular undesirable personality characteristics. We now discuss some anomalous results.

3.6 ANOMALOUS RESULTS AND THEIR INVESTIGATION

It was found that variables relating to the mother's malaise condition resulted in different accident frequency trends when compared to the earlier results. In this section we will examine this anomaly further. Using the serious reported accidents only we found that the reported accident mean was actually lower for children whose mothers suffered from certain malaise conditions than for children whose mothers did not suffer, in direct contradiction of the results from the full reported accident distribution. The malaise conditions of the mothers for which these anomalous results were obtained were 'being keyed-up and jittery', and having 'suffered at one time from a nervous breakdown'. For females this same phenomenon was seen when looking at the whether the child's mother had 'suffered at one time from a nervous breakdown'.

The results for the serious reported accidents appear plausible since one would expect mothers of a keyed-up and jittery nature to be more wary and sensitive about potential danger to their children. This could mean that the child protected from the more dangerous situations which result in serious accidental injury. In contradiction however, we also note that females with mothers who suffered from nerves, i.e. were nervous, were included in the five female highest reported serious accident mean sub-populations. From this we believe that

there may be some accident misreporting with respect to some aspects of the mother's malaise conditions.

To explain how the contradictory results arise we examine Tables 12 and 13 which give respective breakdowns of male reported accidents with respect to whether the mother was keyed-up and jittery, and whether she had ever suffered a nervous breakdown.

TABLE 12. A breakdown of the type of male reported accidents that occurred with respect to whether the male's mother was keyed-up and jittery. The figures in brackets are percentages of the total number of reported accidents in the male population.

Type of accident	Whether the male's mother was keyed-up and jittery	
	YES	NO
Bone fracture	120 (25%)	944 (25%)
Lacerations/stitches	47 (10%)	535 (14%)
Dislocation/sprain	12 (3%)	121 (3%)
Burn/scald	43 (9%)	365 (9%)
Ingestion	62 (13%)	489 (13%)
Superficial injury	88 (18%)	697 (18%)
Other/unidentified	108 (23%)	692 (18%)
Mean number of accidents	0.74	0.69
Size of sub-population	647	5575

From Table 12 we see that males of mothers who were keyed-up and jittery had a higher reported accident mean than males of mothers who were not keyed-up and jittery. However, looking just at the serious accident category 'lacerations requiring stitches' we observe that males of mothers who were not keyed-up and jittery had a higher mean number of reported lacerations requiring stitches

than males of mothers who were keyed-up and jittery. This would account for the result observed when modelling the 'serious' reported accidents alone.

Looking now at Table 13 which gives a breakdown of male reported accidents with respect to whether the mother ever had a nervous breakdown we see that males of mothers who had a nervous breakdown had a higher reported accident mean than males of mothers who had not had a nervous breakdown. Looking just at the accident category 'bone fracture' we observe that males of mothers who did not have a nervous breakdown had a higher mean number of reported 'bone fractures' than males of mothers who did have a nervous breakdown. This would account for the result observed when modelling the 'serious' reported accidents only.

TABLE 13. A breakdown of the type of reported male accidents that occurred with respect to whether the male's mother had ever had a nervous breakdown. The figures in brackets are percentages of the total number of reported accidents in the male population.

Type of accident	Whether the male's mother had a nervous breakdown	
	YES	NO
Bone fracture	42 (17%)	1023 (25%)
Lacerations stitches	32 (13%)	547 (13%)
Dislocation/sprain	5 (2%)	127 (3%)
Burn/scald	20 (8%)	389 (10%)
Ingestion	39 (16%)	518 (13%)
Superficial injury	46 (19%)	743 (18%)
Other/unidentified	60 (25%)	741 (18%)
Total number of accidents	244	4088
Size of sub-population	292	5946

A similar investigation of the contradictory results for the female sub-population is seen in seen in Table 14, which gives a breakdown of reported accidents with respect to whether the mother ever suffered a nervous breakdown.

TABLE 14. A breakdown of the type of female reported accidents that occurred with respect to whether the female's mother had ever had a nervous breakdown. The figures in brackets are percentages in terms of the total number of reported accidents in the female population.

	Whether the female's mother had a nervous breakdown	
Type of accident	YES	NO
Bone fracture	24 (16%)	615 (21%)
Lacerations stitches	12 (8%)	318 (11%)
Dislocation/sprain	4 (3%)	123 (4%)
Burn/scald	20 (14%)	292 (10%)
Ingestion	23 (16%)	450 (16%)
Superficial injury	40 (27%)	513 (18%)
Other	24 (16%)	550 (19%)
Total number of accidents	147	2861
Size of sub-population	247	5515

From Table 14 we observe that females with mothers who had a nervous breakdown had a higher reported accident mean than those whose mothers had not had a nervous breakdown. However, looking just at the accident category 'bone fracture' we observe that females with mothers who did not have a nervous breakdown had a higher mean number of reported bone fractures than those whose mother's had a nervous breakdown. Looking also at the category of superficial accidents we observe that females of mothers who had a nervous breakdown had a higher mean number of

reported superficial accidents than females of mothers who had not had a nervous breakdown. This could account for the result observed when modelling the 'serious' reported accident distribution.

From these results it is clear that the full reported accident population can be broken down into smaller sub-populations of different accident types. Hence, in the following section we investigate the serious reported accident distribution in more detail. A redefinition of the reported accident outcome variable is discussed and assessed in Chapter 5. For example, when redefining the reported accident outcome variable we investigate the modelling of only the most reliably reported accidents.

3.7 ASSESSING THE TRADITIONAL PROBABILITY DISTRIBUTIONS FOR THE MALE AND FEMALE SERIOUS ACCIDENT SUB-POPULATIONS

We now examine the results of fitting the traditional probability distributions to the serious reported accident distributions of the male and female sub-populations. The goodness-of-fit results are given in Table 15 and it can be seen that the Poisson model produced an acceptable fit at the one percent significance level for 41 (49%) of 84 male sub-populations and for 48 (69%) of the 70 female sub-populations. At the one percent level the Negative Binomial model was acceptable for 36 (68%) of the 53 male models and for all of the 32 female models. From this we

conclude that in terms of the Negative Binomial model, the serious reported accident distribution can be considered to be 'purer' or more accurately reported than the full reported accident distribution.

TABLE 15. Goodness-of-fit results from fitting the Poisson and Negative Binomial distributions to the serious reported accident distributions of the male and female sub-populations.

		Level of significance for the Chi-squared values obtained from the model fits				
Sex	Distrib.	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
MALE	POISSON	43	7	3	31	84
	N.B.	17	14	6	16	53
FEMALE	POISSON	22	17	9	22	70
	N.B.	0	1	1	30	32

We now examine the parameter estimates for P and N obtained from fitting the Negative Binomial distribution to the serious reported accident distribution of the male and female sub-populations.

It was observed that the parameter estimates for P were generally larger in the extreme values for the males than for the females and there was strong evidence of a relationship between the parameter estimates for P and N in both males and females. It was observed across gender that the larger estimates for P corresponded to the smaller estimates for N, and that the smaller estimates for P corresponded to the larger estimates for N, a result already seen when modelling the full reported accident distribution with the Negative Binomial distribution.

When comparing the respective parameter estimates obtained from modelling the full and serious reported accident distributions of each male and female sub-population it was seen that the estimates for P and N were smaller in the models fitted to the serious reported accident distribution. This is to be expected since the mean of the reported serious accident distribution is smaller than the mean of the full reported accident distribution.

By examining the reported and predicted serious accident counts in the poor fitting models we see at which counts the model performed badly. It was observed that the reported one serious accident count was underestimated by the Poisson model whereas the tail of the reported accident distribution was overestimated by the model. This suggests that there was misreporting from the higher serious accident counts to the single reported accident count.

In the poor fitting Negative Binomial models the poor fit generally occurred in the two accident count and the tail of the distribution. The model usually overestimated the two accident count and underestimated the tail of the distribution. This suggests accident misreporting from the lower serious accident counts to higher serious accident counts which is unacceptable to us since we have already assumed that misreporting can occur only in the opposite direction, i.e. from higher reported accident counts to lower reported accident counts.

This misreporting trend is the complete opposite to the misreporting trend seen when modelling the full

reported accident distribution using the Negative Binomial distribution. We therefore suspect that the Negative Binomial distribution will provide a good model to a reported accident distribution that is somewhere between the full and the serious reported accident distribution.

We now examine the results of fitting the generalised traditional probability distributions to the full reported accident distribution.

3.8 FITTING THE GENERALISED TRADITIONAL DISTRIBUTIONS TO THE FULL REPORTED ACCIDENT DISTRIBUTION

In this section of the analysis the new models derived in Section 2.3 were applied, that is Generalised Poisson and Generalised Negative Binomial distributions were fitted to the full reported accident distributions of the male and female sub-populations.

We recall that no obvious trend emerged from examining the poor fitting Poisson models yet when examining the poor fitting Negative Binomial models a trend was identified. Across gender and in almost every case, the poor fit was attributable to the one and two accident counts where the Negative Binomial model consistently underestimated the one accident count and overestimated the two accident count. An illustration of this has already been given in Chapter 2 (see Section 2.2).

The next step in the model-fitting therefore involved extending the Poisson and Negative Binomial models to take

into account accident misreporting. This would allow for the phenomenon seen in the poor fitting Negative Binomial models at the one accident and two accident counts. The models fitted in this section have already been defined in Chapter 2.3 and are described as follows.

- Model I. The Generalised Poisson distribution with accident misreporting from one to zero accident.
- Model II. The Generalised Poisson distribution with accident misreporting from two to one accident.
- Model III. The Generalised Poisson distribution with accident misreporting from one to zero accident, together with misreporting from two to one accident.
- Model IV. The Generalised Poisson distribution with accident misreporting from one to zero accident, together with misreporting from two to one accident, estimated with one parameter.
- Model V. The Generalised Poisson distribution with accident misreporting from one to zero accident, together with misreporting from two to one accident, estimated with one parameter.
- Model VI. The Generalised Negative Binomial distribution with accident misreporting from two to one accident.

Before we examine the results of modelling the full reported accident distribution with the generalised distributions let us first select a 'typical' sub-population and follow it through the model fitting to illustrate the logic behind our modelling methodology.

From the male population we select the sub-population of 2,702 males who had no household moves during 1970-1975. This sub-population had a reported accident mean of

0.6392 and fitting the traditional accident probability distributions produced poor results when fitted to the full reported accident distribution. The Chi-squared test statistic for the Poisson model was 30.3066 (3 d.f.) which gave a p-value of 0.000, and for the Negative Binomial model, 11.3663 (2 d.f.) which gave a p-value of 0.003.

Since both traditional accident models produced poor fits we then attempted to identify a misreporting trend or pattern by comparing the reported with the predicted accident counts. No trend was visible when examining the Poisson model but using the same technique on the Negative Binomial model we found that the model overestimated the two accident count and underestimated the single accident count.

As accident misreporting was expected we then began to systematically explore the misreporting of particular accident counts. The simple, single parameter Poisson distribution was generalised to take into account the misreporting of a single accident as no accident and this was called Model I. This generalised model resulted in a higher predicted accident mean of 0.6524 which is to be expected since we are in effect increasing the frequency of the one accident count whilst lowering the frequency of the no accident count. Model I produced a Chi-squared value of 27.4986 (2 d.f.) with a p-value of 0.000, still a poor fit. The reduced number of degrees of freedom in the Chi-squared goodness-of-fit test is due to the estimation of the extra parameter to take into account accident misreporting.

No pattern emerged when examining the reported and predicted accident counts for this model so the simple Poisson distribution was again generalised but this time to take into account misreporting from two accidents to one accident. This was called Model II.

From fitting Model II we increased the predicted accident mean to 0.6528 and the model produced a Chi-squared value of 23.8286 (2 d.f.) with a p-value of 0.000, again a poor fit. No pattern emerged when examining the reported and predicted accident counts so the simple Poisson distribution was again generalised but this time to allow for the misreporting of two accidents as one accident, together with the misreporting of one accident as no accident. This was called Model III.

Model III increased the predicted accident mean to 0.7492 and gave a greatly improved fit when compared to the previous two generalised models. This Poisson model produced a Chi-squared test statistic of 2.7160 (1 d.f.), with a p-value of 0.099.

As an extension to Model III it was thought that perhaps one parameter could be used to estimate all the accident misreporting. In this way we could increase the number of degrees of freedom available for the Chi-squared goodness-of-fit test and perhaps improve the fit of the model. In addition, it might also be possible to fit the model to the shorter reported accident distributions. Hence, Model IV and Model V were developed.

To find a relationship between the two misreporting parameters the pairs of parameter estimates obtained from

Model III for each sub-population were examined and compared. From this examination we decided to estimate B in two ways, as follows.

For Model IV, $B = a \times \theta$, and

for Model V, $B = \theta^a$,

where B is the parameter of misreporting from one to zero accident, θ is the parameter of misreporting from two to one accident, and a is a constant.

From Model IV we obtained a predicted accident mean of 0.7570 and a Chi-squared value of 3.6459 (2 d.f.) with a p-value of 0.162. From Model V we obtained a predicted accident mean of 0.7495 and a Chi-squared value of 3.9559 (2 d.f.) which gave a p-value of 0.138. As one would expect there is little difference in the fit of these two models and both provide excellent goodness-of-fit results.

The motivation for developing the Generalised Negative Binomial model, Model VI, was that the poor-fitting Negative Binomial model failed at the one and two accident counts. Model VI was therefore a Generalised Negative Binomial model which took into account the misreporting of two accidents as one accident.

From fitting Model VI we obtained a predicted accident mean of 0.6553 and a Chi-squared value of 0.2775 (2 d.f.) with a p-value of 0.871. As the results obtained from fitting Model VI were so favourable it was thought unnecessary to generalise the Negative Binomial distribution any further.

We now examine the parameters and the predicted

accident means obtained from fitting the generalised distributions to the male and female sub-populations. We then discuss the Chi-squared results in order to select the generalised traditional distribution which best describes the full reported accident distribution.

Table 16 gives the predicted accident means for each generalised distribution fitted and as one would expect, the reported accident mean for each sub-population was lower than than the predicted accident mean obtained for each generalised model. This is quite simply because the generalised models effectively increased the higher reported accident counts and decreased the lower reported accident counts. The overall effect therefore is an increase in the the total number of accidents aathat occurred in each sub-population. It was also found throughout the modelling that the predicted accident means were higher for the male sub-populations than for the corresponding female sub-populations.

When examining the results obtained from fitting Model I across gender we were unable to identify any relationship between the parameter estimates for L , the predicted mean and B the parameter of misreportng between zero and one accident. The estimates of B ranged from 0.00 to 0.2002 in the models fitted to the male sub-populations and ranged from 0.0070 to 0.1961 in the models fitted to the female sub-populations. No trend was seen when comparing the reported and predicted accident counts within each sub-population.

TABLE 16. Predicted accident means obtained from each generalised probability distribution fitted to the full reported accident ditribution of the male and female sub-populations.

SEX	PREDICTED ACCIDENT MEAN RANGE	Frequencies for each Model						
		Sample	I	II	III	IV	V	VI
M A L E	0.00 to < 0.45	0	0	0	0	0	0	0
	0.45 to < 0.50	0	0	0	0	0	0	0
	0.50 to < 0.55	0	0	0	0	0	0	0
	0.55 to < 0.60	4	2	2	0	0	0	1
	0.60 to < 0.65	8	4	4	0	0	1	1
	0.65 to < 0.70	28	18	24	2	3	2	20
	0.70 to < 0.75	25	24	24	8	9	11	26
	0.75 to < 0.80	8	13	10	18	24	22	11
	0.80 to < 0.85	9	4	6	15	16	13	4
	0.85 to < 0.90	6	8	8	12	17	16	7
	0.90 to < 0.95	0	2	2	5	4	8	2
	>=0.95	0	0	0	6	12	8	0
	TOTAL	88	75	80	66	85	85	72
F E M A L E	0.00 to < 0.45	3	1	1	0	0	0	1
	0.45 to < 0.50	18	3	10	0	3	0	2
	0.50 to < 0.55	39	38	35	0	17	4	30
	0.55 to < 0.60	11	13	10	8	29	14	11
	0.60 to < 0.65	6	9	5	22	18	32	4
	0.65 to < 0.70	3	3	4	10	3	15	2
	0.70 to < 0.75	3	3	3	2	4	6	1
	0.75 to < 0.80	0	1	1	0	2	3	0
	0.80 to < 0.85	1	0	0	2	2	2	0
	0.85 to < 0.90	0	0	0	0	0	3	0
	0.90 to < 0.95	0	0	0	0	1	0	0
	>=0.95	0	0	0	0	0	0	0
	TOTAL	84	71	69	44	79	79	51

Similar results were found when fitting Model II and no trend was identified when comparing the reported and predicted accident counts for each poor fitting model. In addition, no relationship was found between the parameter estimates for L, the predicted accident mean and θ the parameter of misreporting from two to one accident. In the models fitted to the male sub-populations, θ ranged

from 0.0103 to 0.4264 and for females from 0.0133 to 0.3696.

As Model I and Model II had produced such poor Chi-squared goodness-of-fit results it was thought that Model III, which is in a sense a combination of Model I and Model II might give more favourable Chi-squared results. As expected, based on the Chi-squared results Model III did prove to be the better model out of these three models. Examining the parameter estimates for Model III we find no evidence suggesting a relationship between the parameter L the predicted mean and θ the parameter for misreporting between two and one accident, or between L and B , the parameter for misreporting one accident as no accident. In the models fitted to the male sub-populations θ ranged from 0.0791 to 0.4753, whilst in the females θ ranged from 0.2610 to 0.4154. In the models fitted to the male sub-populations the estimates for B ranged from 0.0491 to 0.3623, whilst for the females B ranged from 0.1735 to 0.2752. We did observe some evidence to suggest a linear relationship between θ and B for the models fitted to the male sub-populations, but this relationship was not as apparent in the models fitted to the female sub-populations. With the exception of two male sub-populations the estimate for θ was always greater than the estimate for B .

By redefining Model III so all the misreporting can be estimated using only one parameter we increase the number of degrees of freedom available for the Chi-squared goodness-of-fit test. We could have defined B in any

number of ways and because of the extra degree of freedom this provides we anticipated that the model could be fitted to more sub-populations and produce better goodness-of-fit results than Model III. The parameter B was defined as follows;

$$B = a \times \theta,$$

where $a = 0.6738$ for males and $a = 0.6524$ for females.

The constant a was obtained by calculating the ratio of (B/θ) for each of the Model III fits and taking the mean of these ratios. This model was called Model IV and gave excellent goodness-of-fit results. In the models fitted to the male sub-populations, θ ranged from 0.0154 to 0.5779 and in the females θ ranged from 0.0154 to 0.6035. There was no evidence to suggest any relationship between L and θ .

Once again extending Model III so that the misreporting can be estimated using only one parameter, B was redefined as follows;

$$B = \theta^a,$$

where $a = 1.3$ for males and $a = 1.4$ for females.

The values of 1.3 and 1.4 were obtained from an examination of the power relationship between B and θ across sub-populations. This model was called Model V and gave Chi-squared results similar to those obtained from fitting Model IV. In the models fitted to the male sub-populations θ ranged from 0.0087 to 0.4938, and for the females from 0.0085 to 0.5689. Again there was no

evidence to suggest a relationship between L and θ .

As expected there was little difference between the parameter estimates or the Chi-squared results obtained from fitting Model IV or Model V to the sub-populations. In comparison to Model III the extra degree of freedom allowed us to fit these two models to more sub-populations and also gave better Chi-squared goodness-of-fit results.

Examining the results obtained from fitting Model VI (the Generalised Negative Binomial distribution) to the male and female sub-populations we found that this model also produced excellent goodness-of-fit results.

From examining the parameter estimates it was observed that in each of the models fitted to the male and female sub-populations, P increased and N decreased when compared with the respective estimates of P and N in the corresponding Negative Binomial model of the full reported accident distribution. θ , the parameter of misreporting from two to one accident ranged from 0.0052 to 0.3324 for males and from 0.0326 to 0.4388 for females.

We found that large estimates for P corresponded to small estimates of N and that small estimates for P corresponded to large estimates of N , a trend identified earlier. There was no evidence to suggest a relationship between the parameters P and θ and we observed that generally the females produced lower estimates for N than the males and once again there was strong evidence of a relationship between the parameter estimates for P and N .

3.9 COMPARISON OF THE GENERALISED DISTRIBUTIONS

The question we must now ask is which of these generalised traditional distributions gives us the most sensible interpretation of the reported accident distribution? Examining the results for the males first we look at Table 17 which gives the goodness-of-fit results obtained from fitting each of the generalised traditional distributions to the male sub-populations.

TABLE 17. Goodness-of-fit results from fitting the Poisson, Negative Binomial, Generalised Poisson and Negative Binomial distributions to the full reported accident distributions of the male sub-populations. The numbers in brackets are row percentages in terms of the TOTAL.

Distribution Model	Level of significance for the Chi-squared values obtained from the model fits for the males				
	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
Poisson	52 (59)	10 (11)	7	19	88
N.B.	14 (16)	26 (31)	11	34	85
Model I	50 (67)	11 (15)	2	12	75
Model II	50 (63)	10 (13)	5	15	80
Model III	9 (14)	14 (21)	3	40	66
Model IV	5 (6)	8 (9)	10	62	85
Model V	4 (5)	9 (11)	10	62	85
Model VI	0 (0)	2 (3)	1	69	72

From fitting Model I or Model II to the male sub-populations we did not improve upon the simple, single parameter Poisson model Chi-squared fit. Indeed, for some sub-populations the p-value obtained from the simple Poisson model was better than the p-value obtained from the Generalised Poisson model. This was mostly due to the

reduction in the degrees of freedom resulting from the estimation of the extra parameter needed for the accident misreporting. For example, examining the results for the sub-population of males from Wales we see that the Chi-squared test statistic obtained from fitting the simple Poisson distribution was 6.9329 (2 d.f.) giving a p-value of 0.031, whereas the Chi-squared test statistic obtained from fitting Model I to the same male sub-population was 6.9182 (1 d.f.) which gave a p-value of 0.0085. The Chi-squared test statistics obtained from each model were not very different but the removal of a degree of freedom for the estimation of the misreporting parameter made a noticable difference when obtaining the p-value for the goodness-of-fit of the generalised model.

The simple Poisson model was fitted to every male sub-population but gave very poor results. Poor results were also observed for Model I and Model II. Model III was fitted to only 66 (75%) of the 88 male sub-populations and at the one percent significance level gave slightly better goodness-of-fit results than the Negative Binomial model. Model IV and Model V gave very similar results to each other which is understandable since both models were based on Model III.

We can also see from Table 17 that Model VI gave excellent results but was fitted to only 72 (82%) of 88 male sub-populations. This is because the estimation of three parameters reduced the degrees of freedom available for the Chi-squared goodness-of-fit test. A similar situation occurred with Model III although the Chi-squared

results were certainly not as favourable.

When choosing the best fitting generalised traditional model for the male sub-populations we see that Model IV, Model V and Model VI gave the most favourable results. Model VI gave the better Chi-squared results but was fitted to only 82% of the male sub-populations. In terms of misreporting this model suggests that misreporting occurred from the two accident count to the one accident count, and that accidents were reported accurately throughout the rest of the accident distribution.

These assumptions are reasonable when we examine the construction of the full reported accident distribution in more detail in Chapter 5. In Chapter 5 the full reported accident distribution is broken down into non-superficial and superficial/unidentified reported accidents and the simple Negative binomial distribution provides an adequate model for the non-superficial reported accident population. Since the superficial/unidentified reported accident distribution contains fewer reported accidents out of the two distributions and also thought to be less accurately reported it does seem likely that accident misreporting did occur in the lower accident counts (see Section 5.3).

For Model IV and Model V we assume that misreporting occurred from the single accident count to the no accident count together with misreporting from the two accident count to the single accident count. These are also reasonable assumptions since it is quite plausible that misreporting occurs in just these three counts. Model IV

and Model V are extensions of Model III and the exceptionally good results from these two models are due to the systematic way in which they were developed. For example, to obtain either of these models we first had to assess the relationship between the misreporting parameters by examining the results of fitting Model III. For this reason we have therefore selected Model VI as the model which best represents the full reported accident distribution of the male population.

We now consider the best generalised traditional model to describe the full reported accident distribution of the female population.

Table 18 gives the goodness-of-fit details obtained from fitting each of the generalised traditional distributions to the female sub-populations. We can see from the table that the simple Poisson model, the Negative Binomial model, Model I and Model II gave poor goodness-of-fit results. At the one percent significance level the best model fits came from Model III, Model IV, Model V and Model VI. Model VI was fitted to 51 (61%) of the 84 female sub-populations, Model III was fitted to just over half of the female sub-populations and both Model IV and Model V were fitted to almost all the female sub-populations.

As discussed in the earlier section, Model IV and Model V are merely extensions of Model III, so although the results from fitting these two models are quite exceptional we shall reduce our selection to Model III or Model VI. At the five percent significance level we must

select Model VI over Model III and therefore conclude that Model VI is the generalised traditional distribution which best describes the full reported accident distribution of the female population.

TABLE 18. Goodness-of-fit results from fitting the Poisson, Negative Binomial, Generalised Poisson and Negative Binomial distributions to the full reported accident distributions of the female sub-populations. The numbers in brackets are row percentages in terms of the TOTAL.

Distribution Model	Level of significance for the Chi-squared values obtained from the model fits for the females				
	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
Poisson	48 (57)	12 (14)	2	22	84
N.B.	22 (28)	18 (23)	5	34	79
Model I	49 (69)	9 (13)	2	11	71
Model II	50 (72)	8 (12)	8	8	69
Model III	2 (5)	12 (27)	6	24	44
Model IV	0 (0)	4 (5)	12	63	79
Model V	0 (0)	5 (6)	13	61	79
Model VI	0 (0)	3 (6)	1	47	51

The assumptions inherent in using Model VI are reasonable when we recall that in Chapter 5 the full reported accident distribution can be broken down into non-superficial and superficial/unidentified reported accidents. Since the superficial/unidentified reported accident distribution is the shorter of the two distributions and also thought to be less accurately reported it does seem likely that accident misreporting did occur in the lower accident counts.

Having selected Model VI as the most appropriate generalised model for both the male and the female

populations we now discuss the results obtained from fitting this generalised model in more detail. When examining the predicted accident means we observed that the majority of the predicted accident means for males were between 0.65 and 0.75, with a range from 0.5882 to 0.9200 and recall that the range for the male full reported accident means was from 0.5625 to 0.8880.

The males with the five highest predicted accident means are as follows with the sub-population predicted accident mean in brackets.

- i) Males from a household with a low/very low standard of household equipment (0.9200).
- ii) Males who certainly squirmed and fidgetted (0.9075).
- iii) Males who certainly destroyed belongings (0.8977).
- iv) Males whose mothers often had violent rages (0.8935).
- v) Males with more than three household moves (0.8814).

Comparing these five male sub-populations with the five male highest reported accident means we see that they have in common males from a household with a low/very low standard of household equipment, males who certainly squirmed and fidgetted, males who certainly destroyed belongings and males whose mothers often had violent rages.

For the females we observed that the majority of the predicted accident means were between 0.50 and 0.55, with a range from 0.4792 to 0.7337. We recall that the range for the female full reported accident means was from 0.3506 to 0.8095.

The females with the five highest predicted accident means were are as follows with the sub-population

predicted accident mean in brackets.

- i) Females whose mothers often had violent rages (0.7337).
- ii) Females whose mothers were keyed-up (0.6861).
- iii) Females whose mothers were nervous (0.6533).
- iv) Females who somewhat bullied others (0.6450).
- v) Females whose household was untidy/chaotic (0.6395).

Comparing these five female sub-populations with the five highest female reported accident sub-populations we see that they have in common only those females whose mothers often had violent rages.

Now comparing the five highest male and female predicted accident mean sub-populations we see that they have in common only those children whose mothers often had violent rages.

In conclusion we have shown that for both the male and female populations the full reported accident distribution can be successfully modelled using Model VI. This is a Generalised Negative Binomial distribution which takes into account accident misreporting from the two to the one accident count, and where we also assume that the rest of the reported accident distribution is reported accurately.

This is a sensible model for the data since and it was found in the later analysis (see Section 5.2) that the full reported accident distribution could be broken down into a non-superficial and superficial reported accident distribution and that each of these two reported accident distributions could be modelled adequately with a separate Negative Binomial distribution.

3.10 DISCUSSION

In this Chapter the traditional accident models (the Poisson and Negative Binomial probability distributions) were fitted to the reported accident distributions of sub-populations of the full data set and produced poor fits. However, after generalising the probability distributions to take into account accident misreporting we found that the new Model VI, developed by including a misreporting parameter that takes into account misreporting two accidents as one accident, was found to provide a much better fit to the data than the other accident models.

In terms of misreporting, Model VI suggests that misreporting occurred from the two accident count to the one accident count, and that accidents were reported accurately throughout the rest of the accident distribution. The assumptions in this model are supported by the results discussed later in Chapter 5.

In Chapter 5 the full reported accident distribution is broken down into non-superficial and superficial/unidentified reported accidents and the simple Negative binomial distribution provides an adequate model for the non-superficial reported accident population. Since the superficial/unidentified reported accident distribution contains fewer reported accidents out of the two distributions, and also thought to be less accurately

reported it does seem likely that accident misreporting did occur in the lower accident counts (see Section 5.3).

CHAPTER 4

ANALYSIS OF THE BRITISH BIRTH DATA SET, PART II; ALTERNATIVE PROBABILITY DISTRIBUTIONS

4.1 INTRODUCTION

In this Chapter we model the full reported accident distribution using probability distributions which have been used in the past as alternative accident probability distributions. These two parameter probability distributions were introduced in Section 2.4 and include the Neyman Type A distribution, the Poya-Eggenberger distribution, the Consul/Jain distribution (a Generalised Poisson distribution) and the compounded Poisson-Binomial distribution.

From these four alternative probability distributions the best fitting alternative model for both sexes was selected as being the Consul and Jain Generalised Poisson distribution. It was found that the poor fitting Consul/Jain models usually overestimated the two accident count and underestimated the one accident count, hence the distribution was generalised to take into account accident misreporting from the two to the one accident count.

The next step in the analysis was to investigate the results obtained when truncating the reported accident

distribution at zero and modelling the smaller reported accident distributions with the truncated Poisson and truncated Negative Binomial distributions. These distributions were introduced in Chapter 2.5. Analysing the data after truncating the reported zero count is analagous to the situation where data for the zero count cannot be collected, i.e. in the analysis of hospitalised accidents where the number of patients with zero accidents is not usually observed. The results from modelling the truncated data provide support for the hypothesis that accident misreporting is present in the full data set.

We searched for evidence of accident proneness in males and females separately by calculating the Pearson correlation coefficients between the numbers of accidents occurring in two consecutive time periods. Three different sets of consecutive time periods were used. High correlation would perhaps suggest that accident proneness was present and further investigation would be neccessary. However, recall of reported accident times was such that 'heaps' of accidents were reported to have occurred at six monthly intervals, i.e., the child's age at the time of the accident was rounded to either 6 months old, 12 months, 18 months, 24 months, and so on. This casts doubt on any analysis based on the reported accident times.

We now discuss the results of the distribution modelling.

4.2 FITTING ALTERNATIVE PROBABILITY DISTRIBUTIONS TO THE FULL REPORTED ACCIDENT DISTRIBUTION

In this section we fit the alternative probability distributions to the male and female sub-populations and select the best fitting alternative probability model from these distributions. These alternative two parameter probability distributions have already been introduced and discussed in Section 2.4.

Table 19 gives the goodness-of-fit results obtained from fitting the alternative distributions to the male sub-populations and from the table we are able to compare and select the best fitting of the alternative probability models.

TABLE 19. Goodness-of-fit results from fitting the alternative probability distributions to the full reported accident distributions of the male sub-populations. The numbers in brackets are row percentages in terms of the TOTAL.

Distribution Model	Level of significance for the Chi-squared values obtained from the model fits to the males				
	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
Poisson	52 (59)	10 (11)	7	19	88
N.B.	14 (16)	26 (31)	11	34	85
Polya-Eggen.	31 (36)	24 (28)	2	28	85
Neyman Type A	27 (33)	30 (37)	2	23	82
Consul/Jain	12 (14)	36 (42)	9	29	86
Poisson-Binom.	25 (29)	22 (26)	8	30	85

From Table 19 we can see that at the one percent significance level the best fitting alternative

probability model for the male sub-populations was the Consul/Jain Generalised Poisson distribution. However, there was little difference between the Chi-squared results obtained from fitting the Negative Binomial or the Consul/Jain models.

From Table 20 we can select the best fitting alternative probability distribution to the female sub-populations based on the p-values obtained from the Chi-squared goodness-of-fit test. At the one percent significance level the best fitting alternative model to the female sub-populations was the Consul/Jain Generalised Poisson distribution, although once again there was little difference between the fit of Negative Binomial model and the Consul/Jain model. The Consul/Jain distribution was also the best fitting alternative distribution for the male sub-populations.

TABLE 20. Goodness-of-fit results from fitting the alternative probability distributions to the full reported accident distributions of the female sub-populations. The numbers in brackets are row percentages in terms of the TOTAL.

Distribution Model	Level of significance for the Chi-squared values obtained from the model fits to the females				
	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
Poisson	48 (58)	12 (14)	2	21	83
N.B.	22 (28)	18 (23)	5	34	79
Polya-Eggen.	31 (39)	14 (18)	8	27	80
Neyman Type A	29 (37)	17 (22)	12	20	78
Consul/Jain	22 (28)	20 (25)	7	31	80
Poisson-Binom.	26 (33)	17 (22)	7	29	79

Several similarities were found when looking at the Chi-squared results obtained from fitting these alternative distributions to the male and female sub-populations. When comparing the p-values obtained from fitting the Negative Binomial model with the p-values obtained from fitting the alternative models we found that with the exception of the Consul/Jain distribution the Negative Binomial distribution usually provided the better model. With the exception of the Poisson-Binomial distribution it was also found that the alternative models overestimated the two accident count and underestimated the one accident count. The Poisson-Binomial model generally underestimated the zero accident count and overestimated the single accident count.

The next step in the analysis was to generalise the Consul/Jain distribution to take into account the misreporting of two accidents as one accident, estimated by the parameter θ .

Table 21 gives the goodness-of-fit results obtained from fitting this generalised model to the male and female sub-populations and we observe that at the one percent significance level the Generalised Consul/Jain model gave an acceptable fit to all of the 69 male and all of the 50 female sub-populations to which it was fitted. No relationship was seen between the parameter estimates for L_1 , L_2 or θ the parameter of misreporting. The estimates for θ for the models fitted to the male sub-populations ranged from 0.0029 to 0.3394, and for the females, from 0.0997 to 0.3676. The estimates for θ were very similar

to those estimates for θ obtained from fitting Model VI, the Generalised Negative Binomial model with misreporting from the two to the single accident counts.

TABLE 21. Goodness-of-fit results from fitting the Generalised Consul/Jain distribution to the full reported accident distributions of the male and female sub-populations.

Level of significance for the Chi-squared values obtained from the Generalised model fits					
Sex	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
MALE	0	0	1	68	69
FEMALE	0	1	3	46	50

Summarising this section we found that for both the males and females the best alternative probability model for the full reported accident distribution was the Consul/Jain model. This model gave similar goodness-of-fit results as those obtained from fitting the Negative Binomial distribution. The Consul/Jain distribution was then generalised to take into account the misreporting seen in the two and the one accident counts and this produced excellent goodness-of-fit results.

We now examine the results obtained when truncating the full reported accident distribution at the zero count and fitting the truncated traditional distributions to this shortened reported accident distribution.

4.3 FITTING THE TRUNCATED POISSON AND NEGATIVE BINOMIAL PROBABILITY DISTRIBUTIONS TO THE TRUNCATED REPORTED ACCIDENT DISTRIBUTION

In this section we investigate the results obtained after restricting our data set to include only those children who had one or more reported accident. This is analagous to a situation where an accident data set consists only of hospitalised accidents, i.e. there would be no reported zero count since hospitals generally only come into contact with injured patients. We have an advantage with our accident data in that after fitting the truncated distributions we are able to compare the predicted zero count with the reported zero count. The truncated distributions considered here were introduced in Section 2.5 and are the truncated Poisson distribution and the truncated Negative Binomial distribution.

Before fitting the truncated distributions we note that by removing children who had no reported accidents, the male population was reduced from 6,310 to 3,058, a reduction of 52%, and the female population from 5,830 to 2,295, a reduction of 61%. We also note that by truncating the population at zero we reduce the degrees of freedom available for the Chi-squared goodness-of-fit test.

A separate truncated Poisson distribution was fitted to 84 male and 73 female sub-populations based on the earlier sub-population splits. For 21 of the male and 16

of the female sub-populations a better fit in terms of the p-value was obtained from the full Poisson model than from the truncated Poisson model. For the remaining 62 male and 57 female sub-populations, the p-value from the truncated Poisson model was at least equal to the p-value obtained from the full Poisson model.

Examining the truncated Poisson models (21 male and 16 female) which produced a better fit for the truncated rather than the full model we found that in all but two models (one male and one female) the truncated Poisson Chi-squared test statistic was generally the same as the full Poisson model Chi-squared test statistic, however the degrees of freedom were fewer for the truncated model which made the difference when calculating the p-value.

The predicted zero count obtained from each truncated Poisson model was compared with the reported zero count obtained from the corresponding full sub-population and it was found that the predicted zero count was below the reported zero count for all but six of the 84 male sub-populations, and for all but one of the 71 female sub-populations. It should also be noted that for each of these seven models the ratio of the predicted zero count over the reported zero count was only just larger than one which should indicate that the full Poisson distribution produces a satisfactory model. This is confirmed from examining the Poisson model p-values where it was found that the full Poisson model did give an adequate fit at the five percent significance level to each of these seven sub-populations.

It was also found that the ratio of the predicted zero count over the reported zero count was not always a good indicator of the fit of either the full Poisson or truncated Poisson models. One would have thought that if the predicted zero count was very close to the reported zero count then both the full Poisson and truncated Poisson models would provide a good fit to the data, however this was not observed. For example, excluding the six above mentioned sub-populations and selecting the male sub-population with the next highest ratio of 0.9583, the p-value obtained from fitting the full Poisson model was 0.014 and from the truncated Poisson model was 0.007. This is one of the 21 sub-populations for which the full Poisson distribution gave a better fit than the truncated Poisson.

A summary of the goodness-of-fit results obtained from fitting the truncated Poisson distribution to the male and female sub-populations is given in Table 22.

TABLE 22. Goodness-of-fit results obtained from fitting the truncated Poisson distribution to the truncated reported accident distributions of the male and female sub-populations.

Level of significance for the Chi-squared values obtained from the truncated Poisson model					
Sex	< 1%	1% - 5%	5% - 10%	>10%	TOTAL
MALE	46	14	6	18	84
FEMALE	39	13	3	16	71

From Table 22 we observe that at the one percent

significance level the truncated Poisson distribution gave an adequate fit to 38 (45%) of the 84 male sub-populations, and to 32 (45%) of the 71 female sub-populations.

Table 23 gives a summary of the predicted accident means obtained from fitting the truncated Poisson model to the male and female sub-populations. The predicted accident means for the males ranged from 0.6059 to 1.1458, and for the females from 0.4801 to 0.8653. We recall that the range of the male full reported accident means was from 0.5625 to 0.8880 and for females was from 0.3506 to 0.8095. Three of the truncated male sub-populations had predicted accident means above 1.0 and two of the female truncated sub-populations had predicted accident means above 0.8. Clearly there is a difference in the predicted accident means between gender.

TABLE 23. Predicted accident means obtained from fitting the truncated Poisson distribution to the truncated reported accident distributions of the male and female sub-populations.

PREDICTED MEANS	Frequency	
	MALE	FEMALE
0.45 to < 0.50	0	2
0.50 to < 0.55	0	11
0.55 to < 0.60	0	29
0.60 to < 0.65	3	14
0.65 to < 0.70	8	9
0.70 to < 0.75	18	2
0.75 to < 0.80	17	2
0.80 to < 0.85	18	1
0.85 to < 0.90	6	1
0.90 to < 0.95	8	0
>= 0.95	6	0
TOTAL	84	71

The males with the five highest truncated accident means were from the following sub-populations with the predicted sub-population accident mean in brackets.

- i) Males who certainly bullied others (1.1458).
- ii) Males whose mothers had experienced a nervous breakdown (1.0510).
- iii) Males who certainly destroyed belongings (1.0113).
- iv) Males whose mother's highest educational qualification was 'A' level standard (0.9725).
- v) Males with more than three household moves (0.9640).

Comparing these five male sub-populations with the five male sub-populations with the highest full reported accident means we note that they only have one sub-population in common, i.e. males who certainly destroyed belongings.

The females with the five highest truncated accident means are as follows with the sub-population accident mean in brackets We should note that the five highest female truncated accident means are from different sub-populations than the five highest male truncated accident means.

- i) Females from a household with a low/very low standard of household equipment (0.8653).
- ii) Females with mothers who often had violent rages (0.8322).
- iii) Females with mothers who were keyed up (0.7783).
- iv) Females with father's under 20 years old in April 1970 (0.7732).
- v) Females with nervous mothers (0.7321).

Comparing these five female sub-populations with the five highest full reported accident mean sub-populations we note that they include females whose mother's often had violent rages and females with father's under 20 years old

in April 1970.

We now examine the results obtained from fitting the truncated Negative Binomial distribution to the male and female sub-populations. A separate truncated Negative Binomial distribution was fitted to each of 70 male and 48 female sub-populations. For 14 male and three female sub-populations the full Negative Binomial model provided a better model than the truncated Negative Binomial model and for the remaining 56 male and 45 female sub-populations the p-value for the full Negative Binomial model was at least equal to the p-value obtained from the truncated Negative Binomial model.

Examining the full Negative Binomial models which produced a better fit than the truncated models we found that in all but one female model the Chi-squared test statistic for the truncated model was generally the same as the Chi-squared test statistic obtained from the full model but that the degrees of freedom were fewer in the truncated model. Hence, this made the difference when calculating the p-value. This phenomenon was also seen when fitting the full and truncated Poisson distributions to the data.

When examining parameter estimates for P and N we see that regardless of gender small estimates for P corresponded to large estimates for N and large estimates for P corresponded to small estimates for N. This association was also seen across gender when fitting the full Negative Binomial distribution and also the Generalised Negative Binomial distribution (Model VI) to

the sub-populations.

Comparing the reported zero count with the predicted zero count we found that the predicted zero count was above the reported zero count for 69 of the 70 male sub-populations and for all of the 48 female sub-populations. This is in direct contrast to the trend observed when fitting the truncated Poisson distribution to the truncated data where the predicted zero count was generally below the reported zero count.

If we recall that the full reported accident distribution can be adequately modelled with a Generalised Negative Binomial distribution which allows for accident misreporting from the two to the one accident count, then the results of fitting the truncated Negative Binomial model to the truncated accident data are to be expected. Since the misreporting occurs in the two and the one accident counts we would expect the truncated Negative Binomial model to overestimate the reported zero count and underestimate the reported accident mean, a result which was observed.

We would also expect the truncated Negative Binomial model to provide a good fit to the truncated data and a summary of the goodness-of-fit results is given in Table 24. From this table we observe that the truncated Negative Binomial distribution produced an excellent model for the truncated data. At the one percent significance level the model gave an adequate fit to all the 70 male sub-populations and to 47 (98%) of the 48 female sub-populations.

TABLE 24. Goodness-of-fit results obtained from fitting the truncated Negative Binomial distribution to the truncated male and female reported accident distributions.

Level of significance for the Chi-squared values obtained from the truncated N.B. model					
Sex	< 1%	1% - 5%	5% - 10%	>10%	TOTAL
MALE	0	2	6	62	70
FEMALE	1	6	0	41	48

Comparing Table 22 and Table 24 we see that for both the male and female sub-populations the truncated Negative Binomial distribution produced more favourable Chi-squared results than the truncated Poisson distribution. Table 25 gives a summary of the predicted accident means obtained from fitting the truncated Negative Binomial distribution to the male and female sub-populations.

TABLE 25. Predicted accident means obtained from fitting the truncated Negative Binomial distribution to the truncated male and female reported accident distributions.

PREDICTED MEANS	Frequency	
	MALE	FEMALE
0.00 to < 0.20	1	3
0.20 to < 0.25	0	15
0.25 to < 0.30	1	6
0.30 to < 0.35	3	12
0.35 to < 0.40	2	7
0.40 to < 0.45	6	3
0.45 to < 0.50	11	0
0.50 to < 0.55	19	2
0.55 to < 0.60	12	0
0.60 to < 0.65	7	0
0.65 to < 0.70	2	0
0.70 to < 0.75	3	0
0.75 to < 0.80	2	0
0.80 to < 0.85	1	0
TOTAL	70	48

The predicted male accident means for the truncated Negative Binomial models ranged from 0.1322 to 0.8361, and for the females from 0.1429 to 0.5452. These can be compared with the predicted accident means obtained from the truncated Poisson models which ranged from 0.6059 to 1.1458 for the males, and from 0.4801 to 0.8653 for the females. If we also recall that the range of the full reported accident means for the males was from 0.5625 to 0.8880 and from 0.3506 to 0.8095 for the females we see that truncated Poisson model predicts the reported accident mean above the full reported accident mean which is in contrast with the truncated Negative Binomial model which gives a prediction below the full reported accident mean.

From the truncated Negative Binomial models the male sub-populations with the five highest predicted accident means are as follow with the predicted accident mean in brackets.

- i) Males with more than three household moves (0.8361).
- ii) Males who certainly destroyed belongings (0.7761).
- iii) Males who certainly squirmed or fidgeted (0.7585).
- iv) Males from the North of England (0.7345).
- v) Males from the North West of England (0.7259).

Comparing these five male sub-populations with the five male sub-populations with the highest full reported accident means we note that they include males who certainly destroyed belongings and males who certainly squirmed and fidgeted. Comparing them with the five male sub-populations with the highest reported accident means from the truncated Poisson models we see only one sub-

population in common, i.e. males who certainly destroyed belongings.

The females with the five highest predicted accident means obtained from the truncated Negative Binomial models are as follows with the predicted accident mean in brackets. We note that there are no sub-populations in common with the males.

- i) Females with mothers who were keyed up (0.5452).
- ii) Females with nervous mothers (0.5447).
- iii) Females with mothers who smoked (0.4485).
- iv) Females whose mothers often had violent rages (0.4264).
- v) Females with mother who were depressed (0.4048).

Comparing these five female sub-populations with the five female sub-populations with the highest full reported accident means we note that they have only one sub-population in common, i.e. females whose mother's often had violent rages. Comparing them with the five female sub-populations with the highest reported accident means from the truncated Poisson models we see that they include females with mother's who were keyed up, females with nervous mothers and females whose mother's often had violent rages.

From the results of fitting the truncated Poisson and truncated Negative Binomial models to the male and female sub-populations, we conclude that a good fitting full model does not automatically mean a good fitting truncated model, and vice-versa. It was found that the truncated Poisson distribution underestimated the zero count and overestimated the full reported accident mean which was in

direct contrast to the results obtained when fitting the truncated Negative Binomial distribution. The truncated Negative Binomial model overestimated the zero count and hence underestimated the full reported accident mean.

The results of fitting the truncated traditional accident modelling distributions confirmed the presence of accident misreporting and validated the use of the generalised models which allowed for accident misreporting. We recall that the full reported accident distribution can be adequately modelled with a Generalised Negative Binomial distribution which allows for accident misreporting from the two to the one accident count. Hence, the results of fitting the truncated Negative Binomial model to the truncated accident data are to be expected. Since the misreporting occurs in the two and the one accident counts we would expect the truncated Negative Binomial model to overestimate the reported zero count and underestimate the reported accident mean, a result which was observed.

The truncated Poisson distribution provided a poor model for the truncated data and the truncated Negative Binomial model produced an excellent model. However, when deciding whether or not the truncated Negative Binomial model was a good model for the truncated data it is thought that too much information is lost by the truncation of the zero accident count for the model to be considered appropriate.

In the next part of the analysis we attempted to find evidence of accident proneness by calculating the

coefficients of correlation between the numbers of accidents occurring in consecutive time periods.

4.4 LOOKING FOR EVIDENCE OF PRONENESS

The aim of this part of the analysis was to search for evidence of accident proneness in the data. We calculated the coefficients of correlation between the number of reported accidents that occurred in two consecutive time periods. If strong correlation were to appear then this might suggest support for the 'proneness' assumptions of the Negative Binomial models fitted to the reported accident distribution.

Those children with missing accident dates were excluded from this part of the analysis. Hence, the male population was reduced from 6,310 to 6,245 males, a reduction of one percent and the females from 5,830 to 5,788, also a reduction of one percent.

Correlation coefficients were also calculated using a sub-set of this data set which contained only those children who had one or more accident. The male population was reduced from 6,245 to 2,993 males, a reduction in the total male sub-population of 53% and the female population from 5,788 to 2,253, a reduction in the total female population of 61%.

Three different sets of consecutive time periods were used in the correlation analysis. The first set of time periods were defined on a yearly basis, i.e. 0-12 months

inclusive, 13-24 months inclusive, 25-36 months inclusive, 37-48 months inclusive and 49-60 months inclusive.

The second set of time periods were also defined on a yearly basis but started at six months since it was observed that few accidents were actually reported during the first six months of a child's life. The time periods were redefined as 6-17 months inclusive, 18-29 months inclusive, 30-41 months inclusive and 42-53 months inclusive.

The third set of time periods were defined to split the five year accident reporting period into two consecutive time periods of 30 months each, i.e. 0-30 months inclusive and 31-60 months inclusive. In addition a Poisson distribution and a Negative Binomial distribution were fitted to each marginal reported accident distribution obtained from the third set of time periods. This was to investigate the possibility of bivariate distribution modelling.

Before examining the correlation coefficients we should first assess the reliability of the reported accident times. The reported accident times are of primary importance when placing each reported accident in proper time sequence. Frequency histograms of reported accident times were constructed for both the male and female sub-populations.

It was clear from the separate histograms of reported male and reported female accident times that there was a 'heaping' of reported accident times beginning at six-months and repeated at six-monthly intervals. For both of

the reported accident time distributions the median was at 36 months and 62% of all the reported accidents occurred between 24 months and 48 months inclusively.

A likely explanation for this heaping of reported accident times is that the parents recorded the accident times by rounding them up or down to the nearest year or half year. We speculate that some parents rounded the times up whilst others rounded the times down. This highlights the fact that we are dealing with reported rather than actual accidents and also casts doubt on the reliability of an analysis based on these reported accident times.

Another type of accident misreporting which might be present is in the situation where the chronological order of the accidents is incorrect. For example, consider a male who has a non-superficial accident at 14 months and a superficial accident at 15 months. It is possible that the parents misreport the non-superficial accident to occur at 12 months and then misreport the superficial accident to occur later at 18 months. Hence, the chronological order of the accidents is incorrect. This type of misreporting is also likely to be present in our data.

Examining the reported accident correlation coefficients for the full male population we saw that regardless of which set of consecutive time periods were considered, the coefficients were always positive and below 0.10. The highest correlation coefficient was 0.09595 (significant at the one percent level) for the

time periods 25-36 months inclusive and 37-48 months inclusive.

Examining the reported accident correlation coefficients for the full female population we observed one negative correlation coefficient of -0.0421 (significant at the one percent level) for the time periods 0-30 months inclusive and 31-60 months inclusive. With this exception the remaining coefficients were always positive. The highest was 0.0602 (significant at the one percent level) for the time periods 37-48 months inclusive and 49-60 months inclusive.

When comparing correlation coefficients between gender it was noticed that there was no significant correlation between the consecutive time periods 0-12 months and 13-24 months, or between 6-17 months and 18-29 months for the males but for the females we observed significant correlation coefficients at the one percent level of 0.0436 and 0.0520 respectively. Although overall some correlation coefficients were significant none were above 0.10 and it was thought that the large sample size contributed to this. It was also noted that the male reported accident mean for each time period was always larger than the female reported accident mean for the corresponding time period.

The Poisson and Negative Binomial distributions were fitted to the marginal reported accident distributions obtained from the third set of time periods, 0-30 months inclusive and 31-60 months inclusive.

At the one percent significance level it was found

that the Poisson model gave an adequate fit to the reported male accident distribution for the first time period, but not the second time period. At the one percent level the Poisson distribution provided a poor fitting model for each of the two female marginal reported accident distributions. The Negative Binomial distribution was found to give a satisfactory fit at the ten percent significance level to all four of the marginal reported accident distributions.

The next part of the correlation analysis involved restricting the data set to those children who had one accident or more. Using the 16 different time periods already defined we observed that the female reported accident mean was greater than the male reported accident mean for the corresponding sub-population on five occasions.

Examining the accident correlation coefficients for the male population we see that regardless of which set of consecutive time periods we consider, the coefficients were always significant at the one percent level, always negative and always below -0.1035 . The best coefficient was -0.4370 for the time periods 0-30 months inclusive and 31-60 months inclusive. The negative correlation here suggests that a high number of reported accidents in the first thirty months was associated with a low number of reported accidents in the last thirty months, and that a low number of accidents in the first thirty months was associated with a high number of accidents in the last thirty months.

Examining the reported accident correlation coefficients for the female population over all the time periods we note that the coefficients were always significant at the one percent level, always negative and always below -0.0852. The best correlation coefficient was -0.5230 for the time periods 0-30 months inclusive and 31-60 months inclusive.

When examining the male and female marginal reported accident distributions obtained from the third set of time periods it was noted that the reported accident mean was always greater than the variance. It was also observed that at the one percent significance level the Poisson model gave a very poor fit to each male and female marginal reported accident distribution.

In summary, it is difficult to interpret the correlations found here as we have little confidence in the reported accident times which clearly are not recorded accurately. For example, there is heaping of reported accident times in six monthly intervals and it is also likely that the chronological order of the reported accident times are also inaccurate.

4. 5 DISCUSSION

The results from this Chapter have shown that the Consul/Jain distribution provided the best fitting 'alternative' model to the full data set. This distribution was then generalised to take into account the

misreporting seen in the two and the one accident counts and produced excellent goodness-of-fit results. The results confirm the presence of accident misreporting and emphasise the need for the 'misreporting' parameter.

The truncated traditional accident modelling also confirmed the presence of accident misreporting and validated the use of Model VI used in Section 3.8. For example, since the misreporting occurs in the two and the one accident counts we would expect the truncated Negative Binomial model to overestimate the reported zero count and underestimate the reported accident mean of the full data set, a result which was observed.

The results from the correlation analysis proved inconclusive. This is because it is difficult to have confidence in the reported accident times which clearly are not recorded accurately. For example, there is heaping of reported accident times in six monthly intervals and it is also likely that the chronological order of the reported accident times are also inaccurate.

CHAPTER 5

ANALYSIS OF THE BRITISH BIRTH DATA SET, PART III; NON-SUPERFICIAL REPORTED ACCIDENTS

5.1 INTRODUCTION

In this Chapter an alternative approach is presented to the problem of unreliably reported accidents. Anomalous results in Section 3.6 were uncovered which were found to be due to differences between the full reported accident population and the sub-set of accidents reported as 'serious'. In Section 3.7 we stated the possibility that the Negative Binomial distribution might provide a good model to a reported accident distribution that is somewhere between the full and the serious reported accident distribution. To investigate this possibility the reported accident population has been redefined and the earlier accident distribution modelling repeated.

In this Chapter the analysis of reported accident populations is built up cumulatively from the very serious through to the less serious and the unidentified, and led to the definition of the 'non-superficial' reported population. It was found that data on non-superficial reported accidents can be successfully modelled by the simple Negative Binomial model. A plausible explanation for this finding is that misreporting is minimal for these

accidents.

The non-superficial reported accident distribution is defined to include reported accidents which resulted in fractures, lacerations requiring stitches, burns, scalds, ingestions, dislocations and sprains. The remaining reported accidents, that is the superficial and unidentified reported accidents will be called the 'superficial' reported accidents. The full reported accident population is therefore obtained from the addition of the non-superficial reported accidents and the superficial/unidentified reported accidents.

The reasons for removing superficial and unidentified reported accidents are simple ones. We recall that by definition we are examining reported accidents which require medical advice or treatment. We would therefore expect superficial accidents to require little or no medical advice/treatment. We might even consider the presence of superficial reported accidents as a reflection of how the definition of the 'seriousness' of an accident varies according to parents.

In terms of the unidentified reported accidents we note that a reported accident may not have been identified for a number of reasons, i.e. misreporting, forgotten details, too trivial, etc., and that the 'true' reported accident might have been anything from a superficial accident to a bone fracture. For this reason we believe that the category of unidentified reported accidents contains a wide range of accident outcomes, some of which may have been serious and some superficial, and should

also be removed.

The conclusions from analysing the non-superficial reported accidents are compared with those from fitting Model VI, the Negative Binomial model with an extra parameter to account for misreporting two accidents as one accident, to the full data set. Finally a comparison of the approaches in Chapter 5 and Chapter 3 is made.

5.2 MODELLING DIFFERENT ACCIDENT OUTCOMES

Before examining the results of the 'non-superficial' reported accident modelling we will first examine the cumulative effects of adding in less and less serious reported accidents into the reported accident distribution. In doing so we further justify our definition of the 'non-superficial' reported accident distribution.

In our rating of accident 'seriousness' we will rate bone fractures as being the most serious accident, and then in descending order of 'seriousness', lacerations requiring stitches, dislocations, burns or scalds, ingestions or foreign body accidents, sprains, superficial and then unidentified accidents. Note that this is not a strict medical rating of accident seriousness. We first fit the traditional probability distributions to reported bone fractures alone, and then to bone fractures and lacerations requiring stitches, and then to bone fractures and lacerations requiring stitches and dislocations, and

so on. For example, we fit the Poisson and Negative Binomial distributions to the following outcome variables.

- (a) Bone fractures.
- (b) (a) + lacerations requiring stitches.
- (c) (b) + dislocations.
- (d) (c) + burns and scalds.
- (e) (d) + ingestions and foreign body type accidents.
- (f) (e) + sprains, i.e. non-superficial accidents.
- (g) (f) + superficial accidents.
- (h) (g) + unidentified accidents, i.e. the full reported accident distribution.

The goodness-of-fit results from fitting the traditional accident modelling distributions to these cumulative reported accident distributions are given in Table 26.

Looking at the results for the males we see from Table 26 that the Negative Binomial model gave an excellent fit for reported 'bone fractures'. However, after then adding in 'lacerations requiring stitches' into the reported accident outcome variable the fit became much worse. Thereafter, with the addition of the less serious reported accidents into the reported accident outcome variable the fit of the Negative Binomial gradually improved and for the 'non-superficial' reported accident distribution the Negative Binomial model produced a Chi-squared value of 7.11 (2 d.f.) and a p-value of 0.029. With the addition of superficial reported accidents the fit improved further, shown by a Chi-squared value of 5.22 (3 d.f.) and

a p-value of 0.157. After adding in 'unidentified' reported accidents the Negative Binomial then produced a poor fit shown with a Chi-squared value of 15.72 (3 d.f.) and a p-value of 0.001.

TABLE 26. The modelling details of fitting the Poisson and Negative Binomial models to cumulative reported accident distributions from the male and female populations. 'OUT' is the reported accident outcome variable defined as follows; a=bone fractures, b=a+lacerations requiring stitches, c=b+dislocations, d=c+burns or scalds, e=d+ingestions or foreign body accidents, f=e+sprains, g=f+superficial accidents, h=g+unidentified accidents.

SEX	OUT	Mean	Var.	Poisson fit (df)	P- value	N.B. fit (df)	P- value
M A L E	a	0.1707	0.1866	17.50(1)	0.000	0.16(1)	0.693
	b	0.2643	0.3007	84.27(2)	0.000	9.76(1)	0.002
	c	0.2697	0.3054	79.64(2)	0.000	10.12(1)	0.002
	d	0.3349	0.3657	36.92(2)	0.000	9.62(2)	0.008
	e	0.4239	0.4643	44.55(3)	0.000	7.15(2)	0.028
	f	0.4396	0.4807	43.94(3)	0.000	7.11(2)	0.029
	g	0.5662	0.6299	35.35(3)	0.000	5.22(3)	0.157
	h	0.6957	0.8029	63.43(3)	0.000	15.72(3)	0.001
F E M A L E	a	0.1110	0.1200	7.42(1)	0.006	-	-
	b	0.1681	0.1900	12.38(1)	0.000	1.35(1)	0.246
	c	0.1775	0.2013	11.82(1)	0.001	1.82(1)	0.177
	d	0.2317	0.2587	21.94(2)	0.000	1.53(1)	0.216
	e	0.3134	0.3511	23.73(2)	0.000	4.66(2)	0.097
	f	0.3257	0.3675	33.33(2)	0.000	6.11(2)	0.047
	g	0.4176	0.4668	39.16(3)	0.000	8.38(2)	0.015
	h	0.5208	0.5903	54.87(3)	0.000	15.52(2)	0.000

Looking now at the results for the females we see from Table 26 that the Negative Binomial model was not fitted to reported bone fractures because the tail of the distribution was too short to be modelled. However, after adding in 'lacerations requiring stitches' into the reported accident outcome variable the tail lengthened and

the Negative Binomial model produced an excellent fit. Thereafter, with the addition of the less serious reported accidents the fit of the Negative Binomial grew steadily worse and for the 'non-superficial' reported accident distribution the Negative Binomial model produced a Chi-squared value of 6.11 (2 d.f.) and a p-value of 0.047. With the addition of superficial reported accidents the fit became worse, shown by a Chi-squared value of 8.38 (2 d.f.) and a p-value of 0.015. After adding in 'unidentified' reported accidents the Negative Binomial model then produced a Chi-squared value of 15.52 (2 d.f.) and gave a p-value of 0.000, a poor fit.

We now redefine our accident 'seriousness' rating and again examine the cumulative effects of adding in less and less serious reported accidents into the reported accident distribution. Again we stress that this is not a strict medical rating of accident seriousness. By adding in less serious reported accidents we can monitor how the full reported accident distribution takes shape and precisely where it fails.

In order of accident 'seriousness' we will now rate head fractures together with lacerations of the head requiring stitches as being the most serious types of accident. In descending order of 'seriousness' we now rate accident 'seriousness' as follows.

ACC1 = Head fractures together with lacerations of the head requiring stitches.

ACC2 = ACC1 + other bone fractures/dislocations.

ACC3 = ACC2 + other lacerations requiring stitches.

ACC4 = ACC2 + ingestions, burns, scalds and foreign body type accidents.

ACC5 = ACC4 + other lacerations requiring stitches.

ACC6 = ACC5 + sprains, i.e. non-superficial accidents.

ACC7 = ACC6 + superficial and unidentified accidents, i.e. the full reported accident distribution.

We again fit the traditional accident modelling probability distributions to the cumulative reported accident distributions. The goodness-of-fit results from this cumulative reported accident distribution modelling are given in Table 27.

Looking first at the results for the males we see that the Negative Binomial model gave an excellent fit for reported 'head fractures' together with 'head lacerations requiring stitches'. The fit was also excellent after adding in 'other bone fractures' and 'dislocations'. However, after then adding in 'other lacerations requiring stitches' the Negative Binomial model then produced a Chi-squared value of 10.1236 (1 d.f.) and a p-value of 0.002. The Negative Binomial provided an excellent model for the ACC4 outcome variable, i.e. all bone fractures, all lacerations requiring stitches, ingestions, burns, scalds and foreign body type accidents, and then continued to prove adequate when adding in less and less serious accidents until we obtained the full reported accident distribution. Here the Negative Binomial model fails at the one percent significance level.

TABLE 27. The modelling details of fitting the Poisson and Negative Binomial models to cumulative reported accident distributions from the male and female populations.

SEX	ACC.	Mean	Var.	Poisson fit (df)	P- value	N.B. fit (df)	P- value
M A L E	ACC1	0.1745	0.1983	24.96(1)	0.000	1.80(1)	0.179
	ACC2	0.2271	0.2520	37.17(2)	0.000	0.33(1)	0.568
	ACC3	0.2697	0.3054	79.64(2)	0.000	10.12(1)	0.002
	ACC4	0.3813	0.4103	18.92(2)	0.000	2.93(2)	0.232
	ACC5	0.4239	0.4643	44.55(3)	0.000	7.15(2)	0.028
	ACC6	0.4396	0.4807	43.94(3)	0.000	7.11(2)	0.029
	ACC7	0.6957	0.8029	63.43(3)	0.000	15.72(3)	0.001
F E M A L E	ACC1	0.0926	0.1022	8.67(1)	0.003	-	-
	ACC2	0.1472	0.1640	9.63(1)	0.002	2.33(1)	0.127
	ACC3	0.1775	0.2013	11.82(1)	0.001	1.82(1)	0.177
	ACC4	0.2830	0.3100	19.18(2)	0.000	3.85(1)	0.050
	ACC5	0.3134	0.3511	23.73(2)	0.000	4.66(2)	0.097
	ACC6	0.3257	0.3675	33.33(2)	0.000	6.11(2)	0.047
	ACC7	0.5208	0.5903	54.87(3)	0.000	15.52(2)	0.000

Looking now at the results for the females we see from Table 27 that the Negative Binomial model was not fitted to reported 'head fractures' and 'lacerations of the head requiring stitches' because the tail of the distribution was too short to be modelled. However, after adding in 'other bone fractures' and 'dislocations' into the accident outcome variable the tail lengthened and the Negative Binomial model produced an excellent fit. The Negative Binomial then continued to prove adequate when adding in less and less serious accidents until we obtained the full reported accident distribution. Here (as in the male population) the Negative Binomial model fails at the one percent significance level.

Comparing the results shown in Table 26 and Table 27 we note that other levels of reported accident

'seriousness' could be defined and investigated in this way. However, from the results of these two tables and the earlier analysis we now feel justified in exploring the 'non-superficial' reported accident distribution in more detail.

5.3 MODELLING THE NON-SUPERFICIAL REPORTED ACCIDENT DISTRIBUTION USING TRADITIONAL MODELS

We shall now discuss the results obtained from fitting the traditional probability distributions to the 'non-superficial' reported accident distributions from the male and female sub-populations. For each sub-population we would expect the non-superficial reported accident mean to be lower than the full reported accident mean and yet higher than the reported serious accident mean. In addition, the tail of the distribution will be shorter than the tail of the full reported accident distribution and yet longer than the reported serious accident distribution tail.

The male and female populations were again split into sub-populations as defined in Section 3.2 and the non-superficial reported accident distribution for each sub-population was modelled with the simple Poisson distribution and the Negative Binomial distribution.

5.3.1 SUMMARY OF OVERALL RESULTS

In considering the results overall, we observe once again that the reported accident means for the male sub-populations are higher than for the corresponding female sub-populations. Table 28 gives a summary of the non-superficial reported accident means.

TABLE 28. Reported non-superficial accident means of the male and female sub-populations.

SAMPLE MEAN RANGE	Frequency	
	MALE	FEMALE
0.00 to < 0.28	0	2
0.28 to < 0.30	0	4
0.30 to < 0.32	0	23
0.32 to < 0.34	0	24
0.34 to < 0.36	0	14
0.36 to < 0.38	5	5
0.38 to < 0.40	2	5
0.40 to < 0.42	8	1
0.42 to < 0.44	24	1
0.44 to < 0.46	20	1
0.46 to < 0.48	7	1
0.48 to < 0.50	5	0
0.50 to < 0.52	5	0
0.52 to < 0.54	4	2
0.54 to < 0.56	5	0
>= 0.56	2	0
TOTAL	87	83

As expected the non-superficial reported accident means are higher than the corresponding serious reported accident means, yet lower than the corresponding full reported accident means. For the males we observed a range of non-superficial reported accident means from 0.3620 to 0.5872, with the majority between 0.42 and 0.46.

We recall that the range of the male full reported accident means was from 0.5625 to 0.8880 and for the male serious reported accident means was from 0.2059 to 0.3605.

Comparing the five highest male accident mean sub-populations for the non-superficial and the full reported accident distributions we found that they have in common males who certainly squirmed and fidgeted and males who certainly destroyed belongings.

Now comparing the five highest male accident mean sub-populations for the non-superficial and the serious reported accident distributions we see that they include males from an untidy/chaotic household, males who certainly squirmed and fidgeted and males who certainly destroyed belongings.

For the females we see the majority of the non-superficial reported accident means were between 0.16 and 0.20, with a range from 0.2240 to 0.5238. We recall that the range of the female full reported accident means was from 0.3506 to 0.8095, and for the serious reported accident means was from 0.1386 to 0.2306.

Comparing the five highest female accident mean sub-populations for the non-superficial and the full reported accident distributions we found that they include females who bullied others, females whose father's were less than 20 years old in April 1970, females who certainly destroyed belongings and females whose mother's often had violent rages.

Now comparing the five highest female accident mean sub-populations for the non-superficial and the serious

reported accident distributions we see that they include females who certainly destroyed belongings and females whose mothers often had violent rages.

We would expect some similarities when comparing the five highest non-superficial reported accident means across gender and these sub-populations include children whose mother's were less than 20 years old in April 1970 and children who certainly destroyed belongings.

Several reported accident trends were identified which were similar to the earlier reported accident trends identified using either the full reported accident distribution or the serious reported accident distribution. For example, the age of the parents produced the same accident frequency associations as did the parents' smoking habits.

When considering the child's neighbourhood we found that for males the earlier results were repeated. The results obtained from the females differed slightly from earlier results in that females from an average neighbourhood had the highest non-superficial reported accident mean, followed by poor, well-to-do and rural neighbourhoods. In fact was actually very little difference between the reported non-superficial accident means of the females from the poor and from the average neighbourhoods.

When examining characteristics of the child's personality as assessed by the child's mother, we found that higher reported non-superficial accident means were present in children who displayed particular undesirable

personality characteristics.

It was found that, with one exception, variables relating to the mother's malaise condition were associated with the same accident frequency trends seen in the full reported accident population. Recall from the earlier analysis that the malaise conditions of the child's mother which, when present, were associated with high reported accident means for both males and females were as follows: miserable or depressed moods, often had violent rages, generally being easily upset, being keyed-up and jittery, nervous, having suffered at one time from a nervous breakdown, and other problems not specified in the questionnaire. We found that the non-superficial reported accident mean was slightly lower for children whose mothers were keyed-up and jittery than for children whose mothers were not keyed-up. This anomaly has already been discussed in Chapter 3, Section 6, with respect to serious reported accidents.

The goodness-of-fit results obtained from fitting the Poisson and Negative Binomial distributions to the male and female sub-populations are summarised in Table 29. The Poisson model gave an acceptable fit at the one percent significance level for 48 (55%) of 87 male sub-populations and for 48 (58%) of the 83 female sub-populations. The Negative Binomial model gave an acceptable fit at the one percent level to 70 (96%) of the 73 male sub-populations and to 58 (94%) of the 62 female sub-populations. Clearly the Negative Binomial model gives very favourable results to the non-superficial reported accident distribution.

TABLE 29. Goodness-of-fit results from fitting the Poisson and Negative Binomial distributions to the non-superficial reported accident distributions of the male and female sub-populations.

		Level of significance for the Chi-squared values obtained from the model fits				
Sex	Distrib.	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
MALE	POISSON	39	6	6	36	87
	N.B.	3	10	13	47	73
FEMALE	POISSON	35	7	8	33	83
	N.B.	4	9	7	42	62

Examining the reported and predicted accident counts in the poor fitting Poisson models, we see at which counts the model performed badly. For the males we observed that the zero accident count and the tail of the reported distribution were usually underestimated by the Poisson model whereas the one and two accident counts were usually overestimated. In contrast, the predicted two accident count for the female models was usually very close to the reported two accident count. In addition, for females the zero accident count and the tail of the reported distribution was usually underestimated by the Poisson model, whereas the one accident count was usually overestimated. For the Poisson model this suggests a mixture of accident misreporting which cannot be identified without further investigation.

We now examine the parameter estimates for P and N obtained from fitting the Negative Binomial distribution to the non-superficial reported accident distributions of

the male and female sub-populations.

It was observed that the parameter estimates of P were generally larger in the extreme values for the males than for the females and there was strong evidence of a relationship between the parameter estimates for P and N for both males and females. Further, for both sexes the larger estimates of P corresponded to smaller estimated parameters of N , and smaller estimates for P corresponded to larger estimates of N . This result had already been observed when modelling the full and the serious reported accident distribution with the Negative Binomial distribution (see Section 3.4).

Since the Negative Binomial distribution provides a good model for the non-superficial reported accident distribution it seems sensible to fit the traditional accident modelling distributions to the superficial/unidentified reported accident distribution. We now examine the fitting of the traditional accident probability distributions to the superficial reported accident distribution

5.4 MODELLING THE SUPERFICIAL REPORTED ACCIDENT DISTRIBUTION USING TRADITIONAL MODELS

The male and female populations were again split into sub-populations as defined in Section 3.2, and the superficial/unidentified reported accident distribution for each sub-population was modelled with the simple

Poisson distribution and the Negative Binomial distribution. We shall first describe the results from the social index split. Secondly, we give an overview of the results for all the splits.

5.4.1 SUMMARY OF OVERALL RESULTS

For males we found that after splitting by social index the lowest reported superficial/unidentified accident mean was seen in the most disadvantaged social index sub-population and the highest reported superficial/unidentified accident mean was seen in the average social index sub-population. This is to be expected since we suspect that many male superficial/unidentified reported accidents are misreported.

The Poisson model was found to be acceptable at the one percent significance level for the most advantaged, advantaged and the most disadvantaged sub-populations, yet at the ten percent level was acceptable for the most disadvantaged sub-population only.

The Negative Binomial model was acceptable at the ten percent significance level for the advantaged, average and disadvantaged sub-populations. The Negative Binomial was not fitted to the most advantaged or the most disadvantaged sub-populations because the tail of each reported superficial accident distribution was too short to be modelled.

For females we found that the lowest reported superficial accident mean was seen in the most advantaged social index sub-population and the highest reported superficial accident mean was seen in the most disadvantaged social index sub-population. The Poisson model was acceptable at the ten percent level for the most advantaged and advantaged sub-populations, and at the one percent significance level we also accept the average and the disadvantaged sub-populations.

The Negative Binomial model was only fitted to the average social index sub-population and was acceptable at the ten percent level. The Negative Binomial was not fitted to the other sub-populations because the tail of each reported superficial accident distribution was too short to be modelled.

Examining the results over all the splits we again observed that the reported accident means for the male sub-populations were higher than for the corresponding female sub-populations and Table 30 gives a summary of the reported accident means. Comparing the five highest male accident mean sub-populations for the superficial and the full reported accident distributions we found that they included males whose father's were less than 20 years old in April 1970 and males whose mothers often had violent rages. Comparing the five highest male accident mean sub-populations for the superficial and the non-superficial reported accident distributions we found that there were none in common.

TABLE 30. Superficial reported accident means of the male and female sub-populations.

SAMPLE MEAN RANGE	Frequency	
	MALE	FEMALE
0.14 to < 0.16	0	1
0.16 to < 0.18	0	11
0.18 to < 0.20	1	33
0.20 to < 0.22	4	12
0.22 to < 0.24	7	8
0.24 to < 0.26	32	3
0.26 to < 0.28	17	4
0.28 to < 0.30	10	1
0.30 to < 0.32	5	0
0.32 to < 0.34	1	0
>= 0.34	3	0
TOTAL	80	73

In a comparison of the five highest female accident mean sub-populations for the superficial and the full reported accident distributions we found that included females with more than three household moves and females whose mother's often had violent rages. Comparing the five highest female accident mean sub-populations for the superficial and the non-superficial reported accident distributions we found that they have in common only those females whose mothers often had violent rages.

In terms of superficial reported accidents the only high risk sub-population in common between gender was children whose mother's often had violent rages.

Several superficial reported accident frequency trends were identified which were similar to the earlier reported accident trends. For example, for both males and females it was seen that the superficial reported accident mean decreased as mother's age increased. The

same accident frequency trend was also seen for in father's age for both sexes.

When considering the child's neighbourhood we found that the earlier results were not repeated. For the males we found that males from an average neighbourhood had the highest superficial reported accident mean, then poor and well-to-do neighbourhoods which had very similar reported accident means, and then the rural neighbourhood. The results from the females also differed slightly in that females from a poor neighbourhood had the highest superficial reported accident mean, then average, well-to-do and rural neighbourhoods.

When examining characteristics of the child's personality as assessed by the child's mother we found that higher reported superficial accident means were present in children who displayed particular undesirable personality characteristics.

It was also found that variables relating to the mother's malaise condition produced the same accident frequency trends as found earlier when modelling the full reported accident distribution. The malaise conditions of the child's mother which when present were associated with high reported accident means for both males and females were as follows; miserable or depressed moods, often had violent rages, generally being easily upset, being keyed up and jittery, nervous, having suffered at one time from a nervous breakdown, and other problems not specified in the questionnaire.

The goodness-of-fit results obtained from fitting the

Poisson and Negative Binomial distributions to the male and female sub-populations are given in Table 31. The Poisson model gave an acceptable fit at the one percent significance level for 36 (45%) of 80 male sub-populations and for 47 (64%) of 73 female sub-populations. The Negative Binomial model gave an acceptable fit at the one percent level to all of the 50 male and 33 female sub-populations. Again we see that the Negative Binomial model produced excellent goodness-of-fit results.

TABLE 31. Goodness-of-fit results from fitting the Poisson and Negative Binomial distributions to the superficial reported accident distributions of the male and female sub-populations.

		Level of significance for the Chi-squared values obtained from the model fits				
Sex	Distrib.	< 1%	1% - 5%	5% -10%	> 10%	TOTAL
MALE	POISSON	44	8	9	19	80
	N.B.	0	0	1	49	50
FEMALE	POISSON	26	17	5	25	73
	N.B.	0	0	2	31	33

When examining the reported and predicted accident counts across gender in the poor fitting Poisson models we found that for both the male and female sub-populations the zero count and the tail of the distribution were usually underestimated by the Poisson model whereas the one accident count was usually overestimated. This suggests a mixture of accident misreporting which cannot be identified without further investigation.

When examining the parameter estimates for P and N

obtained from fitting the Negative Binomial distribution we observed the relationship between the parameter estimates already described in Section 3.3.

5.5 DISCUSSION

From the results of this Chapter we have strong evidence to suggest that the reported accident population of our data set is made up two different reported accident distributions. This provides support for the fitting of Model VI (see Section 3.8), which allows for accident misreporting from the two accident count to the one accident count. In addition, it is suspected that some variables will affect each of the reported accident distributions in a different manner and this phenomenon can be observed when examining the male reported accident frequency trends associated with whether the mother is keyed-up and jittery.

There is strong evidence to suggest that the non-superficial reported accident distribution can be adequately described by the two parameter Negative Binomial model. However, when including superficial and unidentified reported accidents into the reported accident population the Negative Binomial model fails.

By definition, we are looking at accidents which resulted in injury requiring medical advice or treatment, yet the reported accident population still includes superficial and unidentified injuries. The presence of

superficial accidents in the data provides evidence of different threshold levels used by the parents when deciding whether or not an injury require medical adide ore treament. Since the superficial reported accident distribution contains fewer reported accidents than the non-superficial reported accident distribution and is also thought to be less accurately reported, it seems likely that accident misreporting did occur in the lower accident counts. This is confirmed by examining Tables 32 and 33 which describe the full, non-superficial and superfical reported accident distributions for the male and female sub-populations respectively.

TABLE 32. The full, non-superficial and superficial/unidentified reported accident distributions for the male population.

COUNT	ACCIDENT TYPE		
	ALL	NON SUPERFICIAL	SUPERFICIAL/ UNIDENTIFIED
0	3,252	4,134	4,971
1	2,112	1,700	1,113
2	658	373	189
3	218	84	30
4	53	19	4
5	12		0
6	2		2
7	1		1
8	1		
9	1		
TOTAL	4,390	2,774	1,616

From both Table 32 and Table 33 we see that accident misreporting is feasible in the lower counts simply because of the shape of the reported accident

distributions.

TABLE 33. The full, non-superficial and superficial/unidentified reported accident distributions for the female population.

COUNT	ACCIDENT TYPE		
	ALL	NON SUPERFICIAL	SUPERFICIAL/ UNIDENTIFIED
0	3,535	4,272	4,836
1	1,743	1,285	867
2	408	220	113
3	113	42	12
4	20	9	2
5	9	1	
6	1	0	
7	1	1	
TOTAL	3,036	1,899	1,137

It should be noted that there may be significant correlation between non-superficial and superficial/unidentified reported accidents but any correlation would be difficult to interpret and might be misleading since we have strong evidence of accident misreporting. It is therefore thought that modelling the full reported accident distribution with a bivariate Negative Binomial distribution to take into account the two reported accident distributions, namely the non-superficial and superficial/unidentified reported accidents, is not feasible.

CHAPTER 6

ANALYSIS OF THE BRITISH BIRTH DATA SET, PART IV; REGRESSION ANALYSES

6.1 INTRODUCTION

In this Chapter we examine the results of the regression analysis which was described in detail in Section 2.6. The regression analysis was completed in two parts. The software package GLIM was used for the first part of the analysis and these results are presented in Section 6.1. The statistical package SAS was used for the logistic modelling and these results are presented in Section 6.2.

A comparison of the two types of regression models is presented in Section 6.3, and the results show that different regression techniques produce different regression models.

The earlier analysis has shown that some types of accidents are likely to be reported more accurately than others. To continue this line of investigation several different reported accident outcome variables are used in both types of regression modelling and these are defined as follows;

- (i) = Head fractures together with lacerations of the head requiring stitches.
- (ii) = (i) + other bone fractures/dislocations.
- (iii) = (ii) + other lacerations requiring stitches.
- (iv) = (iii) + ingestions, burns, scalds and foreign body type accidents.
- (v) = (iv) + other lacerations requiring stitches.
- (vi) = (v) + sprains, i.e. non-superficial accidents.
- (vii) = Superficial and unidentified injuries.
- (viii) = (vi) + (vii), i.e. all reported accidents.

From the earlier distribution modelling (see Chapter 5, Section 3) we found that the full reported accident distribution could be broken down into two separate reported accident distributions (the non-superficial and the superficial/unidentified reported accident distributions), each of which could be modelled with a separate Negative Binomial distribution. Hence, the results of modelling the non-superficial, the superficial/unidentified and the full reported accident distributions are discussed in detail.

The results from the regression analysis support the work presented in the earlier Chapters and provide further evidence that the full accident population comprises of smaller accident sub-populations.

We now discuss the GLIM regression analysis.

6.2 GLIM REGRESSION ANALYSIS

In this section we model the reported accident data using GLIM modelling techniques. One of the drawbacks of using GLIM is that the size of the data set must not exceed 200,000 data points, i.e. the number of observations multiplied by the number of variables must not exceed 200,000 data points. The data set available to us contains 4,840,815 data points which comprises of 13,135 children and 369 variables. This was far too large to analyse simply so as a convenient result of the GLIM data point restriction the data set was split by gender into two smaller populations of males and females and each was analysed separately. This was quite acceptable since the earlier analysis revealed a natural division between gender. This division, together with the restriction of using only those children of U.K. ethnic origin left us with a male population of 6,310 and a female population of 5,830.

The variables available for use in the analysis were not all used and only 30 variables were selected for the GLIM regression analysis. The variables were selected on the basis of the previous results, curiosity and intuition. Taking into account missing values the male population was reduced to 5,152, (82% of the 6,310 male population), and the female population to 4,759, (82% of the 5,830 female population).

The variables used in the GLIM regression analysis are

as follows.

- 1) The child's mother's age in April 1970 (a continuous variable).
- 2) The child's father's age in April 1970 (a continuous variable).
- 3) The social class of the child's father based on his occupation, i.e. manual or non-manual.
- 4) The standard of the household equipment in the child's home, i.e. adequate, well equipped or low/very low standard.
- 5) The level of tidiness of the child's home, i.e. average, very/over tidy or untidy/chaotic.
- 6) The type of neighbourhood of the child's home, i.e. average, poor, well-to-do or rural.
- 7) The number of household moves (a continuous variable).
- 8) Whether the child's mother smoked (smoker or non-smoker).
- 9) Whether the child's father smoked (smoker or non-smoker).
- 10) The type of accommodation in which the child lived in i.e. owned outright, being bought, a council property or other.
- 11) The number of younger siblings (a continuous variable).
- 12) The number of older siblings (a continuous variable).
- 13) The person-per-room ratio (a continuous variable).
- 14) The child's social index, i.e. most advantaged, advantaged, average, disadvantaged or most disadvantaged.
- 15) The number of hospital admissions (not accidents) for the child (a continuous variable).
- 16) The mother's malaise score (on a continuous scale) which was obtained from a series of questions answered by the mother relating to the mother's health.

The following variables refer to characteristics of the child's personality and are assessed by the child's mother. There are three responses as follows; not applicable, applied somewhat and certainly applied.

- 17) Whether the child was restless.
- 18) Whether the child squirmed or fidgeted.
- 19) Whether the child destroyed belongings.
- 20) Whether the child fought with other children.
- 21) Whether the child was not much liked.
- 22) Whether the child was irritable.
- 23) Whether the child was miserable.
- 24) Whether the child took things.
- 25) Whether the child bit his/her nails.

- 26) Whether the child was disobedient.
- 27) Whether the child could not settle.
- 28) Whether the child told lies.
- 29) Whether the child bullied others.
- 30) The Healy Goldstein score which is a measure of the child's behaviour, character and personality and is given in a continuous scale.

During the analysis it was evident when adding variables into the regression models that mother's age and father's age were closely associated, hence it was difficult to choose between them. As a result three additional variables were created from parents' age and used in the modelling. The first variable was the mean of the parents' age, the second was the age of the youngest parent, and the third was the age of the eldest parent.

In the GLIM analysis we have to specify an error structure in the model fitting and in our analysis we select the Poisson error structure. This is traditional when modelling accident data. An alternative to the Poisson error structure is the Negative Binomial error. However, the Poisson error was used primarily for ease and speed of calculation since the extra parameter needed for the Negative Binomial error has to be recalculated for each fit. Once the GLIM modelling using the Poisson error was completed for the 'all reported accidents' models the Poisson error structure was replaced with a Negative Binomial error and the parameter estimates for the models were recalculated.

It has already been shown that the parameter estimates obtained from using the Poisson error are not significantly different to the parameter estimates

obtained from using the Negative Binomial error in the same model (Maycock and Hall, 1984). The parameter estimates are practically identical but the parameter estimate errors are smaller when using the Poisson error. The use of a Negative Binomial error however, does decrease the model difference as one would expect.

The Poisson error is one of the standard errors available in the GLIM package but not unfortunately the Negative Binomial error. Maycock and Hall (1984) have already explored the use of the Negative Binomial error in their study on roundabout accidents and we take advantage of this in our analysis. We use the GLIM macros provided by Maycock and Hall (1984) to fit models with the Negative Binomial error.

We now examine the results of the GLIM regression models fitted to the male population. We begin by examining the models obtained from using the different reported accident outcome variables and then discuss the modelling of all reported accidents, non-superficial reported accidents and then the superficial/unidentified reported accidents.

6.2.1 GLIM MODELLING OF DIFFERENT REPORTED ACCIDENT OUTCOMES

We recall that the different dependent accident outcome variables were defined in Section 6.1. Table 34 shows the GLIM Poisson regression models obtained from

modelling the different reported accident outcome variables for the male population.

TABLE 34. Details of the GLIM regression models for the male population using a Poisson error structure and a 0.1 percent acceptance criteria for inclusion in the model. The variables in brackets are the variables that would be included in the models if the 0.1 percent criteria was extended to the one percent level.

OUTCOME	Variables included in the regression model
(i)	Youngest parent's age (Neighbourhood)
(ii)	Whether male bullied others (Youngest parent's age)
(iii)	Healy Goldstein score (Youngest parent's age) (Whether male bullied others) (Neighbourhood)
(iv)	Youngest parent's age Whether male bullied others Mother's malaise score
(v)	Mother's malaise score Youngest parent's age Whether male bullied others (Healy Goldstein score) (Neighbourhood)
(vi)	Mother's malaise score Youngest parent's age Whether male bullied others (Neighbourhood) (Healy Goldstein score)
(vii)	Mother's malaise score Eldest parent's age
(viii)	Mother's malaise score Mean parents' age Neighbourhood Whether male was restless (Whether male bullied others)

From the table we can clearly see the differences in

the models of the different reported accident outcome variables and note that no single variable appears consistently in all the male GLIM regression models.

Table 35 shows the GLIM Poisson regression models obtained from modelling the different reported accident outcome variables for the female population and again we note that no single variable appears consistently in all the female GLIM regression models.

TABLE 35. Details of the GLIM regression models for the female population using a Poisson error structure and a 0.1 percent acceptance criteria for inclusion in the model. The variables in brackets are the variables that would be included in the models if the 0.1 percent criteria was extended to the one percent level.

OUTCOME	Variables included in the regression model
(i)	Whether female could settle
(ii)	Whether female could settle (Mean parents' age) (Neighbourhood)
(iii)	Whether female could settle (Neighbourhood)
(iv)	Mother's malaise score Youngest parent's age (Whether female destroyed belongings)
(v)	Mother's malaise score Youngest parent's age Whether female destroyed belongings
(vi)	Mother's malaise score Youngest parents' age Whether female destroyed belongings
(vii)	Mother's malaise score Mother's age
(viii)	Mother's malaise score Mother's age (Whether female fought others) (Whether female destroyed belongings)

The results of this 'cumulative' reported accident modelling supports our hypothesis that the full reported accident distribution actually comprises of smaller accident distributions. Since we also suspect that the superficial/unidentified reported accident distribution would be the most misreported of the accident distributions we shall now discuss in more detail the modelling of the full reported accident distribution, the non-superficial reported accident distribution and the superficial/unidentified reported accident distribution. We first discuss the results for the males and then the results for the females.

6.2.2 GLIM MODELLING OF ALL REPORTED ACCIDENTS FOR THE MALES

Using all reported accidents as the dependent variable the general mean was fitted to the male population and the first regressor variable was added into the model based on the most significant improvement in fit. Mother's malais score clearly gave the most significant improvement in fit with a Chi-squared value of 55.52 (1 d.f.). After including mother's malaise score in the model we then had some difficulty in selecting the next variable to add into the model. The variables which gave the most significant improvement in fit according to the acceptance criterion were as follows, where the value of the Chi-squared statistic (on 1 d.f.) is shown in brackets:

- i) parents' mean age (43.90),
- ii) youngest parent's age (43.84),
- iii) father's age (38.62),
- iv) mother's age (38.38),
- v) eldest parent's age (36.72).

Clearly the choice between parents' mean age and youngest parent's age is difficult. Parents' mean age was selected for addition into the model because it produced the most significant improvement in fit and was also representative of both parents' age. Once parents' mean age was included the effect of the other age-related variables decreased greatly. Clearly some confounding is present here since mother's and father's age are closely associated.

After including parents' mean age it was found that a further significant improvement in fit could be obtained with the addition of child's neighbourhood into the model. This gave an improvement in fit measured by a Chi-squared value of 23.6 with 3 d.f.

The final regressor variable to be added into the model was the variable relating to whether the male was restless. This gave a further improvement in fit measured by a Chi-squared value of 16.73 (2 d.f.) and the details for this Poisson regression model are given in Table 36. No further significant improvement in fit was possible using the 0.1 % criterion but if we extend this to the one percent level we can improve the model further with the addition of the variable relating to whether the child bullied others.

TABLE 36. Details of the 'all reported accidents' regression model for the male population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-1.6240	0.1096	
1) Mother's malaise score	0.0270	0.0047	32.29
2) Parents' mean age	-0.0193	0.0032	37.68
3) Neighbourhood			
Poor	-0.0484	0.0725	22.85
Well-to-do	-0.0269	0.0422	
Rural	-0.2193	0.0474	
4) Child was restless			
Applied somewhat	0.0660	0.0433	16.74
Certainly applied	0.1780	0.0450	
Scaled Deviance (d.f.)	5788.3	(5144)	

It can be seen from Table 36 that the male regression model for all reported accidents contains four variables. The mean deviance reduction for each variable was calculated and each variable was acceptable at the one percent significance level. Looking at the effects of each variable in turn we begin by examining mother's malaise score.

The mother's malaise score was obtained from a series of mother's health questions answered by the child's mother or mother figure. A mother who scored highly for this variable was considered at high risk of having psychiatric problems. The effect of mother's malaise score in the model suggests that males of mother's with high malaise scores had a higher reported mean accident rate than males whose mothers had lower scores. As the malaise score increases so does the male mean reported accident rate. This is not suprising since reported

accident trends with respect to mother's health were identified in the early analysis. If a mother is unhealthy in some way then the child may have less supervision or attention and hence be at higher risk of having an accident.

It is also possible that a child whose mother is at high risk of psychiatric problems may develop some of his mother's personality characteristics. Alternatively it is also possible that the mother is influenced by the child's nature and that some mothers might have high malaise scores because of the pressures of coping with particular characteristics of their personality.

Looking now at the effect of parents' mean age (in years) in April 1970 we observe that the male reported mean accident rate decreased as parents' mean age increased. Hence, a male of parents with a low mean age would be more at risk of having an accident than a male whose parents had a higher mean age. This type of parental age-related accident trend was seen earlier when examining the reported accident means after splitting the data by mother's or father's age. Parents' age would also be closely linked with other variables such as parents income, neighbourhood, standard of household equipment, etc., so the effects may be confounded.

Examining the neighbourhood variable we note that the standard errors of the parameter estimates suggest that the only type of neighbourhood which had a significant effect on the male mean reported accident rate was the rural neighbourhood. The results indicate that males from

this type of area had a lower mean reported accident rate than males from any other neighbourhood. We note that this is the only environmental variable in the model.

Examining the variable relating to whether the child was restless we recall that variables of this type were obtained from the mother's own assessment of characteristics of her child's personality. The standard errors of the parameter estimates suggest that the mean reported accident rate for males who were certainly restless was higher than the mean reported accident rate for males in the other categories. This is a logical result since restless children would be more active and perhaps get into more high-risk accident situations. This may also provide evidence of accident proneness since proneness is considered to be associated with personal rather than environmental attributes.

6.2.3 GLIM MODELLING OF NON-SUPERFICIAL REPORTED ACCIDENTS FOR THE MALES

We now examine the modelling results obtained after changing the dependent outcome variable to accidents which resulted in non-superficial injuries. After fitting the general mean the first variable to be added into the model was the youngest parent's age with an improvement in fit measured with a Chi-squared value of 29.14 (1 d.f.). The same confounding effects with respect to the age-related variables seen in the previous model were not encountered

to the same extent here and the Chi-squared values (on 1 d.f.) for each of the age-related variables were as follows; parents' mean age, 24.34, youngest parent's age, 29.14, father's age, 20.15, mother's age, 22.80, eldest parent's age, 17.03.

After including the youngest parent's age it was found that a further significant improvement in fit could be obtained from the addition of mother's malaise score into the model. This gave an improvement in fit measured with a Chi-squared value of 24.62 (1 d.f.).

The next regressor variable added into the model was the variable relating to whether the child bullied others. This gave a further improvement in fit measured by a Chi-squared value of 16.81 with 2 d.f. and the details for this Poisson model are given in Table 37. No further significant improvement in fit was obtained using the 0.1% criterion but the model could be improved further by extending this to the one percent significance level. This would allow us to include the child's neighbourhood in the model, and after including child's neighbourhood a further improvement could be obtained with the addition of the Healy Goldstein score. This is a continuous variable ranging from 0 to 58 and measures the child's behaviour, character and personality.

It can be seen from Table 37 that the male regression model for non-superficial accidents contains three variables. The mean deviance reduction for each variable was calculated and each variable was acceptable at the one percent significance level. The effects of mother's

malaise score and youngest parent's age are not very different to the effects of mother's malaise score and parents' mean age discussed in the previous model, in fact the respective parameter estimates are very similar.

TABLE 37. Details of the non-superficial accident regression model for the male population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-2.0860	0.1137	
1) Youngest parent's age	-0.0202	0.0042	20.80
2) Mother's malaise score	0.0268	0.0058	18.25
3) Child bullied others			
Applied somewhat	0.2192	0.0551	7.48
Certainly applied	-0.1778	0.1823	
Scaled Deviance (d.f.)	5079.7	(5147)	

A variable not included in the previous model is the variable relating to whether the child bullied others. Examining this variable we note that this personality characteristic only has a significant effect in the model if it applied somewhat to the male. Males in this category had a higher mean accident rate than males in the other bullying categories. This is a logical result since males who bullied others would get into more situations from which injury could result and this was also discussed in the early analysis. This may also provide evidence of accident proneness.

6.2.4 GLIM MODELLING OF SUPERFICIAL REPORTED ACCIDENTS FOR THE MALES

We now examine the modelling results obtained when using accidents which resulted in superficial or unidentified injuries as the dependent outcome variable. After fitting the general mean the mother's malaise score was the first variable to be added into the model with an improvement in fit measured with a Chi-squared value of 28.85 (1 d.f.). The same difficulty with respect to the age-related variables seen in the first regression model is again encountered when trying to select the next variable for the model. The improvement in fit obtained for each of the age-related variables (with 1 d.f.) was as follows with the Chi-squared values in brackets;

- i) parents' mean age (23.00),
- ii) youngest parent's age (17.96),
- iii) father's age (21.86),
- iv) mother's age (18.36),
- v) eldest parent's age (23.48).

Eldest parent's age was added into the model on the grounds that it gave a very slightly more significant improvement in fit but there is little difference between choosing the eldest parent's age or the parents' mean age. After including this variable it was found that no further significant improvement in fit was possible at either the 0.1% or one percent criterion. The details for this Poisson model are given in Table 38.

TABLE 38. Details of the superficial accident regression model for the male population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-2.5380	0.1389	
1) Mother's malaise score	0.0391	0.0073	28.26
2) Eldest parent's age	-0.0220	0.0046	29.19
Scaled Deviance (d.f.)	4141.9	(5149)	

The effects of mother's malaise score in this model is the same as the effect of mother's malaise score in the previous two models. Similarly, the effect of eldest parent's age is the same as the effect of the age-related variables seen in the previous two models.

Making a comparison of the parameter estimates of these three male models only we see that mother's malaise score and an age-related variable are included in each of the models. No further variables were included in the final model for superficial accidents but for non-superficial accidents the model also included the variable relating to whether the child bullied others. The 'all accident' regression model included the following variables; mother's malaise score, a parental age-related variable, the male's neighbourhood and the variable relating to whether the male was restless.

We also note that by splitting the 'all reported accidents' dependent variable into non-superficial and superficial reported accidents and modelling them separately we keep mother's malaise score in the respective models and also a parental-age related

variable. In addition, we remove 'neighbourhood' and 'whether the child was restless' from the models and in the non-superficial reported accident model, include the variable relating to 'whether the child bullied others'.

The effects of mother's malaise score and the parental age-related variable in each model were generally the same across the three models. As mother's malaise score increased so the reported mean accident rate increased, and as the age-related variable increased, so the mean reported accident rate decreased. The presence of variables relating to particular characteristics of the male's personality in the GLIM regression models, .i.e. male bullied others, restless, might be considered as evidence of accident proneness.

6.2.5 GLIM MODELLING OF ALL REPORTED ACCIDENTS FOR THE FEMALES

We now examine the GLIM regression results from modelling the female population using a Poisson error structure. Using all accidents as the dependent outcome variable the general mean was fitted to the female population and the next regressor variable was added. Mother's malais score clearly gave the most significant improvement in fit with a Chi-squared value of 55.94 (1 d.f.) so was the first variable added into the model.

After including mother's malaise score we again get into difficulty when selecting the next variable to

include in the model. The variables which gave the most significant improvement in fit (with 1 d.f.) are as follows with the Chi-squared values in brackets; parents' mean age (31.08), youngest parent's age (34.35), father's age (21.97), mother's age (34.51), eldest parent's age (23.60).

It was decided to include mother's age in the model since it gave the most significant improvement in fit, and when it was included in the model the effect of the other age-related variables disappeared, in fact no further significant improvement of fit was possible at the 0.1% criterion level. The details for this Poisson model are shown in Table 39. By extending the acceptance criterion to the one percent significance level a further significant improvement of fit could be obtained by adding the variable relating to whether the child fought others. The model could be improved even further with the addition of the variable that related to whether the child destroyed belongings.

TABLE 39. Details of the 'all reported accidents' regression model for the female population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-1.8600	0.0961	
1) Mother's malaise score	0.0379	0.0047	47.30
2) Mother's age	-0.0227	0.0028	32.60
Scaled Deviance (d.f.)	5034.3 (4756)		

It can be seen from Table 39 that the regression model

for all accidents occurring in the female population contains two variables. The mean deviance reduction for each variable was calculated and each variable was acceptable at the one percent significance level. Looking at each variable in turn we begin with mother's malaise score.

The effect of the child's mother malaise score suggests that females of mothers with high malaise scores had a higher reported mean accident rate than females whose mothers had lower scores. As the malaise score increased so did the mean reported accident rate of the females. This variable also appeared in the regression modelling of the male population with very similar effects.

Now looking at mother's age we observe that the female reported mean accident rate decreased as the mother's age increased. Hence a female of a young mother was more at risk of having an accident than a female with an older mother. The effect of mother's age in this model is very similar to the effect of the age-related variables seen in the previous GLIM regression models of the male population.

6.2.6 GLIM MODELLING OF NON-SUPERFICIAL REPORTED ACCIDENTS FOR THE FEMALES

We now examine the modelling results obtained after changing the dependent outcome variable to accidents which

resulted in non-superficial injuries.

After fitting the general mean to the population the mother's malaise score was the variable to be added into the model with an improvement measured by a Chi-squared value of 26.73 (1 d.f.)

When selecting the next variable to be included in the model we again encountered difficulty when examining age-related variables. The improvement in fit (on 1 d.f.) for each of the age variables was as follows with the Chi-squared values in brackets; parents' mean age (16.26), Youngest parent's age (17.52), father's age (12.36), mother's age (16.82), eldest parent's age (12.66), and including the youngest parent's age into the model it was found that a further significant improvement in fit could be obtained from the addition of the variable relating to whether the child destroyed belongings. This gave an improvement in fit measured by a Chi-squared value of 15.30 with 2 d.f. The details for this Poisson model are shown in Table 40 and no further significant improvement in fit was possible using the 0.1% or one percent criterion.

It can be seen from Table 40 that the regression model for female non-superficial accidents contains three variables. The mean deviance reduction for each variable as calculated and each variable was acceptable at the one percent significance level. The effects of mother's malaise score and the age-related variable in the one percent significance level. The effects of mother's malaise score and the age-related variable in this model are the same as the respective variables seen in the previous model.

TABLE 40. Details of the non-superficial accident regression model for the female population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-2.3840	0.1352	
1) Mother's malaise score	0.0283	0.0067	18.74
2) Youngest parent's age	-0.0199	0.0050	17.97
3) Child destroyed belongings			
Applied somewhat	0.1969	0.0695	8.22
Certainly applied	0.4539	0.1457	
Scaled Deviance (d.f.)	4304.2	(4754)	

The new variable not included in the previous model is the variable relating to whether the child destroyed belongings. The effect of this variable is such that females who certainly destroyed belongings had a higher reported accident mean than those who sometimes destroyed belongings, who in turn had a higher reported accident mean than those who never destroyed belongings. This trend was also discussed in the early analysis and it is thought that children who destroyed belongings were in a higher accident risk group because they were involved in a greater number of accident-risk situations. Again we could consider that the inclusion of this type of variable in the model is evidence of accident proneness.

6.2.7 GLIM MODELLING OF SUPERFICIAL REPORTED ACCIDENTS FOR THE FEMALES

We now examine the modelling results obtained when looking at accidents which resulted in superficial or unidentified injuries. After fitting the general mean, mother's malaise score was added into the model with an improvement in fit measured by a Chi-squared value of 30.82 (1 d.f.). The same difficulty with respect to the age-related variables seen in the first regression model are again seen when selecting the next variable to include in the model and the improvement in fit (on 1 d.f.) for each of the age-related variables are as follows;

- i) parents' mean age (15.30),
- ii) youngest parent's age (17.51),
- iii) father's age (9.76),
- iv) mother's age (18.66),
- v) eldest parent's age (17.51).

There is little difference between choosing mother's age, the youngest parent's age or the eldest parent's age. However, mother's age was added into the model on the grounds that it gave a very slightly more significant improvement in fit and after including it in the model it was found that no further significant improvement in fit could be obtained at either the 0.1% or one percent level. The details for this Poisson model are given in Table 41 where it can be seen that the mean deviance reduction for each variable was calculated and was acceptable at the one percent significance level.

TABLE 41. Details of the superficial accident regression model for the female population using a Poisson error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-2.7790	0.1770	
1) Mother's malaise score	0.0460	0.0085	40.37
2) Mother's age	-0.0278	0.0066	27.34
Scaled Deiance (d.f.)	3245.6	(4756)	

The effect of mother's malaise and mother's age are very similar to the effects of the respective variables in the female GLIM regression model of the full reported accident distribution.

Making a general comparison of the parameter estimates of these three female GLIM regression models we see that mother's malaise score and an age-related variable were included in each of the models. No further variables were included in the superficial or 'all acccident' regression models but in the non-superficial accidents model we also included the variable relating to whether the child destroyed belongings. Once again the effects of mother's malaise score was the same across models, and similarly for the parental age-related variable.

We also note that by splitting 'all reported accidents' into non-superficial and superficial reported accidents we keep a parental age-related variable and mother's malaise score in each model and also add in 'whether the child destroyed belongings' into the non-superficial reported accident model.

6.2.8 GLIM MODELLING OF ALL REPORTED ACCIDENTS USING A
NEGATIVE BINOMIAL ERROR STRUCTURE

Using all reported accidents as the outcome variable we now replace the Poisson error with a Negative Binomial error and recalculate the parameter estimates for the appropriate final male and female models.

Table 42 gives the parameter estimates and the standard errors for the male regression model using the full reported accident distribution when the Poisson error is replaced by the Negative Binomial error.

TABLE 42. Details of the regression model for the male population using the full reported accident distribution and a Negative Binomial error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-1.6270	0.1001	
1) Mother's malaise score	0.0272	0.0049	28.43
2) Parents' mean age	-0.0192	0.0033	28.15
3) Neighbourhood			
Poor	-0.0469	0.0759	6.57
Well-to-do	-0.0270	0.0441	
Rural	-0.2190	0.0493	
4) Child was restless			
Applied somewhat	0.0665	0.0451	7.61
Certainly applied	0.1788	0.0469	
Scaled deviance (d.f.)	5365.7 (5144)		

As expected, there is little difference between the parameter estimates from the Negative Binomial error model and the Poisson error model. The standard errors are very slightly larger for the model with the Negative Binomial error and the scaled deviance is now smaller. The mean

deviance reduction was calculated for each variable and each variable was acceptable at the one percent significance level.

We now look at the female regression model using the full reported accident distribution and the Negative Binomial error and examine Table 43. Again, we observe similarities with the Poisson error model as expected. The mean deviance reduction was calculated for each variable and each was acceptable at the one percent significance level.

TABLE 43. Details of the regression model for the female population using the full reported accident distribution and a Negative Binomial error structure.

Parameter	estimate	s.e.	m.d.r.
General mean	-1.8620	0.1109	
1) Mother's malaise score	0.0377	0.0055	47.30
2) Mother's age	-0.0226	0.0041	32.61
Scaled deviance (d.f.)	4629 (4756)		

Summarising this section of the GLIM regression analysis we compare the Poisson error regression models across gender and see that different GLIM regression models are obtained from the use of different outcome variables. However, we are particularly interested in the full, non-superficial and superficial/unidentified reported accident distributions and note that regardless of gender or which of these three particular outcome variables was used, the final regression model always included mother's malaise score and a parental age-related

variable.

Looking at the models for 'all accidents' we see that the model for the male population includes mother's malaise score, parents' mean age, child's neighbourhood and whether the male was restless. For females the model contains mother's malaise score and mother's age.

The final model for the male population when using non-superficial accidents as the dependent variable contains mother's malaise score, the age of the youngest parent and whether the child bullied others. For females the model includes mother's malaise score, the age of the youngest parent and whether the child destroyed belongings.

Using superficial accidents as the dependent variable the male population model contains mother's malaise score and the age of the eldest parent, whereas the model for the females contains mother's malaise score and mother's age.

The effects of mother's malaise score and the age-related variables have already been discussed and it was found that the general effects were same in each model. As mother's malaise score increases, the reported mean accident rate increases and as the parental age-related variable increases so the mean reported accident rate decreases. There was some evidence of accident proneness since variables relating to characteristics of the child's personality were included in three of the final models.

We now examine the results of the logistic regression modelling.

6.3 LOGISTIC REGRESSION ANALYSIS

In the second part of the regression analysis the accident data is modelled using logistic regression models. Hence, no assumptions are made about the underlying reported accident distribution.

The variables available for the analysis were not all used and only 15 variables were selected for the logistic regression analysis. These variables were selected from the results of the GLIM regression analysis and taking into account missing data the male population was reduced to 5,452, (86% of the 6,310 male population), and the female population to 5,026 (86% of the 5,830 female population). The male and female populations were modelled separately and the variables used in this part of the analysis are as follows;

- 1) The child's mother's age in April 1970 (a continuous variable).
- 2) The child's father's age in April 1970 (a continuous variable).
- 3) The standard of the household equipment in the child's home, i.e. adequate, well equipped or low/very low standard.
- 4) The tidiness of the child's home i.e. average, very/over tidy or untidy/chaotic.
- 5) The type of neighbourhood of the child's home, i.e. average, poor, well-to-do or rural.
- 6) The number of household moves(a continuous variable).
- 7) The child's social index, i.e. most advantaged, advantaged, average, disadvantaged or most disadvantaged.
- 8) The mother's malaise score (on a continuous scale) relating to the mother's health which was obtained from a series of questions answered by the mother.

The following variables relate to characteristics of the child's personality and are assessed by the child's

mother or mother-figure. There are three responses as follows; not applicable, applied somewhat and certainly applied.

- 9) Whether the child was restless.
- 10) Whether the child squirmed or fidgeted.
- 11) Whether the child destroyed belongings.
- 12) Whether the child fought with other children.
- 13) Whether the child was solitary.
- 14) Whether the child could not settle.
- 15) Whether the child bullied others.

During the GLIM analysis it was evident that mother's age and father's age were closely associated and it was difficult to choose between them when adding variables into the regression models. As a result three additional variables were created from parents' age and were used in the modelling. The first variable was the mean of the parents' age, the second was the age of the youngest parent, and the third was the age of the eldest parent. These three variables were used in the GLIM regression modelling and are also used in the logistic regression modelling.

As in the GLIM regression analysis several different accident outcome variables have used and these are defined in Chapter 6, Section 1.

The logistic regression also requires the definition of a binary grouped response variable. When modelling the full reported accident distribution four different binary response groupings were used and these are defined as follows.

- 1) [Zero reported accidents] or [one or more reported accidents].
- 2) [Zero or one reported accident] or [two or more reported accidents].
- 3) [Zero, one or two reported accidents] or [three or more reported accidents].
- 4) [Zero reported accidents] or [three or more reported accidents].

When modelling the other reported accident outcome variables the binary response variable was defined as 1) above. It is suspected that the response groupings 3) and 4) may be inadequate when modelling the 'all reported accident' female population because the tail of the female reported accident distribution may be too short.

The best fitting logistic models were selected in the same stepwise selection process used in the stepwise GLIM regression modelling. Variables were added into the logistic regression model based on the most significant improvement in fit at the 0.1% significance level.

In the logistic regression analysis we first look at the results of modelling the 'cumulative' reported accident distribution and then discuss in detail the modelling of the full, the non-superficial and then the superficial/unidentified reported accident distribution for the males and the females. The final models across gender are also compared. We now examine the results of modelling the 'cumulative' reported accident distribution from the male population and then the female population.

6.3.1 LOGISTIC MODELLING OF DIFFERENT REPORTED ACCIDENT OUTCOMES

Table 44 shows the logistic regression models obtained from modelling the different reported accident outcome variables for the male population. From the table we can clearly see the differences in the models of the different reported accident outcome variables and note that no single variable appears consistently in all the male logistic regression models.

Table 45 shows the logistic regression models obtained from modelling the different reported accident outcome variables for the female population and again we note that no single variable appears consistently in all the models.

The results of this 'cumulative' reported accident modelling supports our hypothesis that the full reported accident distribution is comprised of smaller accident distributions.

TABLE 44. Details of the logistic regression models for the male population using a 0.1 percent acceptance criterion for inclusion in the model and binary response grouping(i). The variables in brackets are the variables that would be included in the models if the 0.1 percent criterion was extended to one percent.

OUTCOME	Variables included in the regression model
(i)	(Neighbourhood)
(ii)	Whether male bullied others (Youngest parent's age)
(iii)	(Youngest parent's age) (Whether male destroyed belongings)
(iv)	Youngest parent's age Mother's malaise score (Whether male bullied others)
(v)	Mother's malaise score Youngest parent's age (Whether male squirmed)
(vi)	Mother's malaise score Youngest parent's age (Whether male squirmed)
(vii)	Mother's malaise score Parents mean age
(viii)	Mother's malaise score Youngest parents' age (Whether male was restless)

TABLE 45. Details of the logistic regression models for the female population using a 0.1 percent acceptance criterion for inclusion in the model and binary response grouping(i). The variables in brackets are the variables that would be included in the models if the 0.1 percent criterion was extended to one percent.

OUTCOME	Variables included in the regression model
(i)	(Whether female settled)
(ii)	Whether female settled
(iii)	(Whether female settled) (Whether female fought others)
(iv)	Youngest parent's age Mother's malaise score
(v)	Youngest parents' age Mother's malaise score (Whether female fought others)
(vi)	Mother's malaise score Youngest parent's age (Whether female fought others)
(vii)	Mother's malaise score Parents mean age
(viii)	Mother's malaise score Youngest parents' age

Since we also suspect that the superficial/unidentified reported accident distribution would be the most misreported of the accident distributions we shall now discuss in more detail the modelling of the full reported accident distribution, the non-superfical reported accident distribution and the superficial/unidentified reported accident distribution. We first discuss the results for the males and then the results for the females.

6.3.2 LOGISTIC MODELLING OF ALL REPORTED ACCIDENTS FOR
THE MALES

Using all reported accidents as the dependent outcome variable and specifying (i) as the response grouping we fit the intercept to the male population. The first regressor variable was added to the model based on the most significant improvement in fit and this was mother's malaise score with a Chi-squared value of 31.81 (1 d.f.). After including mother's malaise score in the model the next variable was selected which was the age of the youngest parent. This gave the most significant improvement in fit with a Chi-squared value of 25.51 (1 d.f.). The parameter estimates for this model are given in Table 46 and no further significant improvement in fit was possible using the 0.1% criterion, however extending this to the one percent level we can improve the model further with the addition of the variable relating to whether the child was restless.

TABLE 46. Details of the logistic regression model for the male population using all reported accidents as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	-0.4237	0.1430	
1) Mother's malaise score	-0.0418	0.0078	21.05
2) Youngest parent's age	0.0265	0.0053	18.55
Model log-likelihood (d.f.)	7492 (5449)		

From Table 46 we see that the logistic regression model for the males contains just two variables. The mean deviance reduction for each variable was calculated and each was significant at the one percent level.

Examining each variable in turn we begin with mother's malaise score. We recall that a mother who scored highly for this variable was considered at high risk of having psychiatric problems. The effect of the mother's malaise score suggests that males of mothers with high malaise scores had a greater probability of having an accident than males whose mothers had lower scores. As the mother's malaise score increases so does the probability of having an accident. This is not suprising since reported accident trends with respect to mother's health have already been identified in the early analysis and in the GLIM regression analysis.

Now looking at the effect of youngest parents' age in years in April 1970 we observe that the probability of having an accident decreased as the youngest parent's age increased. Hence, a male with a particularly young parent would be more at risk of having an accident than a male whose youngest parent was older. This type of parental age-related trend was also identified earlier.

We now change the response variable to be (ii) and after fitting the intercept we encountered some difficulty when selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit on 1 d.f. are as follows with the Chi-squared values in brackets;

- i) mother's malaise score (31.55),
- ii) parents' mean age (35.81),
- iii) youngest parent's age (34.27),
- iv) father's age (32.00),
- v) mother's age (30.86),
- vi) eldest parent's age (31.19).

The parents' mean age was added into the model on the grounds that it produced a very slightly more significant improvement in fit. Mother's malaise score was the next variable to be added and gave a significant improvement in fit measured with a Chi-squared value of 28.33 (1 d.f.). The final regressor variable to be added into the model using the 0.1% criterion was the child's neighbourhood. This variable gave an improvement in fit measured by the Chi-squared value of 16.46 (3 d.f.) and the parameter estimates for this model are given in Table 47. By extending the acceptance criterion to the one percent level we can improve the model further with the addition of the variable relating to whether the child was restless.

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effects of mother's malaise score and parents' mean age are very similar to the effects of mother's malaise and youngest parent's age seen in the previous model shown in Table 46. The neighbourhood of the male was not included in the previous model though and when examining the neighbourhood variable we note that the s.e. of the parameter estimates suggest that the only type of neighbourhood which had a significant effect on the model was the poor neighbourhood. The results indicate that a

male from this type of area had a higher probability of having an accident than a male from any other area. The child's neighbourhood also appeared in the GLIM male regression model for 'all accidents'.

TABLE 47. Details of the logistic regression model for the male population using all reported accidents as the outcome variable and {zero or one reported accident} and {two or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	0.9837	0.2127	
1) Mother's malaise score	-0.0525	0.0105	29.51
2) Parents' mean age	0.0400	0.0075	36.61
3) Neighbourhood			
Poor	-0.1722	0.0638	6.70
Well-to-do	-0.0171	0.1236	
Rural	-0.0760	0.0779	
Model log-likelihood (d.f.)	4463 (5446)		

We now change the response variable to be (iii) and after fitting the intercept we again have some difficulty in selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit according to the acceptance criterion are as follows where the value of the Chi-squared statistic is shown in brackets, on 1 d.f.

- i) mother's malaise score (16.90),
- ii) parents' mean age (20.79),
- iii) youngest parent's age (19.99),
- iv) father's age (17.26),
- v) mother's age (19.59),
- vi) eldest parent's age (18.12).

Clearly there is little to choose between these variables and parents' mean age was added into the model

on the grounds that it gave a very slightly more significant improvement in fit. After including parents' mean age in the model the next variable, mother's malaise score was added and gave a significant improvement measured with a Chi-squared value of 28.33 (1 d.f.). The details for this model are shown in Table 48 and no further significant improvement was possible using the 0.1% or one percent criterion.

TABLE 48. Details of the logistic regression model for the male population using all reported accidents as the outcome variable and {zero, one or two reported accidents} and {three or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	1.9247	0.3706	
1) Mother's malaise score	-0.0682	0.0171	42.52
2) Parents' mean age	0.0560	0.0135	53.63
Model log-likelihood (d.f.)	1908 (5449)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effects of mother's malaise score and the parental age-related variable are again very similar to the effects seen in the model described in Table 46.

We now change the response variable to be (iv) and after fitting the intercept we again have some difficulty in selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit (on 1 d.f.) according to the acceptance

criterion are as follows, with the Chi-squared statistic in brackets;

- i) mother's malaise score (24.53),
- ii) parents' mean age (27.54),
- iii) youngest parent's age (27.46),
- iv) father's age (23.34),
- v) mother's age (25.97),
- vi) eldest parent's age (23.62).

On the grounds that it gave a very slightly more significant improvement in fit, the parents' mean age was added into the model. Following this, mother's malaise score was added and gave a significant improvement measured with a Chi-squared value of 23.63 (1 d.f.). The details for this model are shown in Table 49 and no further significant improvement was possible using the 0.1% criterion. However, extending the acceptance level to the one percent significance level we can further improve the model with the addition of the variable relating to whether the male fought others.

TABLE 49. Details of the logistic regression model for the male population using all reported accidents as the outcome variable and {zero reported accidents} and {three or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	1.0847	0.3755	
1) Mother's malaise score	-0.0892	0.0177	79.83
2) Parents' mean age	0.0676	0.0138	77.73
Model log-likelihood (d.f.)	1613 (5449)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent

level. The effects of mother's malaise score and the parental age-related variable are again very similar to the effects seen in the model described in Table 46.

6.3.3 LOGISTIC MODELLING OF NON-SUPERFICIAL REPORTED ACCIDENTS FOR THE MALES

We now change the outcome variable to be 'non-superficial injuries received from the reported accidents' and select the response variable to be (i). After fitting the intercept to the male population we select the first variable to be added into the model and this was mother's malaise score which gave a Chi-squared value of 27.42 (1 d.f.). The next regressor variable to be included in the model was youngest parent's age with a Chi-squared value of 20.59 (1 d.f.). The parameter estimates for this model are given in Table 50 and no further improvement in the model was possible using the 0.1% criterion. Extending the criterion to the one percent level we can further improve the model by including the variable relating to whether the child squirmed. After the addition of this variable we were still able to improve the model further with the addition of the variable relating to whether the male bullied others.

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. Again, the effects of mother's malaise score and the age-related variable are very similar to the effects

seen in the model described in Table 46.

TABLE 50. Details of the logistic regression model for the male population using non-superficial injuries (from the reported accidents) as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	0.1842	0.1512	
1) Mother's malaise score	-0.0402	0.0080	19.53
2) Youngest parent's age	0.0253	0.0056	16.13
Model log-likelihood (d.f.)	6955 (5449)		

6.3.4 LOGISTIC MODELLING OF SUPERFICIAL REPORTED ACCIDENTS FOR THE MALES

We now change the outcome variable to be 'superficial injuries received from the reported accidents' and select the response variable to be (i). After fitting the intercept to the male population we select the first variable to be added into the model and once again we have some difficulty in selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit at the 0.01 percent level (on 1 d.f.) are as follows with the Chi-squared statistic in brackets; mother's malaise score (17.84), parents' mean age (19.22), youngest parent's age (16.36), father's age (17.84), mother's age (15.78), eldest parent's age (18.40).

Parents' mean age was the first variable put into the model. Following this the next regressor variable to be selected was mother's malaise score with a Chi-squared value of 16.05 with 1 d.f. The parameter estimates for this model are given in Table 51 and no further improvement in the model was possible using the 0.1% or the one percent criterion.

TABLE 51. Details of the logistic regression model for the male population using superficial injuries (from the reported accidents) as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	0.7874	0.1787	
1) Mother's malaise score	-0.0371	0.0092	15.71
2) Parents' mean age	0.0259	0.0063	17.06
Model log-likelihood (d.f.)	5567 (5449)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. Again, the effects of mother's malaise score and the age variable are very similar to the effects seen in the model described in Table 46.

Making a comparison of these final six male models we must note that each model included mother's malaise score. Four of the six models included parents' mean age and the remaining two models included the age of the youngest parent. Only one of the models contained more than two variables and that model included mother's malaise score, parents' mean age and the male's neighbourhood. The

effects of mother's malaise score and the parental age-related variables are generally the same across models.

We also note that using different response groupings for the response variable when modelling all reported accidents resulted in slightly different models. All four groupings resulted in models which included mother's malaise score and a parental age-related variable, and the model using binary grouping (ii), i.e. two or more reported accidents, also included the variable 'neighbourhood'.

6.3.5 LOGISTIC MODELLING OF ALL REPORTED ACCIDENTS FOR THE FEMALES

We now examine the results of the logistic regression modelling of the female population. Using all reported accidents as the outcome variable and specifying (i) as the response variable we first fit the intercept to the female population. After fitting the intercept there was some difficulty in selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit (on 1 d.f.) at the 0.01 acceptance level are as follows, with the Chi-squared statistic in brackets; mother's malaise score (28.63), parents' mean age (27.08), youngest parent's age (31.01), father's age (18.07), mother's age (31.46), eldest parent's age (19.67).

Mother's age gave the most significant improvement in

fit and was therefore included in the model, but clearly the other variables make similar contributions to the model. After including mother's age in the model the next variable to be selected was mother's malaise score with the Chi-squared value of 25.24 (1 d.f.). The parameter estimates for this model are given in Table 52 and no further significant improvement in fit was possible using the 0.1% criterion or the one percent criterion.

TABLE 52. Details of the logistic regression model for the female population using all reported accidents as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	-0.1169	0.1516	
1) Mother's malaise score	-0.0405	0.0081	19.06
2) Youngest parent's age	0.0290	0.0055	21.20
Model log-likelihood (d.f.)	6651 (5023)		

The logistic regression model for the females shown in Table 52 contains two variables. The mean deviance reduction for each variable was calculated and each was significant at the one percent level.

Looking at each variable in turn we begin with mother's malaise score which has already been discussed in the previous models. The effect of the mother's malaise score suggests that females of mothers with high malaise scores had a higher probability of accident than females whose mothers had lower scores. As the malaise score increases so does the probability of having an accident.

This is not suprising since reported accident trends with respect to mother's health have already been identified in the analysis.

Examining the effect of the youngest parents' age in years in April 1970 we observe that the probability of having an accident decreased as the youngest parents' age increased. Hence, a female with a particularly young parent would be more likely to have an accident than a female whose youngest parent was older.

We now change the response variable to be (ii) and after fitting the intercept we select the variable which gave the most significant improvement in fit to the model which was mother's malaise score with a Chi-squared value of 35.65 (1 d.f.). After including mother's malaise score in the model, the variables which gave the most significant improvement are as follows with the Chi-squared values in brackets on 1 d.f.

- i) parents' mean age (15.74),
- ii) youngest parent's age (15.29),
- iii) father's age (12.05),
- iv) mother's age (16.03),
- v) eldest parent's age (13.45).

After including mother's age in the model no further improvement in fit was possible at either the 0.1% or one percent level and the parameter estimates for this model are given in Table 53.

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effects of mother's malaise score and the age variable are very similar to the effects seen in the model

described in Table 52.

TABLE 53. Details of the logistic regression model for the female population using all reported accidents as the outcome variable and {zero or one reported accident} and {two or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	1.6474	0.2596	
1) Mother's malaise score	-0.0718	0.0123	53.19
2) Mother's age	0.0378	0.0097	26.34
Model log-likelihood (d.f.)	3057 (5023)		

We now change the response variable to be (iii) and after fitting the intercept the variable which gave the most significant improvement in fit to the model was mother's malaise score with a Chi-squared value of 18.32 (1 d.f.). The parameter estimates for this model are given in Table 54 and no further improvement in the model was possible using the 0.1% criterion. Extending the acceptance criterion to one percent we have some difficulty in selecting the next regressor variable to add into the model. The variables which gave the most significant improvement in fit (on 1 d.f.) were as follows with the Chi-squared statistic in brackets: parents' mean age (9.46), youngest parent's age (8.69), father's age (7.87), mother's age (8.85), eldest parent's age (8.53), female fought others (9.29).

Clearly there is little to choose from between parent's mean age or whether the female fought others. Parent's mean age was added into the model on the grounds

that it gave a very slightly more significant improvement in fit and after including parents' mean age in the model no further improvement in fit was possible from using the one percent level of acceptance.

TABLE 54. Details of the logistic regression model for the female population using all reported accidents as the outcome variable and {zero, one or two reported accident} and {three or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	4.1906	0.1534	
1) Mother's malaise score	-0.0975	0.0216	82.53
Model log-likelihood (d.f.)	1115 (5024)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effect of mother's malaise score is very similar to the effect of mother's malaise score described in Table 52.

We now change the response variable to be (iv) and after fitting the intercept the first variable to be included in the model is mother's malaise score with a Chi-squared value of 23.88 (1 d.f.). After the inclusion of mother's malaise score the variables which gave the most significant improvement in fit (on 1 d.f.) are as follows with the Chi-squared value in brackets;

- i) parents' mean age (13.39),
- ii) youngest parent's age (12.80),
- iii) father's age (10.78),
- iv) mother's age (13.01),
- v) eldest parent's age (11.70).

Again there is little to choose from between these variables and parents' mean age was added into the model on the grounds that it gave a very slightly more significant improvement in fit. The details for this model are shown in Table 55 and no further significant improvement was possible using the 0.1% criterion or the one percent criterion.

TABLE 55. Details of the logistic regression model for the female population using all reported accidents as the outcome variable and {zero reported accidents} and {three or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	1.9812	0.5295	
1) Mother's malaise score	-0.1103	0.0225	110.56
2) Parents' mean age	0.0672	0.0193	68.19
Model log-likelihood (d.f.)	986.3 (5023)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effects of mother's malaise score and the parental age-related variable are again very similar to the effects seen in the model described in Table 52.

6.3.6 LOGISTIC MODELLING OF NON-SUPERFICIAL REPORTED ACCIDENTS FOR THE FEMALES

We now change the outcome variable to be non-superficial injuries received from the reported accidents

and select the response variable to be (i). Once again after fitting the intercept we find we have some difficulty in selecting the first variable to be added into the model. The variables which gave the most significant improvement in fit on 1 d.f., with the Chi-squared statistic in brackets are as follows;

- i) mother's malaise score(16.92),
- ii) parents' mean age (13.55),
- iii) youngest parent's age (15.32),
- iv) father's age (9.82),
- v) mother's age (14.59),
- vi) eldest parent's age (10.00).

The variable which gave the most significant improvement in fit was mother's malaise score with a Chi-squared value of 16.92 (1 d.f.). We then had more difficulty when choosing the next regressor variable to be included and the variables which gave the most significant improvement in fit on 1 d.f. (with the Chi-squared values in brackets) are as follows; parents' mean age (11.72), youngest parent's age (13.07), father's age (8.37), mother's age (12.76), eldest parent's age (8.75).

The next regressor variable to be included in the model was the youngest parent's age and the parameter estimates for this model are shown in Table 56. No further significant improvement in the model could be obtained using the 0.1% significance but relaxing this criterion to the one percent level we could further improve the model by including the variable relating to whether the child fought others.

TABLE 56. Details of the logistic regression model for the female population using non-superficial injuries (from the reported accidents) as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	0.5973	0.1700	
1) Mother's malaise score	-0.0337	0.0087	12.77
2) Youngest parent's age	0.0225	0.0063	11.37
Model log-likelihood (d.f.)	5772 (5023)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent level. The effects of mother's malaise score and the age variable are similar to the effects seen in the model described in Table 52.

6.3.7 LOGISTIC MODELLING OF SUPERFICIAL REPORTED ACCIDENTS FOR THE FEMALES

We now change the outcome variable to be superficial injuries received from the reported accidents and select the response variable to be (i). After fitting the intercept to the female population we select the first variable to be added into the model. Once again we had some difficulty in selecting the first variable to be added into the model and the variables which gave the most significant improvement in fit on 1 d.f. are as follows;

- i) mother's malaise score (19.16),
- ii) parents' mean age (16.37),
- iii) youngest parent's age (18.44),
- iv) father's age (10.97),
- v) mother's age (19.02),
- vi) eldest parent's age (12.15).

Mother's malaise score was added into the model and we again had difficulty when choosing the next regressor variable to be entered into the model. The variables which gave the most significant improvement in fit on 1 d.f. (with the Chi-squared statistic in brackets) are as follows; parents' mean age (14.16), youngest parent's age (15.73), father's age (9.30), mother's age (16.71), eldest parent's age (10.64).

The next regressor variable to be included in the model was mother's age and the parameter estimates for this model are shown in Table 57. No further improvement in fit could be obtained using the 0.1% or one percent criterion.

TABLE 57. Details of the logistic regression model for the female population using superficial injuries (from the reported accidents) as the outcome variable and {zero reported accidents} and {one or more reported accidents} as the response grouping.

Parameter	estimate	s.e.	m.d.r.
Intercept	1.0408	0.2007	
1) Mother's malaise score	-0.0421	0.0101	18.92
2) Parents' mean age	0.0297	0.0074	18.76
Model log-likelihood (d.f.)	4472 (5023)		

The mean deviance reduction for each variable was calculated and each was significant at the one percent

level. The effects of mother's malaise score and the age variable are very similar to the effects seen in the model described in Table 52.

Making a general comparison of these final six female models we must note that mother's malaise score was included in each of the models and an age related variable was included in five of the models. The only model which did not include an parental age-related variable was the 'all reported accident' model using response grouping (iii). The effects of the mother's malaise score and the parental age-related variables were generally the same across models.

The effect of using different binary outcome groupings resulted in slightly different models. Each model included mother's malaise score and in addition the models for binary groupings (i), (ii) and (iv), also included a parental age-related variable.

Examining the logistic results across gender for all reported accidents, non-superficial and superficial/unidentified reported accidents we see that regardless of gender and in all but one female logistic regression model the final model always included mother's malaise score and an age-related variable. We now compare the models obtained from using the different regression techniques.

6.4 A COMPARISON OF THE REGRESSION MODELS

In this section we compare the results from the previous two sections, i.e. the GLIM and logistic regression analyses. We acknowledge that fewer variables and a larger data set were used in the logistic regression modelling. We recall that the cumulative reported accident variables were defined in Section 6.1, and that the binary grouped response variables used in the logistic regression were defined in Section 6.3.

Table 58 gives a summary of the final regression models fitted to the male population and the obvious result from Table 58 is that the use of different modelling methods and reported accident outcome variables results in different regression models.

When comparing the GLIM and logistic regression models for all reported accidents we note that the GLIM model included mother's malaise score, parents' mean age, child's neighbourhood and whether the child was restless. When using different outcome groupings in the logistic modelling of the full reported accident distribution we get slightly different results.

TABLE 58. A summary of the final regression models fitted to the male population using a 0.1 percent inclusion criterion. The variables in brackets are the variables that would be included in the models if the 1 percent criterion was used. OUT is the reported accident outcome variable defined in Section 6.3.

OUT	Variables included in the regression models	
	GLIM	resp LOGISTIC
i	Youngest parent's age (Neighbourhood)	1 (Neighbourhood)
ii	Whether male bullied (Youngest parent's age)	1 Whether male bullied (Youngest parent's age)
iii	Healy Goldstein score (Youngest parent's age) (Whether male bullied) (Neighbourhood)	1 (Youngest parent's age) (Whether male destroyed belongings)
iv	Youngest parent's age Mother's malaise score Whether male bullied	1 Youngest parent's age Mother's malaise score (Whether male bullied)
v	Mother's malaise score Youngest parent's age Whether male bullied (Healy Goldstein score) (Neighbourhood)	1 Mother's malaise score Youngest parent's age (Whether male squirmed)
vi	Mother's malaise score Youngest parent's age Whether male bullied (Neighbourhood) (Healy Goldstein score)	1 Mother's malaise score Youngest parent's age (Whether male squirmed (Whether male bullied)
vii	Mother's malaise score Eldest parent's age	1 Mother's malaise score Parents' mean age
viii	Mother's malaise score Parents' mean age Neighbourhood Whether male was restless (Whether male bullied)	1 Mother's malaise score Youngest parents' age (Whether male was restless)
		2 Mother's malaise sc. Parents' mean age Neighbourhood (Whether male was restless)
		3 Mother's malaise score Parents' mean age
		4 Mother's malaise score Parents' mean age (Whether male fought)

For example, using (i) as the response grouping the logistic model includes mother's malaise and youngest

parents' age. Using (ii) the logistic model includes mother's malaise score, parents' mean age and child's neighbourhood, and using (iii) the logistic model includes mother's malaise score and parents' mean age. The effects of mother's malaise score and the parental age-related variable are generally the same in each type of model but child's neighbourhood is included in the logistic model using outcome grouping (ii). We see therefore that different modelling techniques produce slightly different regression models.

We also observe that for the GLIM 'all reported accidents' regression model the most important type of neighbourhood was the poor neighbourhood where it was seen that males from this area had a higher mean reported accident rate than males from other neighbourhoods. In the logistic model however the most important neighbourhood for the 'all reported accidents' model was the rural neighbourhood where it was seen that males had a lower probability of having an accident than males from the other neighbourhoods.

Variables relating to particular characteristics of the males' personality were included in the 'all reported accidents' and non-superficial reported accidents GLIM regression models. Looking at the male 'all reported accident' GLIM model first we see that the effect of the variable relating to whether the male was restless was such that the males with the highest reported mean accident rate were those who were certainly restless. Looking at the non-superficial reported accident GLIM

model we see that the effect of the variable relating to whether the child bullied others was such that the males with the highest reported mean accident rate were those who somewhat bullied others.

The presence of these two variables could be evidence of accident proneness since they both relate to 'internal' attributes rather than environmental attributes. However, we must acknowledge that the display of these particular personality characteristics may be due to other causes such as particular aspects of the mother's malaise conditions. For example, a child may bully other children to get attention from his mother.

We now examine Table 59 which gives a summary of the regression models fitted to the female population. As seen in the results of the male regression modelling, the obvious feature from Table 59 is that the use of different modelling methods and different reported accident outcome variables produces different models.

When comparing the female GLIM and logistic regression models for all reported accidents we note that the GLIM model included mother's malaise score and mother's age. Using different response groupings of the dependent variable in the logistic modelling we get slightly different results. For example, using (i) as the response grouping the logistic model includes mother's malaise and youngest parents' age. Using (ii) the logistic model includes mother's malaise score and mother's age, and using (iii) the logistic model includes just mother's malaise score. It is thought that the response grouping

(iii) may not be adequate for modelling the female reported accidents since the tail of the reported accident distribution may be too short. We again see that different regression modelling techniques produce slightly different models.

TABLE 59. A summary of the final regression models fitted to the female population using a 0.1 percent inclusion criterion. The variables in brackets are the variables that would be included in the models if the 1 percent criterion was used. OUT is the reported accident outcome variable defined in Section 6.3.

Variables included in the regression models	
OUT	GLIM resp LOGISTIC
i	Whether female could settle 1 (Whether female could settle)
ii	Whether female could settle 1 Whether female could settle
	(Mean parents' age)
	(Neighbourhood)
iii	Whether female could settle 1 (Whether female could settle)
	(Neighbourhood)
iv	Mother's malaise score 1 Mother's malaise score
	Youngest parent's age
	(Whether female destroyed belongings)
v	Mother's malaise score 1 Mother's malaise score
	Youngest parent's age
	(Whether female destroyed belongings)
vi	Mother's malaise score 1 Mother's malaise score
	Youngest parents' age
	(Whether female destroyed belongings)
vii	Mother's malaise score 1 Mother's malaise score
	Mother's age
viii	Mother's malaise score 1 Mother's malaise score
	Mother's age
	(Whether female fought)
	(Whether female destroyed belongings)
	2 Mother's malaise score
	Mother's age
	3 Mother's malaise score
	4 Mother's malaise score
	Parents' mean age

6.5 DISCUSSION

From the regression results in this Chapter we conclude that by using different regression techniques and different reported accident outcome variables we can produce different regression models. However, looking at the full reported accident models, non-superficial and superficial/unidentified reported accident models we note that regardless of which outcome variable was used or which type of regression modelling method was used, the final regression model always included mother's malaise score and with one exception, a parental age-related variable. Using the three outcome variables and different response groupings produced slightly different models but essentially we find that the most influential variables are the mother's malaise score and a parental age-related variable. In addition it should be noted that generally the effects of mother's malaise score and the parental age-related variables are the same across models.

CHAPTER 7

DISCUSSION

In this Chapter we summarise and discuss the conclusions drawn from each stage of the analysis. Throughout the analysis strong evidence of differential accident reporting has been identified across the population and the results suggest that sections of the reported accident data are so unreliable that they lead to misleading conclusions.

7.1 CONCLUSIONS

The exploratory analysis produced strong evidence to indicate ethnic and gender differences in the reported accident distribution. Several reported accident trends and associations were identified and strong evidence of accident misreporting was found. High-risk sub-populations were highlighted and accident recall problems were uncovered. For example, the peculiar reporting phenomenon observed when examining head injuries and the child's social index grouping for which the most likely explanation must be accident misreporting.

The traditional accident modelling probability distributions, the Poisson and Negative Binomial

distributions were fitted to the full reported accident distribution of male and female sub-populations without success. Sub-populations of high risk were once again highlighted and it was found that children who displayed particular personality characteristics were associated with higher reported accident means than children without those characteristics. This could be interpreted as evidence of accident proneness although the display of a particular personality characteristic by a child may be due to other variables, for example, mother's health or living conditions.

Another interpretation is that in reality accidents may have been more frequent in particular sub-populations but reported less accurately by the parents because they considered them everyday and common events. For example, particular sub-populations may report different types of accidents with different degrees of accuracy, and have different thresholds when deciding in the seriousness of each accident.

This hypothesis is supported throughout the analysis. For example, although the full reported accident distribution can not be successfully modelled with a Negative Binomial distribution, it can be broken down into a non-superficial and superfical reported accident distribution, each of which can be modelled with a separate Negative Binomial distribution. In addition, the best fitting probability model for both males and females was the new Model VI, the Generalised Negative Binomial distribution taking into account accident misreporting

between the two and one accident counts. The results presented throughout the analysis strongly suggest that different types of accidents are recorded with different levels of accuracy in different sub-populations.

Further, when modelling the full reported accident distribution we also investigated the use of four alternative accident probability distributions. Evidence of accident misreporting was present in the full reported accident distribution and the introduction of a misreporting parameter into the best fitting alternative model overcame the problem.

It was thought that too much information was lost by the truncation of the zero accident count for the truncated traditional accident probability distributions models to be considered appropriate. However, the results of fitting the truncated traditional accident modelling distributions confirmed the presence of accident misreporting and validated the use of the generalised models which allowed for accident misreporting. For example, it was observed that the truncated Poisson model consistently underestimated the reported zero accident count for both sexes, whilst in contrast, the truncated Negative Binomial generally overestimated the reported zero accident count for both sexes.

In a search for evidence of proneness the coefficients of correlation between the numbers of accidents occurring in two consecutive time periods were calculated. However, it was evident that the accident times were rounded in some way and that any type of analysis which involved time

would be unreliable. Evidence of proneness using this method was therefore inconclusive.

The regression modelling provided support for our hypothesis of differential accident reporting. GLIM and logistic regression models were fitted to the male and female populations and significant influencing factors were identified. Different outcome variables were used and yielded different models. In terms of interpretation this means the full accident data population comprises of smaller accident sub-populations. The cumulative effects of adding in progressively less serious accidents into the reported accident outcome variable were investigated and again strong evidence of differential accident reporting was discovered.

In conclusion, the methodology of the analysis has demonstrated that sections of the reported accident data are so unreliable that they lead to misleading conclusions. Any accident analysis based on the full reported accident distribution must be treated with caution. There is strong evidence suggesting that different types of accidents are reported with different degrees of accuracy by different sections of the population. The regression analysis supports the use of generalised probability models to model the full reported accident distribution and the most sensible model is Model VI, the Generalised Negative Binomial distribution which takes into account accident misreporting from two to one accident.

7.2 FURTHER WORK

There are several areas where further research would be beneficial. For example, different probability models could be found, different parameter estimation methods could be used, and other regression models investigated. However, the most important area for further investigation must be in the area of recall error. Since recall error is something that is present in most types of retrospective studies it is not a problem that is confined to accident analysis.

Lessler (1989) proposed a research agenda of five components to examine the reduction of memory errors in survey research which should be pursued. These are as follows:

- 1) Research into the nature of the recall process.
- 2) Development of mathematical models for memory errors and their impact on survey estimates.
- 3) Development of survey procedures that will improve recall.
- 4) Design of experiments which are imbedded in surveys and allow one to quantify components of the error models.
- 5) Use of estimation procedures that allow one to adjust survey estimates for recall loss.

In particular, Sikkel (1985) has investigated 'Models for memory effects' (see Section 1.3.2) and more recently, Torelli and Trivellato (1990) have addressed the 'heaping' problem and define it as 'abnormal concentrations of responses at certain durations.' She examined

unemployment data using a proportional hazards approach and confirms the view that memory effects are data and model dependent. This is certainly demonstrated in our data since we have additional differential reporting effects to be taken into account across the population. For example, the strong evidence suggesting that different sub-populations of parents have different thresholds of deciding whether an accidental injury requires medical treatment.

Clearly there is a lot of scope for further investigation into the analysis of accident data.

REFERENCES

ADELSTEIN, A.M. (1952).
Accident Proneness: A criticism of the concept based upon
an analysis of shunter's accidents. J.R.S.S.(A), 115,
354-410.

ARBOUS, A.G. and KERRICH, J.E. (1951).
Accident statistics and the concept of accident proneness.
Biometrics, 7, 340-432.

ASHFORD, J.R. and HUNT, R.G. (1974).
The distribution of Doctor-patient contacts in the N.H.S..
J.R.S.S.(A), 137, 347-383.

BACKETT, E.M. (1965).
Domestic accidents, Geneva, World Health Organisation.

BASS, J. and MEHTA, K. (1980).
Developmentally-oriented safety surveys. Clinical
Pediatrics, 19, 83-96.

BHATTACHARYA, S.K. and HOLLA, M.S. (1965).
On a discrete distribution with special reference to the
theory of accident proneness. J.A.S.A., 60, 1060-1066.

BIJUR, P., STEWART-BROWN, S. and BUTLER, N. (1983).
Child behaviour and accidental injury in 11,966 pre-school
children. Paediatrics, *, **-***.

BLUM, A. and AMES, B. (1977).
Flame-retardant additives as possible cancer hazards.
Science, 195, 42-78.

BLUMENTHAL, S., DAHIYA, R.C. and GROSS, A.J. (1978).
Estimating the complete sample size from an incomplete
Poisson sample. J.A.S.A., 73, 182-187.

BOWMAN, K.O. (1984).
Extended moment series and the parameters of the Negative
Binomial Distribution. Biometrics, 40, 249-252.

CALDWELL, J.C. (1966).
Study of age misstatement among children in Ghana.
Demography, 3, 477-490.

CHAMBERS, E.G. and YULE, G.U. (1941).
Theory and observation in the investigation of accident
causation. J.R.S.S. Supplement, 7, 101-107.

CHI P.Y. (1980).
Testing for homogeneity: The Negative Binomial
distribution. Biometrika, 67, 252-254.

CLARKE, A. and WALTON, W.W. (1979)
Effect of safety packaging on aspirin ingestion by
children. Pediatrics, 63 687-693.

COHEN, A.C. (1959)
Estimation in the Poisson distribution when sample values of $(c+1)$ are sometimes erroneously reported as c . Annals of the Institute of Statistical Mathematics, Tokyo, 11, 189-193.

COHEN, A.C. (1960).
Estimating the parameters of a modified Poisson Distribution. J.A.S.A., 55, 139-144.

CONSUL, P.C. (1989).
Generalised Poisson distribution: properties and applications. Statistics: Textbooks and Monographs, Marcel Dekker, N.Y..

CONSUL, P.C. and JAIN, G.C. (1973).
A generalisation of the Poisson distribution. Technometrics, 15, 791-799.

DANEN, J., O'CONNOR, J.F. and BRIGGS, R. (1976).
The consequences of imitative behaviour in children: The 'Evel Knievel' syndrome. Pediatrics, 57, 418-419.

DAVID F.N. and JOHNSON N.C. (1952).
The truncated Poisson. Biometrics, 8, 275-285.

EDWARDS, C. and GURLAND, J. (1961).
A class of distributions applicable to accidents. J.A.S.A., 56, 503-517.

FELLER, W. (1943).
On a general class of "Contagious" distributions. Annals of Mathematical Statistics, 14, 389-400.

FELDMAN, K.W., SCHALLER, R.T., FELDMAN, J.A. et al. (1978).
Tap water scald burns in children. Pediatrics, 62, 1-7.

FROME, E.L. (1983).
The analysis of rates using Poisson regression models. Biometrics, 39, 665-674.

GILLINGS, D.B. (1974).
Some further results for bivariate generalizations of the Neyman type A distribution. Biometrics, 30, 619-628.

GREENWOOD, M. and YULE, G.U. (1920).
An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J.R.S.S.(A), 83, 255-279.

GUSTAFSSON, L.H. (1977)
Childhood accidents. Scandinavian Journal Society of Medicine, 5, 5-13.

- HINZ P. and GURLAMD, J. (1970).
A test of fit for the Negative Binomial and other contagious distributions. J.A.S.A., 65, 887-903.
- ILLINGWORTH, C., BRENNAN, P., JAY, A. et al. (1975).
200 injuries caused by playground equipment. British Medical Journal, 4, 332-334.
- IRWIN, J.O. (1959)
On the estimation of the mean of a Poisson distribution from a sample with the zero class missing. Biometrics, 15, 324-326.
- IRWIN, J.O. (1964)
The personal factor in accidents (a review article). J.R.S.S.(A), 127, 438-451.
- IRWIN, J.O. (1968).
The generalised Waring distribution applied to accident theory. J.R.S.S.(A), 131, 202-225.
- JOHNSON, N.L. and KOTZ, S. (1969).
Distributions in Statistics: discrete distributions. John Wiley and Sons, New York.
- KATTI, S.K., and GURLAND, J. (1962).
Some methods of estimation for the Poisson Binomial distribution, Biometrics, 18, 42-51.
- KEMP, C.D. (1967).
On a contagious distribution suggested for accident data. Biometrics, 23, 241-255.
- KEMP, C.D. and KEMP, A.W. (1956).
The analysis of point quadrant data. Australian Journal of Botany, 4, 167-174.
- LANGLEY, J.D. (1988).
The need to discontinue the use of the term "accident" when referring to unintentional injury events. Accident Analysis and Prevention, 20, 1-8.
- LANGLEY, J.D. and SILVA, P.A. (1982).
Accidents-parents' attitudes to prevention. Australian Paediatric Journal, 18, 247-249.
- LESSLER, J.T. (1989).
Reduction of memory errors in survey research, a research agenda. I.S.I. - 47th session. Paris. August 29 - september 6. 1989.
- LEVIN, D.L., MCLAIN, M.M. and BLAIR, M. (1976).
Hang 'em high: A direct effect of TV violence. Journal of Pediatrics, 89, 1034-1035.

MACKAY, A., HALPERN, J., MCLOUGHLIN E., et al. (1979).
A comparison of age-specific burn injury rates in five
Massachusetts communities. American Journal of Public
Health, 69, 1146-1150.

MARITZ, J.S. (1950).
On the validity of inferences drawn from the fitting of
Poisson and Negative Binomial distributions to observed
accident data. Psychology Bulletin, 47, 434-443.

MAYCOCK, G. and HALL, R.D. (1984).
Accidents at 4-Arm roundabouts. T.R.R.L.: Lab Report
1120, Department of Environment and Department of
Transport.

MCGUIRE, J.U., BRINDLEY, T.A., and BANCROFT, T.A. (1957).
The distribution of the European corn-borer larvae,
Pyrausta Neubilalis (Hbn), in field corn. Biometrics, 13,
65-78.

MCLOUGHLIN, E., CLARKE, N., STAHL, K. et al. (1977).
One pediatric burn unit's experience with sleep-wear-
related injuries. Pediatrics, 60, 405-409.

MILLER, F.J.W., COURT, S.D.M., WALTON, W.S. and KNOX,
E.G. (1960).
Growing up in Newcastle-upon-Tyne. Oxford University
Press, London.

MOHR, D.L. and CLEMMER (1988).
The "accident prone" worker: an example from heavy
industry. Accident Analysis and Prevention, 20, 123-127.

MOORE, P.G. (1952).
The estimation of the Poisson parameter from a truncated
distribution. Biometrika, 39, 247-251.

NEDELMAN, J. (1983).
A Negative Binomial model for sampling mosquitoes in a
malaria survey. Biometrics, 39, 1009-1020.

NELDER, J.A. and WEDDERBURN R.W.M. (1972).
General linear models. J.R.S.S.(A), 135, 370-384.

NEWBOLD, E.M. (1927).
Practical applications of the statistics of repeated
events, particularly to industrial accidents.
J.R.S.S.(A), 90, 487-547.

NEYMAN, J. (1939).
On a new class of 'contagious' distributions, applicable
in entomology and bacteriology. Annals of Mathematical
Statistics, 10, 35-57.

PAUL, S.R. and PLACKETT, R.L. (1978).
Inference sensitivity for Poisson mixtures. Biometrika,
65, 691-602.

- PLIONIS, E.M. (1977).
Family functioning and childhood accident occurrence.
American Journal of Orthopsychiatry, 47, 250-263.
- RAO C.R. and RUBIN H. (1964).
On a characterization of the Poisson distribution.
Sankya(A), 26, 295-298.
- READ, J.H., BRADLEY, E.J., MORISON, J.D., et al(1963).
The epidemiology and prevention of traffic accidents
involving child pedestrians. Canadian Medical Association
Journal, 89, 687-701.
- RIDER, P.R. (1955).
Truncated Binomial and Negative Binomial distributions.
J.A.S.A., 50, 877-883.
- RIVARA, F.P. (1982).
Epidemiology of childhood injuries: review of current
research and presentation. American Journal of Diseases
of Childhood, 136, 399-405.
- ROGHMAN, K., HECHT, P. and HAGGERTY, R. (1975).
Coping with stress, Child Health and the Community.
Haggerty, Roghman and Pless (Eds), New York, Wiley.
- RUDOLFER, S.M. (1990).
A Markov chain model of extrabinomial variation.
Biometrika, 77, 255-264.
- RUTHERFORD, G.W. (1979).
Injuries associated with playground equipment. U.S.
Consumer Product Safety Commission.
- SAMPFORD, M.R. (1955).
Truncated Negative Binomial distribution. Biometrika, 42,
58-69.
- SANATHANAN, L. (1977).
Estimating the size of a truncated sample. J.A.S.A., 72,
669-672.
- SANDELS, S. (1970).
Young children in traffic. British Journal of Educational
Psychology, 40, 111-115.
- SIKKEL, D. (1985).
Models for memory effects. J.A.S.A., 80, 835-841.
- SPENCE, J. , WALTON, W.S., MILLER, F.J.W. and
COURT, S.D.M. (1954).
A thousand families in Newcastle-upon-Tyne. Oxford
University Press, London.

- STARFIELD, B. (1982).
Child health and socio-economic status. American Journal of Public Health, 72, 532-534.
- STATISTICAL BULL. METROPOL LIFE INSURANCE COMPANY (1978).
Accidental drownings by age and activity, 59, 2-5.
- STEWART-BROWN, S., PETERS, T., GOLDING, J. and BIJUR, P. (1985).
Case definition in childhood accident studies. A vital factor in determining results. International Journal of Epidemiology, 15, 352-359.
- SUDMAN and BRADMAN, (1973).
Effects of time and memory factors on response in surveys. J.A.S.A., 68, 805-815.
- TORELLI, N., and TRIVELLATO, U. (1990).
Modelling inaccuracies in job-search duration data. Research report. Dipartimento di Scienze Statistiche. Universita di Padova. Via San Fransesco, 33-35121 Padova, Italy.
- TRIPATHI R.C. and GUPTA, R.C. (1984).
Statistical inference regarding the generalised Poisson distribution, Sankhya(B), 46, 166-173.
- WADSWORTH, J., BURNELL, I., TAYLOR, B. and BUTLER, N. (1983).
Family types and accidents in pre-school children. Journal of Epidemiology and Community Health, 37, 100-104.
- WALLER, J.A. (1987).
Injury as disease. Accident Analysis and Prevention, 19, 13-30.
- WALLER, J.A. and KLEIN, D. (1973).
Society, energy and injury: Inevitable triad? Research Directives Toward the Reduction of Injury, NIH publication, U.S. Department of Health, Education and Welfare, 1-37.
- WEBER, D.C. (1971).
Accident Rate Potential: An application of multiple regression analysis of a Poisson process. J.A.S.A., 66, 285-288.
- W.H.O.
Accidents in childhood, facts as a basis for prevention (1957). Report of an advisory group, W.H.O. technical report, 118, Geneva.
- WILLIAMS, A.F. and ZANDOR, P. (1977).
Injuries to children in automobiles in relation to seating location and restraint use. Accident Analysis and Prevention, 9, 69-76.

WYSHAK, G. (1974).

A program for estimating the parameters of the truncated
Negative Binomial distribution. Applied Statistics,
Statistical Algorithms.